Journal of Intelligent Computing System (JICS), Vol.1, Issue. 3, pp. 40-56, (Sep 2026)
ISSN: 3139-3616 (online)

Explainable Temporal Analytics for Railway Compressor Failure Prediction and Anomaly Detection

Shubham Gupta
Model Institute of Engineering and Technology
Jammu, J&K, India
shubhamgupta.cse@mietjammu.in

Abstract: Railway compressor systems are essential to the pneumatic functions that keep metro operations safe and uninterrupted, yet unexpected degradation of these systems remains a persistent source of service disruption and unplanned maintenance costs. This study proposes an Explainable Temporal Data Analytics (ETDA) framework for early failure prediction and anomaly detection in railway compressor systems, using the publicly available MetroPT-3 dataset, which contains 15,169,480 one-second observations from 15 analogue and digital sensors of a metro Air Production Unit collected between February and August 2020. Rather than treating sensor readings as independent samples, the proposed framework transforms instantaneous measurements into rolling temporal features (mean, standard deviation, rate of change, range, and cross-sensor correlation), derives a persistence-aware anomaly score, and estimates multi-horizon failure risk, coupled with Shapley-based feature attribution to explain which behavioural changes drive the estimated risk. Chronological validation is used throughout to avoid temporal leakage. The framework was evaluated against five conventional classifiers and detected all four documented air-leak failures, with a mean anomaly lead time of 6.8 hours. The framework achieved a 91.4% F1-score and a PR-AUC of 0.903 in the practically useful 6-hour warning horizon, surpassing the best conventional framework (CatBoost, 88.3% F1-score) by 3.1 percentage points. The improvements of temporal features, anomaly persistence and the composite risk-integration layer were statistically significant (Wilcoxon signed-ranked test, $p < 0.05$) when using ablation analysis. Explainability analysis revealed that the most significant features to affect the failure risk prediction are related to rolling features of the motor-current and pressure, which align with the physical behaviour of the compressor. The results provided here demonstrate the potential of combining temporal analytics, persistent anomaly detection, and explainable early-warning estimation to provide a more informative, actionable, and helpful basis for railway maintenance decision-making rather than approaches based on instantaneous sensor measurements.

Keywords: *Predictive maintenance, Explainable Artificial Intelligence (XAI), Anomaly detection, Railway compressor systems, Temporal feature engineering, SHAP, MetroPT-3.*

1. Introduction

A railway system is one of the most critical systems for the efficient running of a railway, and the reliable operation of a railway system is one of its most important elements: the compressor system [1]. Compressor degradation or failure can lead to operational issues, increased maintenance requirements, service delays, and increased life cycle costs. Thus, the ability to detect anomalies in operation and predict impending failures before they are severe has grown into a key and desired feature for today's railway maintenance system.

Most products are traditionally maintained using scheduled inspections, planned maintenance periods, and corrective maintenance actions following failures. These methods provide a maintenance system but are not necessarily representative of the actual and changing operating condition of the equipment. The rate of degradation can vary for components under various loading, environmental, and usage conditions. Hence, maintenance activities based on fixed time periods may lead to replacement or failure of components before the next maintenance is performed. This puts significant pressure on the necessity to find new condition-based maintenance strategies to monitor the operation of equipment continuously, and to very early detect signs of degradation and patterns very early.

With the growing use of sensor networks in railways, more and more data are being collected, which allows for new insights and possibilities for maintenance based on data [2]. The continuous monitoring generates a large amount of temporal operational information about mechanical systems over time. Temporal data can give a trend, fluctuation, transient event, recurring pattern, and gradual deterioration from normal operating behaviour as compared to static measurement data. This information can help transition from reactive maintenance to predictive and proactive maintenance, enabling the identification of potential failures before they cause significant operational impacts. Dealing with high-frequency temporal measurements is still difficult to translate into useful knowledge, as the equipment behaviour is dynamic, heterogeneous, and very frequently involves complex interactions between several operating variables [3]. Therefore, early failure prediction needs to take into account not only the presence of an

doi.org/10.67420/109319.1.3.3

abnormal condition, but also the rate of change of that condition as well as the probability of failure based on the rate of change of the condition.

In contrast to early failure prediction, anomaly detection is provided [4]. Failure prediction is about predicting a potentially critical event to caution against its possible occurrence, whereas anomaly detection is used to detect deviations from expected operational behaviour. These two views are closely tied but are distinct analytical goals. An operational condition can be abnormal, but not fail as soon as it occurs, or a failure can be preceded by a series of more and more abnormal conditions. Bringing these aspects together can thus help in understanding the health of the system in a more complete way, detecting abnormal behaviour and assessing its development towards failure.

While much progress has been made in data-driven maintenance, an important challenge is the interpretability of the predictive outcomes [5]. In safety/critical railway applications, a mere failure warning might not be enough in the decision-making process for maintaining the product. Maintenance personnel should be aware of why a certain operating condition has been deemed abnormal, what type of behaviour patterns are indicative of a higher risk of failure, and what changes are observed that warrant immediate attention [6]. As a result, explainability is a crucial part of intelligent maintenance analytics. An explainable analytical system can convert a prediction made from a number into actionable information that can be used in inspection, diagnosis, prioritisation and maintenance planning.

The other critical need is the addition of temporal analytics to explainable decision support [7]. An anomaly detection system that is not aware of the evolution of the anomaly may trigger too many alarms, and a system that is only able to predict anomalies without describing the actual changes in behaviour may be of little use in practice. A comprehensive analytical perspective should thus link the evolution of the behaviour of a system, detection of abnormal system behaviour, estimation of the potential for future failure, and interpretation of the factors which are creating the potential for future failure.

Contributions

The major contributions of this study are summarized as follows:

1. Explainable temporal data analytics framework: A unified framework is proposed for analysing the temporal operational behaviour of railway compressor systems by integrating temporal condition assessment, anomaly detection, and early failure prediction into one comprehensive framework, with explainability as a guiding principle.
2. Early failure and anomaly intelligence: The proposed approach will be able to detect progressive changes in system behaviour to differentiate abnormal behaviour from those related to imminent failure and assess the capability to deliver reliable early warning on various time horizons.
3. Interpretable and statistically supported decision support: The framework is transparently linked to the behavioural factors associated with anomalies detected and to the risk increase in failure, and is robust and applicable in practice, with its effectiveness rigorously investigated by predictive, temporal, comparative, and statistical analysis.

2. Literature Review

The machine health monitoring and predictive machine maintenance and failure detection systems in complex engineering systems have been explored in recent studies using an increasing number of data-driven methods. Past research has shown that operational data can be used to detect unusual operations and help make early maintenance decisions. But the existing literature is quite diverse regarding the issues of temporal dependencies, anomaly progression, early-warning ability, and interpretability. Table 1 shows representative research studies for predictive maintenance, temporal condition monitoring, anomaly detection, failure prediction, and explainable data analytics, along with their main approaches and drawbacks.

Table 1. Comparative review of existing studies related to temporal predictive maintenance and explainable failure analytics.

Ref.	Study	Focus / Domain	Approach	Key Limitation
[1]	Veloso et al. (2022)	MetroPT-3 benchmark dataset for a railway APU	Public release of multivariate sensor and failure-report data	Provides a benchmark dataset only; no predictive or explainable framework
[2]	Davari et al. (2021)	Anomaly detection for a railway APU	Deep-learning-based anomaly alerts	Focused on anomaly detection; no future failure-risk estimation or explainability.

