

A 3-level interpretable data-driven framework for operational analysis of industrial centrifugal pumps

Alessia D'Ambrosio · Antonio D'Ambrosio · Roberta Siciliano · Angelo Martone · Michele Ferrucci · Gianpaolo Romano · Gaetano Zazzaro 

Received: 20 May 2026 / Accepted: 9 September 2026

© The Author(s) 2026

Abstract

This paper proposes an interpretable framework for industrial condition monitoring, enabling a three-level understanding of machinery behaviour (machine, context, and system). Using multivariate time-series from a WirelessHART sensor network, the framework combines transparent data-mining and statistical-learning components (density-based clustering, principal component analysis, variable-importance modelling, and time-series dependency analysis) to extract readable patterns and rules describing operational dynamics. Unlike approaches that treat machine, context, or system levels in isolation, the proposed workflow unifies them within a single analytical scheme built around intrinsically interpretable modules. At the machine level, DBSCAN and PCA identify and characterise distinct operational states and enable compact decision rules; at the context level, the Kolmogorov-Smirnov test and a consensus ranking of variable importance quantify campaign-to-campaign drift and highlight the role of environmental variables; at the system level, Granger test and cross-correlation analysis identify short-lag directional predictive relationships between concurrently operating machines. The framework is demonstrated on three centrifugal pumps monitored across multiple campaigns in an industrial plant. All data are publicly available, enabling full replication.

Keywords Condition monitoring · Interaction analysis · Multivariate time series · Sensor data · Variable importance

Introduction

Nowadays, monitoring industrial plants, whether simple or highly complex, equipped with dense networks of smart sensors, industrial IoT devices, and advanced control systems, continuously produces large volumes of multivariate time-series data¹. These data streams capture rich information about the operating conditions of machinery, useful for understanding their operational status and health. Effectively exploiting this information requires analytical frameworks that can structure, organise, and interpret heterogeneous signals, rather than focusing on isolated variables or single operating points. A substantial body of research has investigated data-driven analysis of industrial machinery, particularly in the context of condition monitoring and predictive maintenance².

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



In this context, several studies have highlighted that data-driven fault detection on industrial sensor streams is not sufficient per se, and that interpretability and explanation of abnormal behaviours are required to effectively support diagnostic reasoning and maintenance decisions³. Recent surveys on multi-fault diagnosis for industrial rotating machines explicitly report model interpretability as a key barrier to credible deployment in Industry 4.0 environments, motivating interpretable-by-design monitoring pipelines rather than purely black-box predictors⁴.

Within this literature, most contributions address machine-level analysis of rotating equipment such as pumps, bearings, and turbines, using clustering, dimensionality-reduction methods, and supervised or unsupervised classifiers for anomaly detection, fault diagnosis, and health assessment based on vibration, temperature, and pressure data signals^{5–7}. Comprehensive reviews have emphasized that such models are increasingly integrated into broader industrial analytics architectures supporting process optimization, fault diagnosis, and asset management, while also pointing to emerging trends in hybrid physics-informed and explainable approaches⁸.

Environmental conditions can modify measurable responses in rotating electrical equipment. Evidence from electrical-machine studies shows that temperature, pressure and humidity influence key signal characteristics⁹, indicating that ambient factors can drive systematic shifts in observed data and motivating a dedicated context-level analysis.

Beyond single-component analysis, several studies propose plant-wide or system-level frameworks that model interactions among machines and subsystems through multivariate time-series models, graph or network representations, and causal or correlation-based analyses^{10–12}. Collectively, these works show that the behaviour of industrial assets is governed not only by local operating conditions, but also by shared loads, control-loop feedback, and inter-subsystem dependencies. A key contribution of such approaches is the ability to characterise how operational disturbances propagate across shared infrastructures, such as hydraulic circuits or electrical distribution networks, highlighting that the state of a given unit depends on its functional coupling with neighbouring equipment and on system-level operating conditions, rather than on internal dynamics alone.