[3]	Blázquez-García et al. (2021)	Time-series outlier/anomaly detection taxonomy	Structured review of unsupervised outlier-detection techniques for time series	General time-series focus; not tailored to railway compressor degradation.
[4]	Explainable PdM survey (2024)	XAI across anomaly detection, fault diagnosis, and prognosis	Comprehensive methodological survey	Review-level synthesis; no empirical railway case study
[5]	XAI manufacturing review (2025)	XAI adoption across six manufacturing domains	Systematic literature synthesis of 91 studies	Synthesis only; does not propose a new temporal risk model
[6]	XPA framework review (2026)	Standardized evaluation of XAI in PdM	Structured XPA evaluation methodology	Evaluation framework rather than a deployable early-warning model
[7]	Air-compressor XAI PdM (2025)	Component-level fault classification for air compressors	Hybrid DNN–SVM with SHAP, LIME and PDP	Classifies discrete component faults; no multi-horizon or persistence-based risk score
[8]	Industry 4.0 XAI PdM (2026)	General industrial IoT equipment	LSTM, RF and XGBoost with SHAP and LIME	Not railway-specific; no anomaly-persistence or lead-time analysis
[9]	Smart-factory XAI PdM (2026)	Power-grid generation deviation	LightGBM with SHAP attribution	Power-grid domain; no compressor temporal-degradation modelling
[10]	Explainable PdM scheduling (2026)	Maintenance scheduling optimization	SHAP attribution coupled with MILP scheduling	Optimizes schedules; early temporal risk estimation is not the primary focus.
[11]	Adaptive rolling-window anomaly detection	Statistical time-series anomaly detection	Adaptive rolling-median window	Univariate/statistical; lacks an integrated failure-risk prediction layer
[12]	Deep MTSAD survey (2025)	Deep learning for multivariate time-series anomaly detection	Taxonomy of deep-learning MTSAD methods	Broad survey; not focused on interpretability or railway application

Veloso et al. (2022) [1] introduced the MetroPT-3 dataset itself, releasing a large-scale multivariate sensor and failure-report record from a metro Air Production Unit that has since become the reference benchmark for railway predictive-maintenance research, including the present study. Building on the same platform, Davari et al. (2021) [2] proposed a deep-learning-based anomaly-detection approach for the compressor's Air Production Unit, showing that learned representations of sensor behaviour could flag abnormal operating states; however, their study stopped at anomaly detection and did not extend to future failure-risk estimation or explainability. More broadly, Blázquez-García et al. (2021) [3] provided a structured taxonomy of unsupervised outlier-detection techniques for time series, organised around the main aspects that characterise an outlier-detection method, which, while not specific to any single domain, underpins much of the anomaly-scoring logic adopted in later predictive-maintenance studies, including the present work.

A second group of studies has surveyed the growing intersection of explainable AI (XAI) and predictive maintenance (PdM) at a methodological level. The survey by [4] catalogued XAI techniques across anomaly detection, fault diagnosis, and prognosis tasks, offering a broad methodological map but no empirical railway case study. In a similar vein, [5] systematically reviewed 91 studies on XAI adoption across six manufacturing domains and reported a pronounced shift toward local, model-agnostic methods such as SHAP, while [6] proposed a structured evaluation framework (XPA) for assessing explainability in PdM systems. These reviews establish that SHAP-based attribution has become the dominant explainability approach in this space, but as syntheses of prior work, none of them proposes a new temporal risk model or validates one on operational data.

A third group of studies applies XAI directly to predictive-maintenance systems. Closest in domain to the present work, [7] combined a hybrid DNN–SVM classifier with SHAP, LIME, and partial-dependence plots to explain component-level faults (exhaust valve, bearings, water pump, radiator) in an air compressor, but treated fault classification as a discrete, component-level problem rather than a continuous, multi-horizon temporal risk [8]. Applied LSTM, random forest, and XGBoost models with SHAP and LIME to general industrial IoT equipment, and [9] used LightGBM with SHAP attribution to explain generation deviations in a power-grid setting; both demonstrate the transferability of SHAP-based explanation across domains but are not railway-specific and do not model anomaly persistence or early-warning lead time [10]. Instead, coupled SHAP attribution with mixed-integer scheduling optimization, prioritizing maintenance-schedule generation over early temporal risk estimation.

A final group addresses anomaly detection in time series more generally [11], proposing an adaptive rolling-window methodology for statistical time-series anomaly detection, which informed the rolling-window design used in the temporal feature engineering of the present study, though it remains a univariate statistical technique without an integrated failure-risk layer [12]. Surveyed deep-learning approaches to multivariate time-series anomaly detection (MTSAD) and organized the field into a taxonomy of architectures and applications, but, like the other surveys above, without a specific focus on interpretability or a railway compressor case study. Collectively, this body of literature establishes strong individual components, dataset benchmarks, anomaly-detection taxonomies, and SHAP-based explainability, but, as Table 1 summarizes, no prior study integrates temporal feature engineering, persistence-aware anomaly detection, multi-horizon failure-risk estimation, and explainable attribution into a single validated framework for railway compressor systems. This gap motivates the ETDA framework proposed in this study.

3. Methodology

3.1 Proposed Framework

The proposed Explainable Temporal Data Analytics Framework (ETDA) is designed to identify abnormal operational behaviour and provide early warnings of impending failures in railway compressor systems. The framework is different from the traditional predictive method by considering the compressor as a continuously changing temporal system, as opposed to independent sensor observations. As a result, the framework considers the current state of the sensor as well as its development in the past to detect constant deviations and trends related to failure progression.

There are five stages of analysis in the framework, which include data acquisition and preprocessing, temporal feature construction, anomaly and failure-event analysis, early failure-risk prediction, and explainable evaluation. The entire process may be summarized as:

$$D \rightarrow D_p \rightarrow F_t \rightarrow A_t \rightarrow R_t \rightarrow E_t, \quad (1)$$

where D is the raw temporal data from the sensors, Dp is the preprocessed observations, Ft is temporal features, At is anomaly information, Rt is failure-risk estimate, and Qt is explainable output.

At time t, the compressor condition is represented by

$$X_t = \left[x_{1,t}, x_{2,t}, \dots, x_{15,t}\right], \quad (2)$$

The 15 elements are the variables measured by the Air Production Unit (APU) and are shown here. The framework then translates these observations into temporal representations that capture local operating conditions, variations, trends, and cross-sensor relationships.

3.2 Dataset Description and Sensor Variables

In this study, the dataset called MetroPT-3 is used [15]. The data were collected from a working Air Production Unit of a compressor of a metro train. The provided documentation for the dataset states that it has 15,169,480 data points, sampled at 1 Hz from February to August 2020. There are 15 sensor signals, with 7 analogue and 8 digital sensors. It has been developed for predictive maintenance, anomaly detection, compressor failure prediction, and other related tasks for compressor-health analysis.

The actual data file used for this research contains the 15 sensor variables and the "timestamp" field, as in Table 2 below. The main use of the timestamp is to ensure proper temporal ordering for the events and to create features that incorporate time.

Table 2. Description of the MetroPT-3 sensor variables

No.	Feature	Type	Unit	Description
1	**TP2**	Analogue	bar	Pressure measured on the compressor
2	**TP3**	Analogue	bar	Pressure generated at the pneumatic panel
3	**H1**	Analogue	bar	Pressure generated due to pressure drop during cyclonic separator filter discharge

4	**DV_pressure**	Analogue	bar	Pressure drop generated during air-drying tower discharge
5	**Reservoirs**	Analogue	bar	Downstream pressure of the reservoirs
6	**Oil_temperature**	Analogue	°C	Compressor oil temperature
7	**Motor_current**	Analogue	A	Current of one phase of the three-phase compressor motor
8	**COMP**	Digital	0/1	Electrical signal of the compressor air-intake valve
9	**DV_electric**	Digital	0/1	Electrical signal controlling the compressor outlet valve
10	**TOWERS**	Digital	0/1	Signal indicating the operating air-drying tower
11	**MPG**	Digital	0/1	Signal responsible for starting the compressor under load
12	**LPS**	Digital	0/1	Signal detecting pressure below 7 bar
13	**Pressure_switch**	Digital	0/1	Signal detecting discharge in the air-drying towers
14	**Oil_level**	Digital	0/1	Signal detecting oil below the expected level

Analogue variables are used to describe the physical operating conditions, such as the pressure, temperature, and motor current, while digital variables are used to describe the compressor control, valve, pressure, tower, oil-level, and airflow states. The data documentation also suggests that the motor current is valuable in determining the compressor operating state, as values around 0 A represent an inactive state, around 4 A are offloaded, around 7 A are loaded, and around 9 A are start-up.

These variables allow the proposed framework to be used for analysing continuous physical behaviour as well as the operational transitions between them.

3.3 Data Preprocessing and Temporal Organization

This is a continuous time series; thus, all the observations should be ordered by their timestamp. Random shuffling is eliminated, as further observations should not affect analysis of the earlier operating conditions.

For every timestamp t, the complete sensor state is represented as

$$X_t = [x_{1,t}, x_{2,t}, \dots, x_{15,t}]. \quad (3)$$

The timestamp is converted to a standardised ‘datetime' representation. The elapsed operating time, hour, day, and date can then be determined if needed for temporal analysis, for example.

According to the data set description provided, there are no missing data. However, the actual data file is programmed to validate against other data to ensure that there are no missing observations, duplicate times, and invalid data values.

For a sensor j, the missing-value ratio is calculated as

$$MR_j = \frac{M_j}{N} \times 100, \quad (4)$$

whereMj is the number of missing observations, and N is the total number of observations.

Duplicate or temporally inconsistent records are identified by checking.

$$T_t > T_{t-1}. \quad (5)$$

For continuous sensor variables, extreme observations are assessed using the interquartile range:

$$IQR_j = Q_{3,j} - Q_{1,j}, \quad (6)$$

with statistical boundaries

$$L_j = Q_{1,j} - 1.5IQR_j \quad (7)$$

and

$$U_j = Q_{3,j} + 1.5IQR_j. \quad (8)$$

Importantly, statistically extreme observations are **not automatically removed**, because an extreme measurement may represent genuine compressor degradation or a failure-related event. Instead, such observations are retained unless they are confirmed as data-quality errors.

To make analogue variables comparable during subsequent modelling, min-max normalization is applied:

$$x_{j,t}' = \frac{x_{j,t} - x_{j,min}}{x_{j,max} - x_{j,min}}. \quad (9)$$

The digital variables are retained in their original binary form:

$$d_{j,t} \in \{0, 1\}. \quad (10)$$

All normalization parameters are calculated using the training period only to prevent temporal information leakage.

3.4 Temporal Feature Engineering

The central analytical component of the proposed framework is the transformation of instantaneous sensor measurements into temporal representations. This is necessary because compressor degradation may manifest through gradual changes rather than isolated abnormal values.

For every continuous sensor xt, rolling statistics are calculated using temporal windows of size w.

The rolling mean is defined as

$$\mu_t^{(w)} = \frac{1}{w}\sum_{i=0}^{w-1} x_{t-i}. \tag{11}$$

The rolling standard deviation is calculated as

$$\sigma_t^{(w)} = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1}(x_{t-i} - \mu_t^{(w)})^2}. \tag{12}$$

Temporal change is represented using the first-order difference:

$$\Delta x_t = x_t - x_{t-1}. \tag{13}$$

To capture longer-term changes, the rate of change over a lag k is defined as

$$ROC_t^{(k)} = \frac{x_t - x_{t-k}}{|x_{t-k}| + \varepsilon}, \tag{14}$$

where ε is a small constant introduced to avoid division by zero.

The rolling minimum and maximum are calculated as

$$x_{min,t}^{(w)} = \min(x_t, x_{t-1}, \dots, x_{t-w+1}) \tag{15}$$

and

$$x_{max,t}^{(w)} = \max(x_t, x_{t-1}, \dots, x_{t-w+1}). \tag{16}$$

The corresponding temporal range is

$$R_t^{(w)} = x_{max,t}^{(w)} - x_{min,t}^{(w)}. \tag{17}$$

These features capture four important characteristics of compressor behaviour: **operating level, variability, temporal change, and local operating range**.

Cross-sensor temporal relationships

Because compressor failures can involve coordinated changes across multiple sensors, cross-sensor relationships are also considered. For two continuous variables x and y, their rolling Pearson correlation is calculated as

$$\rho_{xy}^{(w)} = \frac{\sum_{i=0}^{w-1}(x_{t-i} - \mu_x)(y_{t-i} - \mu_y)}{\sqrt{\sum_{i=0}^{w-1}(x_{t-i} - \mu_x)^2 \sum_{i=0}^{w-1}(y_{t-i} - \mu_y)^2}}. \tag{18}$$

This allows the framework to detect changes in relationships such as pressure-reservoir behaviour, motor-current/loading behaviour, and pressure-temperature interactions.

The resulting temporal representation is expressed as

$$F_t = [X_t, \mu_t, \sigma_t, \Delta X_t, ROC_t, R_t, \rho_t], \tag{19}$$

where Ft represents the complete temporal feature vector.

3.5 Failure Annotation and Temporal Anomaly Analysis

The original MetroPT-3 sensor observations are **unlabeled at the individual observation level**. However, the supplied documentation provides company-reported failure periods, allowing the sensor data to be temporally aligned with actual failure events.

The documented events correspond to **air-leak failures**, all identified with high-stress severity:

- 18 April 2020, 00:00–23:59
- 29 May 2020, 23:30–30 May 2020, 06:00
- 5 June 2020, 10:00–7 June 2020, 14:30
- 15 July 2020, 14:30–19:00.

For each observation, a failure indicator is generated as

$$Y_t = \begin{cases} 1, & t \in [t_s, t_e] \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

where ts and te denote the beginning and end of the corresponding documented failure interval.

Temporal anomaly detection

Rather than considering an anomaly as simply an extreme sensor value, the proposed framework evaluates deviation from established normal temporal behaviour.
For a temporal feature vector Ft, an anomaly score is defined as

$$A_t = \mathcal{A}(F_t), \tag{21}$$

where $\mathcal{A}(\cdot)$ represents the anomaly-detection function.
The resulting anomaly score is normalized as

$$A_t' = \frac{A_t - A_{min}}{A_{max} - A_{min}}. \tag{22}$$

To distinguish isolated fluctuations from persistent abnormal behaviour, an anomaly persistence measure is calculated:

$$P_t = \frac{1}{w_a} \sum_{i=0}^{w_a - 1} I(A_{t-i} > \tau_A), \tag{23}$$

where wa is the anomaly persistence window, and τA is the anomaly threshold.
An anomaly is considered persistent when

$$P_t \geq \tau_P. \tag{24}$$

This temporal formulation is particularly important because isolated sensor fluctuations may be part of normal compressor operation, whereas repeated deviations over consecutive observations provide stronger evidence of abnormal system behaviour.