In parallel, the rising complexity and multiscale structure of industrial systems has increased the demand for methods that provide both accurate predictions and interpretable insights. This has led to the development of inherently interpretable models, such as fuzzy networks, decision trees, linear models, and causal analysis techniques^{13,14}. Complementarily, hybrid approaches that integrate data-driven learning with physically-motivated or domain-informed components have been proposed to enhance model transparency and generalization, producing models that can both describe and justify operational behaviours¹⁵. Together, these developments aim to produce human-understandable representations of operational regimes, key influencing factors, and dominant interaction mechanisms, improving communication among data scientists, engineers, and plant operators and supporting reliable decision-making in industrial settings.

Taken together, these studies show that operational analysis of industrial machinery benefits from complementary perspectives on sensor data, including local machine behaviour, acquisition-dependent variability, and temporal dependencies among interconnected units. Building on this perspective, this paper proposes an interpretable modular framework for the descriptive analysis of industrial centrifugal pumps. The framework is organised into three analytical levels—machine, context, and system, each producing directly inspectable evidence on operational states, cross-campaign distributional drift, and short-lag lead-lag relationships. The approach is applied to multivariate sensor data collected from three centrifugal pumps over three acquisition campaigns.

Related works and proposed framework

The literature on data-driven analysis of industrial systems spans a wide range of approaches, including signal processing, interpretable machine learning, and causal analysis. In this study, we adopt a conceptual reading of existing contributions based on three complementary levels: (i) *machine-level*, oriented toward the characterization of the local behaviour of individual equipment; (ii) *context-level*, focused on the effects of environmental

dynamics on operating conditions; and (iii) *system-level*, which analyses relationships, dependencies, and propagations of phenomena across machines and subsystems.

Table 1 summarises the representative works, selected to cover a significant range of methodologies (clustering, fuzzy networks, process mining, deep learning with attention mechanisms, causal analysis) and to illustrate the diversity of analytical levels considered in recent studies on industrial condition monitoring.

This overview shows that existing contributions cover only parts of the analytical spectrum needed to understand complex industrial systems: machine-level analyses dominate, context-level aspects are rarely formalised, and system-level dynamics appear only in specialised studies. This fragmentation limits coherent interpretations of operations under interdependencies, non-stationary conditions, and heterogeneous signals.

No published study, to our knowledge, jointly addresses the machine, context and system levels on a public real-plant centrifugal pump dataset using intrinsically interpretable methods.

The proposed framework

To address the fragmentation highlighted in previous works, this paper proposes a modular, interpretable analytical framework integrating statistical learning and data-mining techniques. The framework adopts interpretability as an intrinsic design principle²³, ensuring that the analytical logic and intermediate results remain transparent throughout the entire pipeline. In particular, interpretability is understood here as the property of models whose internal structure and parameters are directly comprehensible (*intrinsic interpretability*), in contrast to explainable machine learning, which typically provides *post hoc* explanations for black-box predictors rather than employing models that are interpretable by construction.

The proposed solution operationalises the three levels outlined in the literature by organising the framework into three analytical modules (summarised schematically in Fig. 1), each corresponding to a distinct engineering

Table 1 Representative data-driven approaches for industrial condition monitoring and operational understanding, grouped by analytical level.

Analytical level(s)	Analytical techniques	Objective/application domain	Multi-level coverage	References
Machine	Evolving fuzzy neural networks (interpretable)	Adaptive structure learning for industrial sensor-based classification tasks	✗	13
	Time-aware clustering (DBSCAN + temporal segmentation)	Sensor data segmentation and machine status monitoring in industrial equipment	✗	16
	Unsupervised ML (PCA + clustering) for sensor data preprocessing	Feature extraction and data structuring from machinery sensor signals in manufacturing case-study	✗	17
Machine, System	Process-mining + interpretable ML on sensor/event logs	Causality modelling of sensor data in a pulp-mill industrial system	✓	18
	Fuzzy-logic-based integration of diagnostic indicators	Fault diagnosis and condition monitoring of mechanical components in Industry 4.0 manufacturing processes	✗	19
	CNN with attention for causal inference on SCADA sensor data	Revealing causal relationships in wind-turbine sensor networks	✗	20
	Multi-level attention GRU (seq2graph) on multivariate sensor time-series	Discovery of dynamic dependencies in industrial machine sensor data	✓	21
Machine, Context	Incremental clustering	Streaming sensor-network analysis for real-time fault/novelty detection in industrial assets	✗	22
Machine, Context, System	Clustering, PCA, KS-test, variable importance, Granger test on pump sensor data	Explainable multi-level analysis of centrifugal pump operations using sensor networks	✓ modular	This work