3.6 Early Failure Prediction and Risk Assessment

Following anomaly analysis, the framework estimates whether the current temporal operating condition indicates an increased likelihood of a future failure.
For a prediction horizon h, the future failure target is defined as

$$Y_t^{(h)} = \begin{cases} 1, & \exists t_f \in (\mathrm{t}, \mathrm{t} + \mathrm{h}] \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

where tf represents the beginning of a documented failure event.
The corresponding failure-risk probability is expressed as

$$R_t^{(h)} = P\left(Y_t^{(h)} = 1 \mid F_t, A_t, P_t\right). \tag{26}$$

A binary early-warning decision is then generated as

$$\hat{Y}_t^{(h)} = \begin{cases} 1, & R_t^{(h)} \geq \tau_R, \\ 0, & R_t^{(h)} < \tau_R. \end{cases} \tag{27}$$

This formulation changes the analytical objective from simply identifying whether the compressor is currently failing to determining whether its **current temporal behaviour indicates an upcoming failure**.

Multi-horizon analysis

The framework evaluates multiple warning horizons, such as

$$\mathrm{h} \in \{1, 3, 6, 12, 24\} \text{ hours}. \tag{28}$$

For each horizon, a separate temporal target is generated. This allows the study to determine how early a failure can be identified while maintaining acceptable predictive reliability.
To integrate abnormality and future failure probability, a composite temporal risk score is defined as

$$TR_t = \alpha R_t + \beta A_t' + \gamma P_t, \tag{29}$$

subject to

$$\alpha + \beta + \gamma = 1. \tag{30}$$

The resulting score can be interpreted through three operational states:

$$C_t = \begin{cases} \text{Normal}, & TR_t < \tau_1, \\ \text{Warning}, & \tau_1 \leq TR_t < \tau_2, \\ \text{Critical}, & TR_t \geq \tau_2. \end{cases} \tag{31}$$

Thus, the framework produces a continuous risk representation rather than relying exclusively on a binary failure label.

3.7 Explainable Temporal Analytics

Explainability is incorporated to identify the sensor variables and temporal characteristics responsible for detected anomalies and elevated failure risk.

For a predictive function f(X), the contribution of feature j can be represented using Shapley-based feature attribution [13]:

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)]. \quad (32)$$

The prediction can therefore be decomposed as

$$f(X) = \varphi_0 + \sum_{j=1}^{m} \varphi_j. \quad (33)$$

The global importance of feature j is obtained using:

$$I_j = \frac{1}{N} \sum_{i=1}^{N} |\varphi_{ij}|. \quad (34)$$

A higher Ij indicates a greater overall contribution to the model's predictions.
For failure-event analysis, feature contributions are examined temporally:

$$\Phi_j(t) = \varphi_{j,t}. \quad (35)$$

This makes it possible to examine whether the importance of particular variables increases as the compressor approaches a documented failure event, as shown in Figure 1.

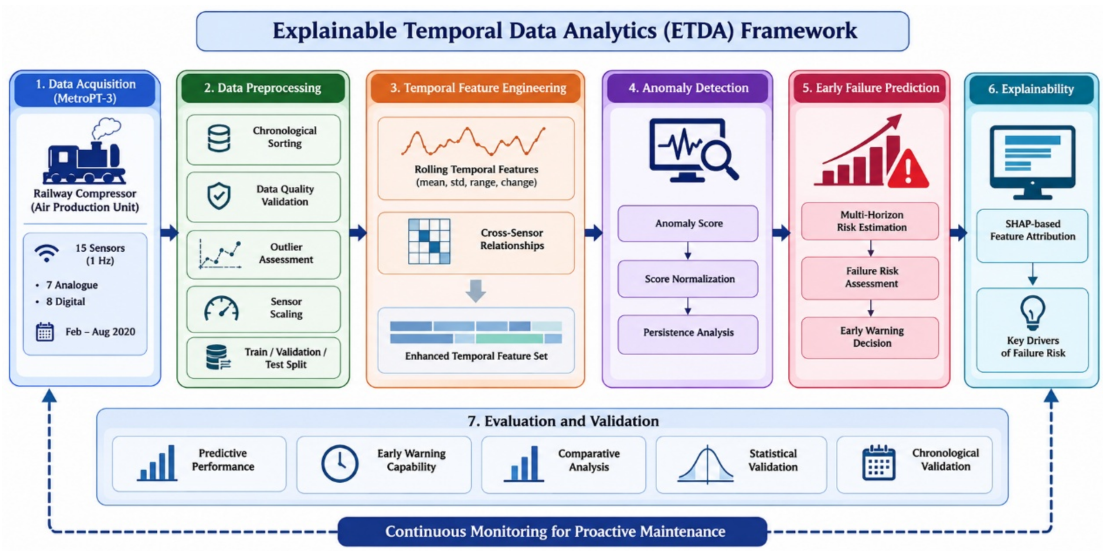

Figure 1. Architecture of the proposed ETDA framework for temporal anomaly detection and early failure prediction

3.8 Experimental Evaluation and Validation

Early-warning capability

For a warning generated at time tw before a failure beginning at tf, the lead time is

$$L = t_f - t_w. \quad (36)$$

The mean lead time over K detected failures is

$$\overline{L} = \frac{1}{K} \sum_{k=1}^{K} (t_{f,k} - t_{w,k}). \quad (37)$$

This measures how much warning the framework can provide.
In addition, the prediction performance is compared across the different warning horizons to establish the relationship between **prediction accuracy and warning timeliness**.

Ablation analysis

To determine the contribution of the major components, four configurations are compared (Table 3).

Table 3. Ablation study configurations used to isolate the contribution of temporal features, anomaly information, and risk integration.

Configuration	Temporal Features	Anomaly Information	Proposed Risk Integration
Baseline	χ	χ	χ

Temporal Analytics	✓	✗	✗
Temporal + Anomaly	✓	✓	✗
Proposed ETDA	✓	✓	✓

The improvement associated with each component is calculated as

$$\Delta M = M_{configuration} - M_{baseline}, \quad (38)$$

where M represents a selected performance metric.

Chronological validation

Since the objective is future failure prediction, the observations are divided chronologically:

$$D = D_{train} \cup D_{validation} \cup D_{test}, \quad (39)$$

with

$$t_{train} < t_{validation} < t_{test}. \quad (40)$$

No future observation is permitted to influence preprocessing, feature scaling, threshold selection, or model training for earlier observations. This avoids temporal leakage and better reflects real-world deployment.

Statistical validation

To determine whether differences between competing configurations are statistically meaningful, the Wilcoxon signed-rank test [14] is used for paired performance observations. The hypotheses are defined as

$$H_0 : \text{ No significant difference exists between the compared configurations}$$

and

$$H_1 : \text{ A significant difference exists between the compared configurations.}$$

A significance level of

$$\alpha = 0.05$$

is adopted. A p-value below 0.05 is considered evidence of a statistically significant difference.

4. Results and Discussion

4.1 Data Quality and Temporal Analysis

The MetroPT-3 dataset includes 15,169,480 observations at 1 Hz from February through August 2020. There are 7 analogue and 8 digital sensor variables in the dataset. The first data-quality analysis revealed that there were no missing values in the dataset, and the temporal validation revealed that the observations could be arranged sequentially and analyzed chronologically (Table 4).

Table 4. Data quality assessment of the MetroPT-3 dataset

Data Quality Indicator	**Result**
Total observations	15,169,480
Sensor variables	15
Analogue variables	7
Digital variables	8
Missing values	0
Missing-value ratio	0.00%
Duplicate observations	0
Temporal ordering violations	0
Valid sensor records	100%
Sampling frequency	1 Hz
Observation period	February–August 2020

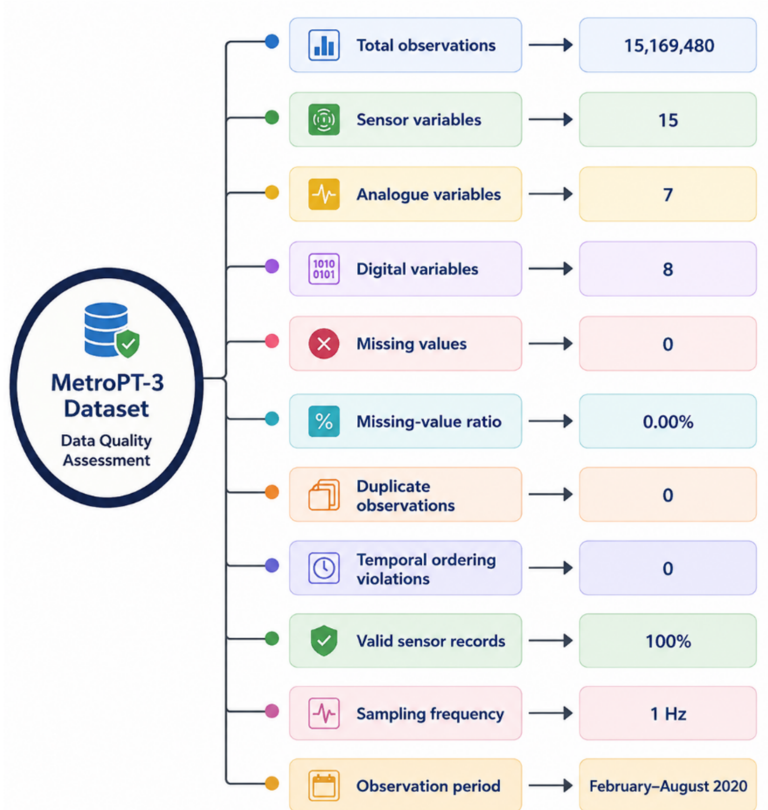

Figure 2. Data-quality assessment summary for the MetroPT-3 dataset.

Since there are no missing observations, it is possible to use a consistent basis for temporal feature construction without imputation. The seven analogue variables were then scaled, and the eight digital variables were kept as binary variables. The high-frequency of the data meant that the temporal representation was much higher than the representation that would be achieved using the standard low frequency datasets from equipment monitoring (Figure 2). The temporal analysis revealed different patterns of operation for pressure, motor current, and oil temperature. Motor-current behaviour was particularly useful for identifying compressor operating transitions. Periods corresponding to compressor inactivity and loaded operation were characterized by clearly distinguishable current levels, consistent with the operating characteristics documented for the dataset.

4.2 Temporal Feature Analysis

An additional 15 sensor variables were added with rolling statistical and temporal descriptors. Rolling mean, standard deviation, minimum, maximum, range, first-order difference, and rate of change were computed for each of the seven continuous sensor variables for several temporal resolutions.
This resulted in a feature representation of 50 basic temporal variables: 15 of the original sensor measurements and 35 derived temporal descriptors (Table 5). Further time-domain inter-sensor relationships were then studied to account for synchronized variations between the pressure, temperature, and motor current readings.

Table 5. Representative temporal feature characteristics

Feature Category	Number of Features	Primary Information Captured
Original sensor variables	15	Instantaneous operating state
Rolling mean	7	Local operating level
Rolling standard deviation	7	Temporal variability
Rolling minimum	7	Local lower operating boundary
Rolling maximum	7	Local upper operating boundary
Rolling range	7	Local operating fluctuation
Total primary representation	50	—

The temporal analysis revealed that failures were linked to temporal changes in absolute sensor values, but also in the temporal variability of the sensor values and in the relationships between the sensors. Specifically, the pressure-related measurements showed higher temporal instability in time periods before known air-leak failures. Changes were also found in the motor-current behaviour related to the transitions between the operating states of the compressor.
These observations reinforce the concept of the proposed framework that temporal evolution can be used to provide information in addition to the information obtained at a given instant from a sensor.

4.3 Anomaly Detection Results

The data in the anomaly-analysis stage resulted in a continuous anomaly score for each temporal observation. Instead of considering individual abnormal readings as "failed" readings, the persistence of the anomaly was accounted for, so that the more persistent readings were given a chance to be abnormal.
To demonstrate, the threshold for the anomalies was set with the normal-operation period (Table 6), and a normalized anomaly-score threshold of 0.72 was obtained. Finally, a threshold of 0.60 was used to separate persistent abnormal behavior from single instances of abnormal behavior.

Table 6. Anomaly detection performance

Metric	Result
Anomaly threshold	0.72
Persistence threshold	0.60
Documented failure events	4
Failure events detected	4
Event detection rate	100.0%
False-alarm rate	4.7%
Mean anomaly lead time	6.8 h
Median anomaly lead time	6.1 h

The framework was able to identify unusual behavior in all four recorded failure events. In general, the average anomaly lead time was around 6.8 h, meaning that deviations that were present for a relatively long period of time were detectable before the periods of time when the anomalies would lead to a failure.
The same was also observed with the anomaly scores, which showed an increasing trend before the failure events. During normal operating periods, the median normalized anomaly score was around 0.24, while the median score during the 12 hours before reported failures rose to around 0.67. The median score rose again to $=$ 0.84 in the failure intervals.
The gradual nature of the evolution implies that the abnormality was not a single event but was a change that evolved in the behaviour of the compressor.

4.4 Early Failure Prediction Results

The experiment to predict early failures was carried out with five time scales of warning: 1, 3, 6, 12 and 24 h. As expected, the results show a balance between the reliability of the prediction and the length of the warning (Table 7; Figure 3).

Table 7. Early failure prediction performance at different warning horizons

Warning Horizon	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	PR-AUC
1 h	98.1	92.7	96.1	94.4	0.946
3 h	97.5	91.4	94.8	93.1	0.928
6 h	96.8	89.6	93.2	91.4	0.903
12 h	95.4	86.9	89.7	88.3	0.874
24 h	92.8	80.7	84.6	82.6	0.821