The column Multi-level coverage marks with ✓ a level addressed by the corresponding work and with ✗ a level not considered. For the present paper, the three levels are addressed by three analytical modules, each applicable on its own to a campaign meeting the corresponding data requirements (modular).

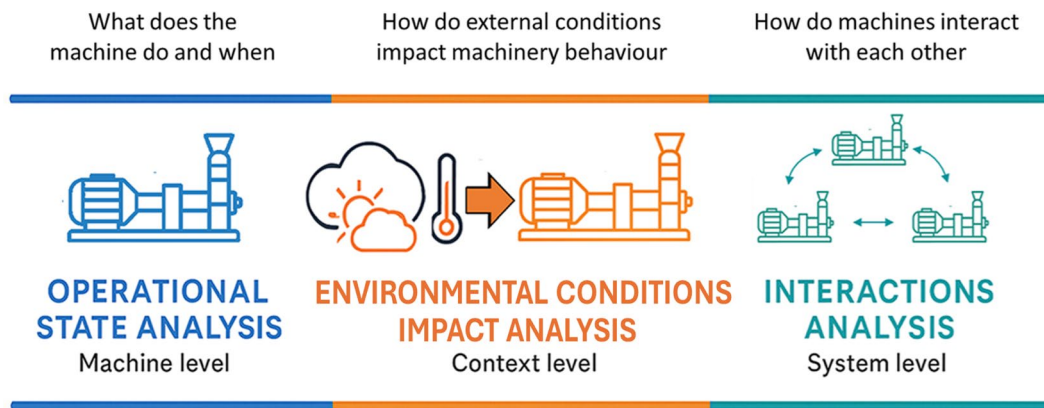


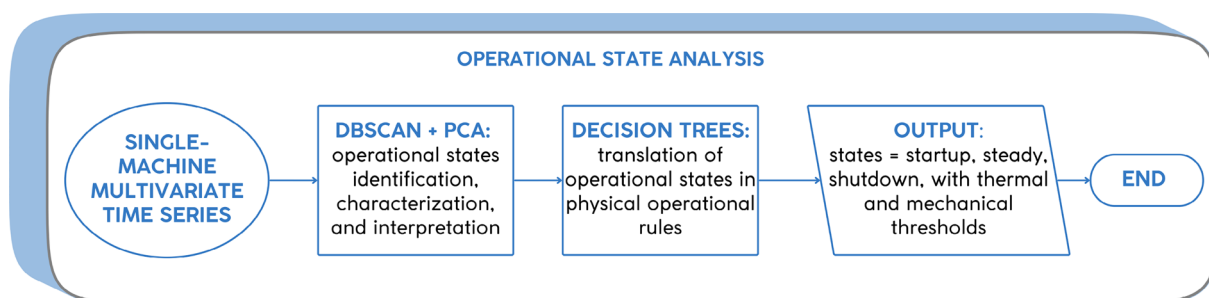
Fig. 1 Schematic overview of the three analytical modules composing the framework (operational state, environmental conditions impact, and interactions) pursued at a specific engineering level (machine, context, or system).

interpretation level. At each module, the analysis is implemented as a structured pipeline of statistical learning and data-mining techniques that integrates modelling algorithms with their interpretation, with outputs such as low-dimensional projections, if-then rules, ranked variable lists and predictive relations that are directly inspectable and technically meaningful for maintenance engineers and domain experts. From an engineering standpoint, each level has the same general goal of optimising machinery monitoring and maintenance operations. The analytical modules with their corresponding engineering use, are the followings:

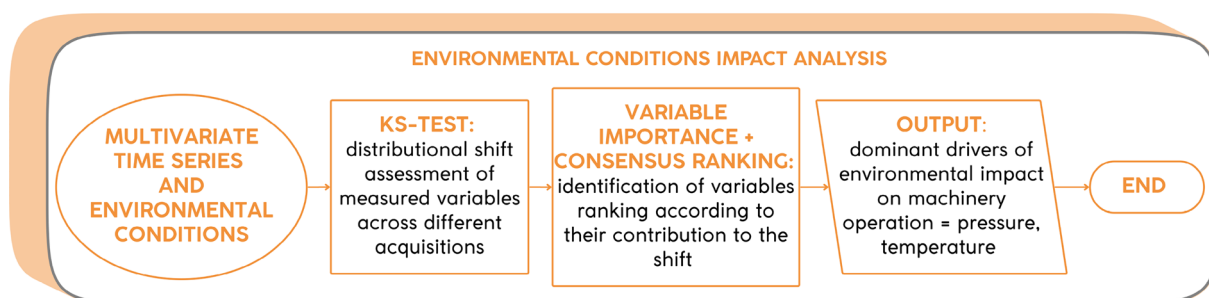
1. *Operational state analysis (machine level)* Identification and characterisation of machinery operating states and recurrent patterns, with a focus on local dynamics (Fig. 2a). The output of this module provides a data-driven baseline of normal operating behaviour against which future deviations can be flagged to plant engineers.
2. *Environmental conditions impact analysis (context level)* Assessment of how operating conditions and environmental factors influence machine behaviour, including seasonal and long-term variations (Fig. 2b). The output of this module can support the decision of whether to adapt the maintenance frequency to the season, for instance by introducing additional inspection shifts during the warmer months.
3. *Interaction analysis (system level)* Examination of machine-to-machine short-lag directional predictive relationships (Fig. 2c). The output of this module reveals how coupled machines actually interact, i.e. which sensed quantities propagate from one unit to another, information that was not available beforehand; notably, no vibration transmission is observed because the foundations are rigid, so the coupling occurs through the shared hydraulic and thermal paths rather than mechanically.

Contributions

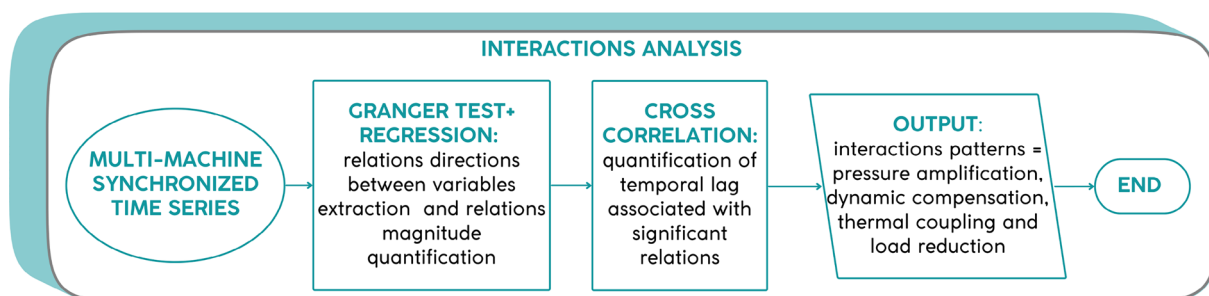
The paper makes three practical contributions: (i) an interpretable, modular three-level pipeline consisting of machine-, context-, and system-level evidence; (ii) a fully reproducible case study on a public industrial dataset²⁴, including campaign metadata and preprocessing/parameter settings; and (iii) a set of engineering-facing interpretable outputs (operational-state clusters, low-dimensional projections and loadings, compact decision rules, consensus importance rankings, and an interaction taxonomy) that support diagnosis beyond single-machine fault detection. Applied to centrifugal pump sensor data from a real heating plant, the three modules produce a coherent description of operating regimes, between-campaign distributional contrasts, and between-pump short-lag directional predictive relationships, each reported in a form that admits direct inspection by a plant engineer.