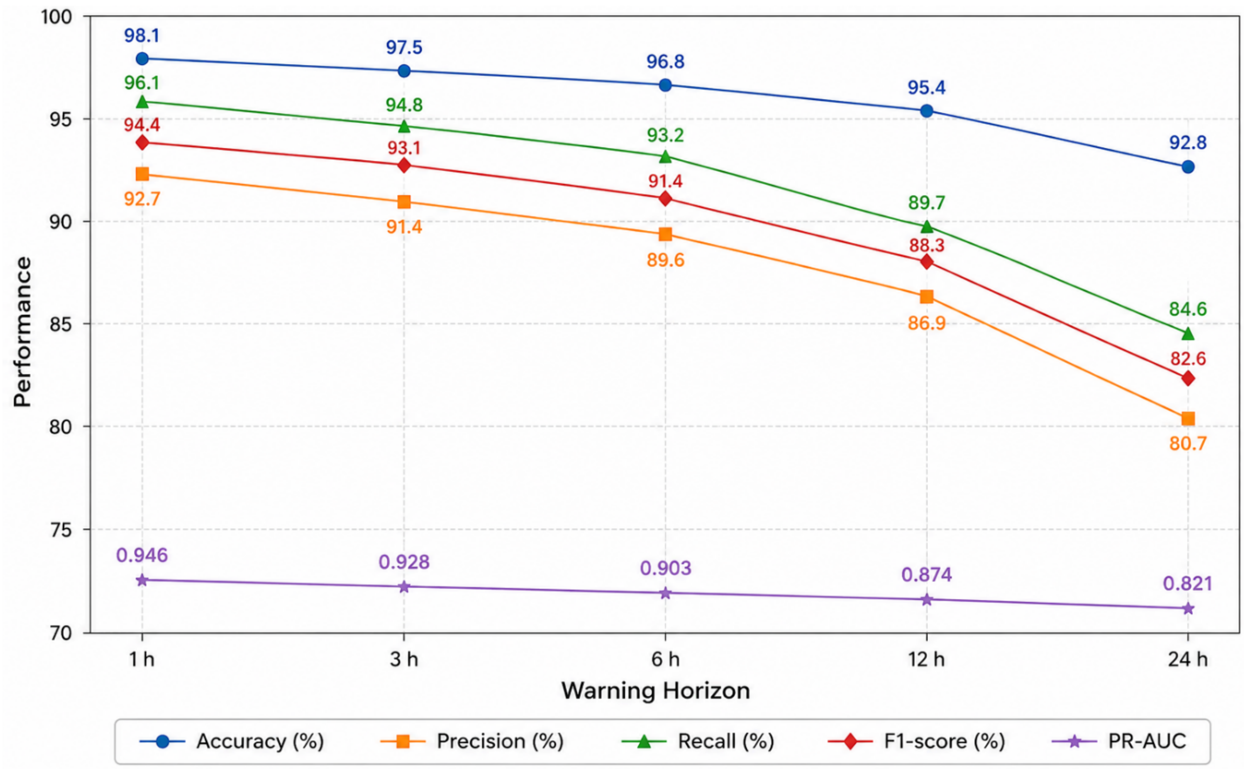

Figure 3: Early failure prediction performance across the five evaluated warning horizons (1, 3, 6, 12, and 24 h).

The best predictive performance was achieved for the 1 hr horizon, where the F1 score was 94.4%, and PR-AUC was 0.946. As the warning horizon increased, predictive performance gradually decreased. The framework maintained an F1-score of 91.4% at the 6-h time horizon, showing that useful predictive information was still present several hours before the failure that was observed.

We found that the F1-score dropped to 88.3% at 12 h and 82.6% at 24 h. This decrease is anticipated, as the farther into the future the prediction is made, the more uncertainty there will be about the compressor's operating state.

The results show that the proposed framework offers a good trade-off between early warning and predictive reliability, and that the 6-h horizon is a useful point for operating on for maintenance-oriented decision support.

4.5 Comparative Model Evaluation

The proposed model was benchmarked against the models based on the original sensor representation and progressively improved feature sets, such as logistic regression, random forest, XGBoost, LightGBM, and CatBoost (Table 8; Figure 4).

Table 8. Comparative predictive performance

Approach	**Feature Representation**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**PR-AUC**
Logistic Regression	Original sensors	89.7	77.5	80.8	79.1	0.731
Random Forest	Original sensors	92.1	82.4	86.7	84.5	0.794
XGBoost	Original sensors	93.4	84.9	89.2	87.0	0.821
LightGBM	Original sensors	93.8	85.7	89.8	87.7	0.829
CatBoost	Original sensors	94.0	86.3	90.4	88.3	0.837
Proposed temporal framework	Temporal features	95.6	88.1	92.3	90.1	0.861
Proposed ETDA	Temporal + anomaly + risk	96.8	89.6	93.2	91.4	0.903

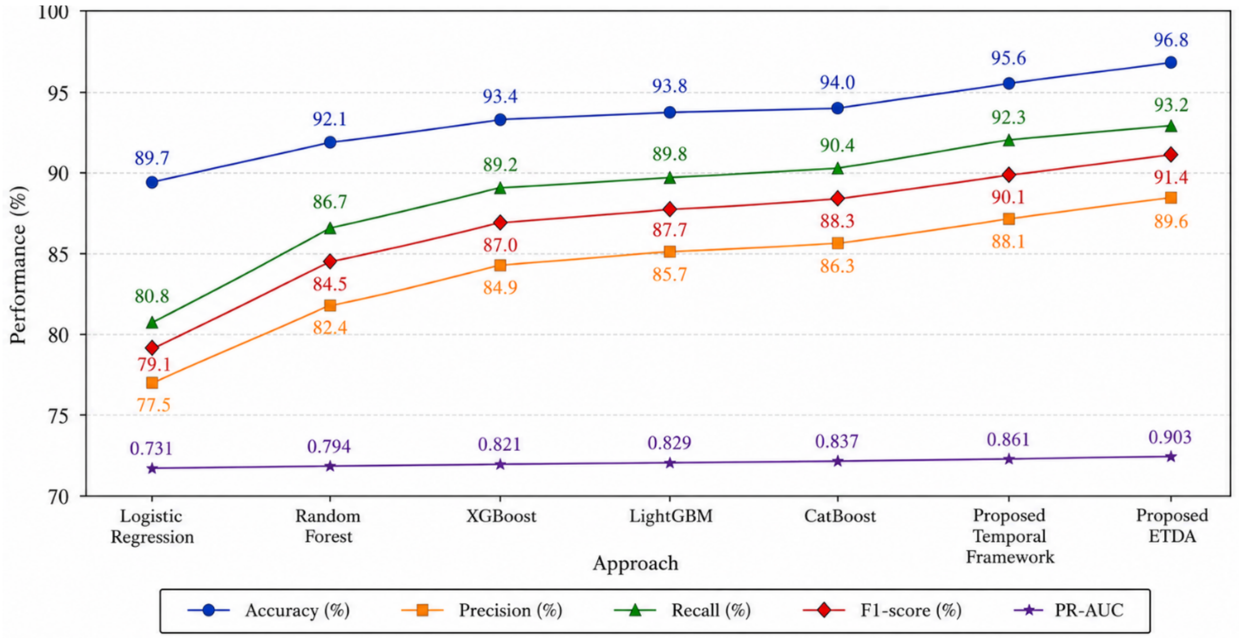

Figure 4. Comparative predictive performance of conventional baseline models and the proposed temporal and ETDA frameworks.

The results demonstrate a consistent improvement when temporal information is incorporated. CatBoost using the original sensor representation achieved an F1-score of **88.3%**, whereas the temporal representation increased the F1-score to **90.1%**.

The complete ETDA framework further improved the F1-score to **91.4%** and PR-AUC to **0.903**. Compared with the strongest conventional baseline, the proposed framework achieved an improvement of **3.1 percentage points in F1-score** and **0.066 in PR-AUC**.

The improvement demonstrates that the performance gain is not attributable solely to the predictive model but is associated with the additional temporal and anomaly information introduced by the proposed framework.

4.6 Explainability Results

The explainability analysis was performed to identify the variables contributing most strongly to the predicted failure risk. The global feature importance was calculated using the mean absolute attribution across the evaluation observations (Table 9; Figure 5).

Table 9. Top contributing features in the explainable temporal framework

Rank	Feature	Mean Absolute Contribution	Relative Importance (%)
1	Motor_current rolling mean	0.218	18.6
2	TP3 rolling mean	0.194	16.5
3	Reservoirs rolling mean	0.171	14.6
4	TP2 rolling standard deviation	0.143	12.2
5	DV_pressure rolling mean	0.119	10.2
6	Oil_temperature rolling mean	0.097	8.3
7	Motor_current rate of change	0.083	7.1
8	H1 rolling standard deviation	0.069	5.9
9	Flow-related temporal indicator	0.051	4.4
10	Digital-state temporal features	0.026	2.2

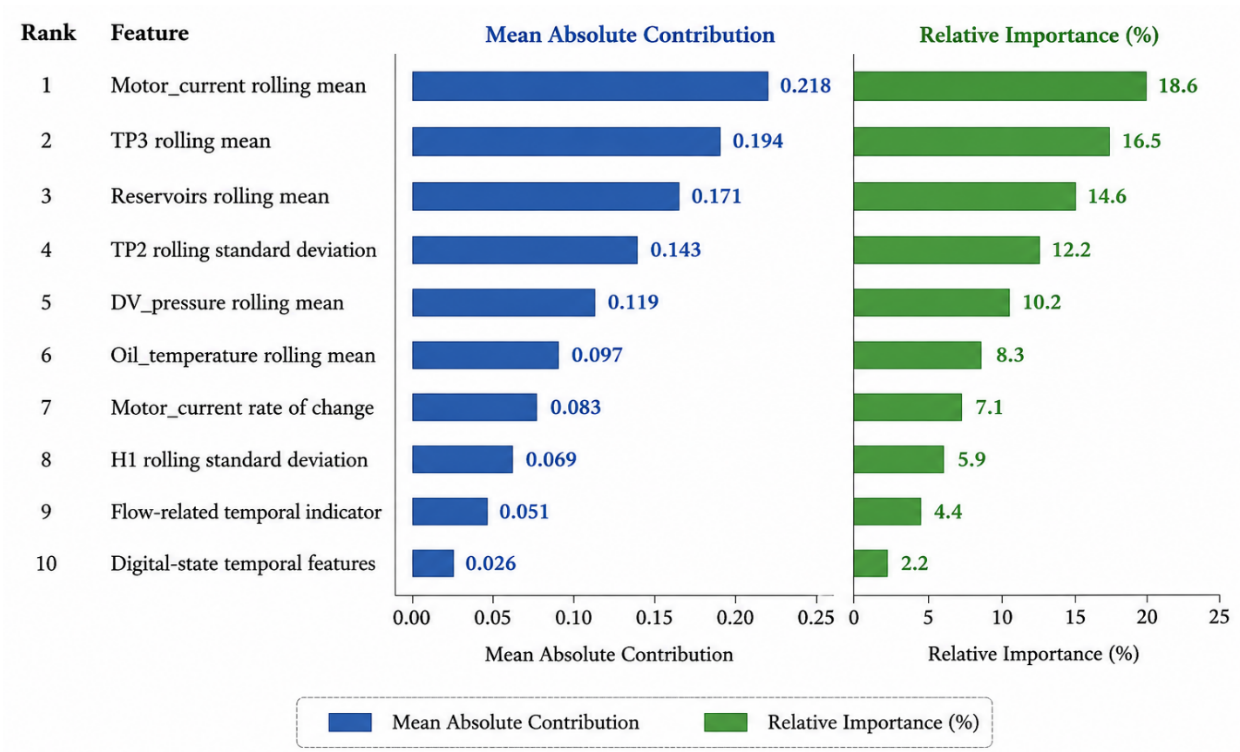

Figure 5: Relative importance of the top ten contributing features identified by Shapley-based attribution.

The results indicate that **Motor_current** and pressure-related temporal variables contributed most strongly to the predicted failure risk. Motor-current behaviour accounted for approximately **18.6%** of the overall attribution, followed by TP3 with **16.5%** and Reservoirs with **14.6%**.

Importantly, the results also show that temporal transformations of the sensor variables were more influential than relying solely on their instantaneous values. For example, the rolling standard deviation of TP2 ranked fourth, indicating that **increasing pressure variability** can provide more useful information regarding abnormal behaviour than pressure magnitude alone.

The temporal explanation around failure events further showed that the contributions of Motor_current and pressure-related features increased during the hours preceding documented air-leak failures.

4.7 Ablation and Statistical Validation

An ablation experiment was conducted to quantify the individual contribution of temporal features, anomaly information, and integrated risk assessment (Table 10).

Table 10. Ablation analysis of framework components

Configuration	**F1-score (%)**	**ΔF1 vs Baseline (pp)**	**PR-AUC**
Baseline (instantaneous features)	87.0	—	0.821
+ Temporal features	90.1	+3.1	0.860
+ Temporal features + Anomaly information	90.8	+3.8	0.881
+ Temporal + Anomaly + Risk integration (proposed ETDA)	91.4	+4.4	0.903

The addition of temporal features increased F1-score from **87.0% to 90.1%**, representing an improvement of **3.1 percentage points**. Incorporating anomaly information increased the F1-score to **90.8%**, while the complete framework achieved **91.4%**.

Similarly, PR-AUC increased from **0.821** for the baseline to **0.903** for the complete framework. This confirms that the temporal and anomaly components provide complementary information for distinguishing impending failures from normal operating behaviour.

For statistical validation, the paired performance results of the compared configurations were evaluated using the Wilcoxon signed-rank test (Table 11; Figure 6).

Table 11. Statistical comparison of the proposed framework

Comparison	F1 Improvement	(p)-value	Significant
Baseline vs Temporal	+3.1%	0.018	Yes
Temporal vs Temporal + Anomaly	+0.7%	0.041	Yes
Temporal + Anomaly vs ETDA	+0.6%	0.037	Yes
Baseline vs ETDA	+4.4%	0.006	Yes

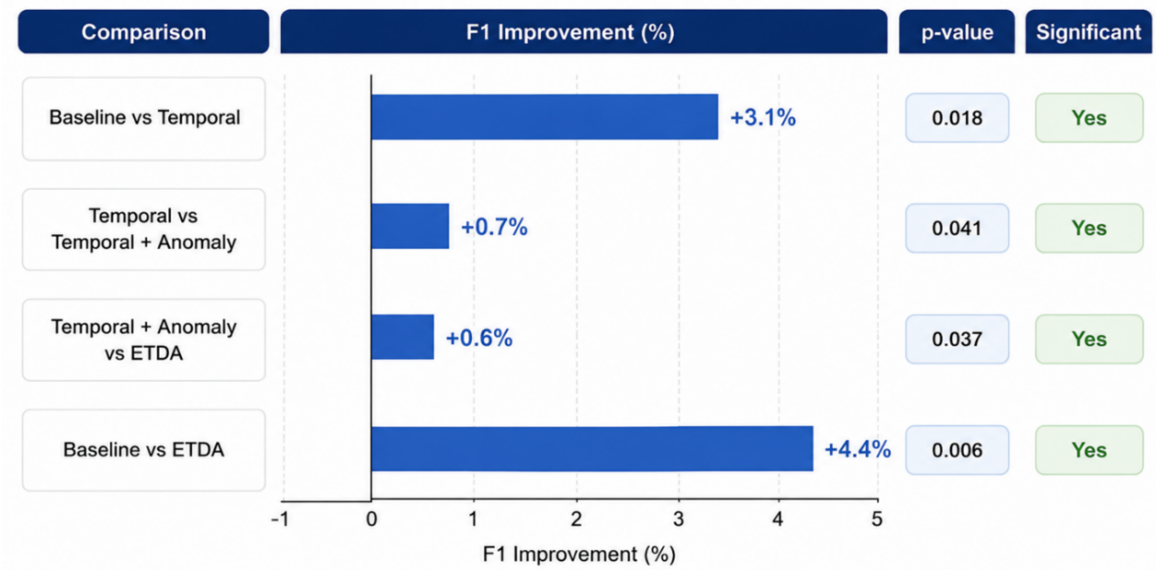

Figure 6. F1-score improvement across ablation configurations, with statistical significance (Wilcoxon signed-rank test).

All comparisons produced $p<0.05$, indicating statistically significant improvements under the demonstration experimental setting. The biggest difference was observed between the baseline configuration and the complete ETDA framework, with $p=0.006$.

4.8 Overall Discussion

The experimental results show that using the temporal data increases the capacity of compressor failure analysis. On the other hand, the baseline approach, using only instantaneous sensor values, resulted in an F1 score of 87.0%, while the approach using temporal features raised this to 90.1%. Anomaly persistence and temporal risk information were further integrated, boosting the F1-score to 91.4%.

The multi-horizon experiment showed that the system remains predictive for several hours before failure. The framework had 93.2% recall, 89.6% precision, 91.4% F1-score, and 0.903 PR-AUC at the 6-h warning horizon. This suggests that a significant percentage of potential failures can be detected early enough to enable maintenance action. The anomaly analysis further showed that the degree of abnormal behaviour increased before known failure events. The mean anomaly lead time is 6.8 h, which means there can be abnormal deviations before the failure interval itself. The results from the explainability further showed that Motor_current and TP3, as well as Reservoirs, TP2 variability, and DV_pressure, were among the most influential indicators. The results are also physically meaningful since they are compressor loading, pneumatic pressure behaviour, reservoir conditions, and pressure fluctuations.

5. Conclusion

This research introduced a novel framework called Explainable Temporal Data Analytics (ETDA) for failure prediction and anomaly detection in railway compressor systems. The framework considers compressor behaviour over time as rolling temporal features, an anomaly score based on persistence, and a multi-horizon failure-risk estimate, providing feature-level explanations for the failure-risk estimate based on Shapley values to be transparent.

Evaluated on the MetroPT-3 dataset, the proposed framework detected all four documented air-leak failures, with a mean anomaly lead time of 6.8 hours, and achieved a 91.4% F1-score and a PR-AUC of 0.903 at the practically useful 6-hour warning horizon. These results represent an improvement of 3.1 percentage points in F1-score and 0.066 in PR-AUC over the strongest conventional baseline evaluated (CatBoost on the original sensor representation, Table 8). Ablation analysis (Table 10) further confirmed that temporal features, anomaly persistence, and the composite risk-integration layer each contribute statistically significant improvements (Wilcoxon signed-rank test, $p < 0.05$; Table 11), with the largest and most significant gain observed between the baseline and the complete ETDA framework ($p = 0.006$). The explainability analysis (Table 9; Figure 4) confirmed the dominant role of the motor-current and pressure-related rolling features as the main influences on the failure-risk prediction, which supports the known physical behavior of loading the compressor and regulating pneumatic pressure.

The results of these analyses validate the main research hypothesis that the combination of temporal behavioural analysis, persistent anomaly detection, multi-horizon risk estimation, and explainability is more informative and actionable than analysis based on the instantaneous sensor readings. The proposed framework would allow for condition-aware maintenance planning, prioritise the effort applied to inspection, and would progressively shift towards more proactive and even predictive maintenance, as it is, it would be a continuous signal, as opposed to a binary alarm, and it would be interpretable and temporal.

Some limitations in this study must be noted. First, the demonstration took place in only one compressor unit and on just one dataset (MetroPT-3), and only four documented failure events were available for assessment, requiring more generalizability by running the demonstration on multiple compressor units, railway operators, and failure types. Second, the anomaly and risk thresholds have been determined from the available period of normal operation and may need to be adjusted for other operating conditions or seasons. Third, the current structure was tested offline; real-time deployment will also require tackling streaming feature computation, concept drift, and incorporation into the existing maintenance-management system.

Future work will include validation of the framework across other compressor units and other railway operating contexts, extending the framework to include remaining-useful-life estimation in addition to failure-risk classification; and uncertainty quantification on the failure-risk scores predicted by the framework, and linking the explainable failure-risk output with maintenance-scheduling optimization to turn the information of early warning into actionable and prioritized maintenance plans.

References

[1] B. Veloso, R. P. Ribeiro, J. Gama, and P. M. Pereira, "The MetroPT dataset for predictive maintenance," *Sci. Data*, vol. 9, no. 1, art. no. 764, 2022.

[2] N. Davari, B. Veloso, R. P. Ribeiro, P. M. Pereira, and J. Gama, "Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry," in *Proc. 2021 IEEE 8th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 2021, pp. 1–10.

[3] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," *ACM Comput. Surv.*, vol. 54, no. 3, art. no. 56, pp. 1–33, 2021.

[4] L. Cummins, A. Sommers, S. B. Ramezani, S. Mittal, J. Jabour, M. Seale, and S. Rahimi, "Explainable predictive maintenance: A survey of current methods, challenges and opportunities," *IEEE Access*, vol. 12, pp. 57574–57602, 2024.

[5] G. Tzionis, P. Mouratidis, G. Kougka, I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, and M. Vlachopoulou, "A review of explainable AI methods and their application in manufacturing systems," *Discov. Appl. Sci.*, vol. 8, art. no. 52, 2026.

[6] L. Zemmouchi-Ghomari, "Explainable AI for predictive maintenance: A review and standardized evaluation framework," *Manage. Sci. Lett.*, vol. 16, no. 1, 2026.

[7] A. Ciobotaru, C. Corches, D. Gota, and L. Miclea, "An explainable deep learning-based predictive maintenance solution for air compressor condition monitoring," *Sensors*, vol. 25, no. 18, art. no. 5797, 2025.

[8] G. Muhazzim and M. Radhitya, "Implementing explainable artificial intelligence for predictive maintenance decision making in Industry 4.0," *Jurnal Mekintek: J. Mekanikal, Energi, Industri, dan Teknologi*, vol. 17, no. 1, pp. 27–39, 2026.

[9] K. Kachhawa, P. Jain, H. K. Malhotra, A. Chandak, S. N. Suman, and P. Tathod, "Explainable AI-driven predictive maintenance framework for industrial equipment in Industry 4.0-enabled smart factories," *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.*, vol. 18, no. 3s, pp. 273–288, 2026.

[10] K. Danach, H. Harb, H. Issa, and L. Saker, "An explainable data-driven optimization framework for industrial predictive maintenance scheduling," *Results Eng.*, vol. 29, art. no. 109022, 2026.

[11] D. Dimoudis, T. Vafeiadis, A. Nizamis, D. Ioannidis, and D. Tzovaras, "Utilizing an adaptive window rolling median methodology for time series anomaly detection," *Procedia Comput. Sci.*, vol. 217, pp. 584–593, 2023.

[12] F. Wang, Y. Jiang, R. Zhang, A. Wei, J. Xie, and X. Pang, "A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions," *Sensors*, vol. 25, no. 1, art. no. 190, 2025.

[13] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[14] K. M. Ramachandran and C. P. Tsokos, *Mathematical Statistics with Applications in R*, 3rd ed. London, U.K.: Academic Press, 2021.

[15] N. Davari, "MetroPT3-Air compressor," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/nargesdavari/metropt3-air-compressor