(a) Operational state analysis



(b) Environmental conditions impact analysis



(c) Interactions analysis

Fig. 2 Flowchart describing the structured pipeline of analytical tools used for: **(a)** operational state analysis, **(b)** environmental conditions impact analysis, **(c)** interactions analysis.

Paper structure

The remainder of this paper is organised as follows. First, the industrial dataset, acquisition setup, data characteristics, and preprocessing steps are described. The proposed analytical framework and the corresponding results are then presented. Finally, the implications of the findings and possible future directions are discussed.

Industrial data: source, understanding and pre-processing

The analytical modules of the framework were applied using real-world data collected from an operational industrial plant, where a dedicated sensor system was deployed to monitor machinery operations under actual working conditions.

Target plant and data acquisition system

The target plant is made up of a group of three industrial centrifugal pumps, that feed the boilers with demineralized water of a heating plant used for the steam production necessary for the CIRA (Italian Aerospace Research Centre)²⁵ (<https://www.cira.it/en>) research facilities. Each centrifugal pump, named A, B, and C, consists of a coupling of electric motor and corresponding mechanical pump, which transfer the fluid by increasing its pressure. Each centrifugal pump is, in turn, instrumented with four sensors: one accelerometer on the pump and one on the motor (both of which additionally incorporate a temperature sensing element for contact temperature measurement), a temperature sensor on the motor casing, and a pressure sensor to measure the flow rate of the outgoing fluid. The choice of sensing modalities follows established practice for the condition monitoring of rotating machinery^{26,27}: the international standard ISO 20816-7:2018 prescribes the measurement and evaluation of mechanical vibration for rotodynamic pumps, which motivates the accelerometers mounted on both motor and pump and supports the detection of the most common mechanical failure mechanisms, whose signatures map onto specific signals, namely unbalance, misalignment and bearing wear (vibration velocity and peak acceleration) and cavitation (vibration together with outlet-pressure fluctuations), while overheating and thermal degradation are captured by the contact- and casing-temperature channels. The two external sensors, although outside the scope of the standard, were added for research purposes. In addition to the twelve sensors installed on the centrifugal pumps, two external sensors are included to measure environmental temperature and pressure. All sensors are interconnected to form a wireless mesh network, based on an industrial standard protocol (WirelessHART), which communicates with a Wi-Fi Gateway. This gateway collects and subsequently stores sensor data in a dedicated Data Center, as shown in Fig. 3. Table 2 summarizes the sensor types, the measured physical quantities and the corresponding units of measurements. All sensor data are timestamped (in UTC) with precision to the second; given this acquisition rate and the limited frequency bandwidth of the sensors themselves, content above their cut-off frequency is not captured, so the time- and frequency-domain features typically extracted from 1D signals^{28,29} are not available and the framework operates directly on the acquired readings.

Data understanding and pre-processing

The data analysed in this study were collected during scheduled industrial monitoring campaigns, each defined by operational needs. In this regard, the public release of real-plant datasets supports reproducibility and enables fair benchmarking, addressing a frequently reported gap between laboratory benchmarks and industrial data constraints³⁰. As a result, the acquisition windows differ in terms of duration, sampling density, and environmental conditions, providing natural variability for multi-level analysis. Table 3 provides a detailed overview of the data acquisition campaigns for each centrifugal pump, organized by operational date. For each session, the corresponding analytical module addressed in the study is also reported. Additional details on the data collection setup and campaign specifications can be found in the original data description³¹. Notably, pump C was turned off on the third campaign day (30/10/2024). The empirical analyses are therefore organised according to the acquisition structure reported in Table 3. Each module is applied to the data window that provides the required observational conditions, rather than being repeated mechanically across all campaigns. This choice keeps the analysis focused on the informative portions of the dataset and avoids redundant applications of the same procedures. Before modelling, a descriptive analysis was performed to characterize the raw sensor data collected from centrifugal pumps A, B and C, across the multiple operating sessions. Although similar descriptive analyses were conducted for the

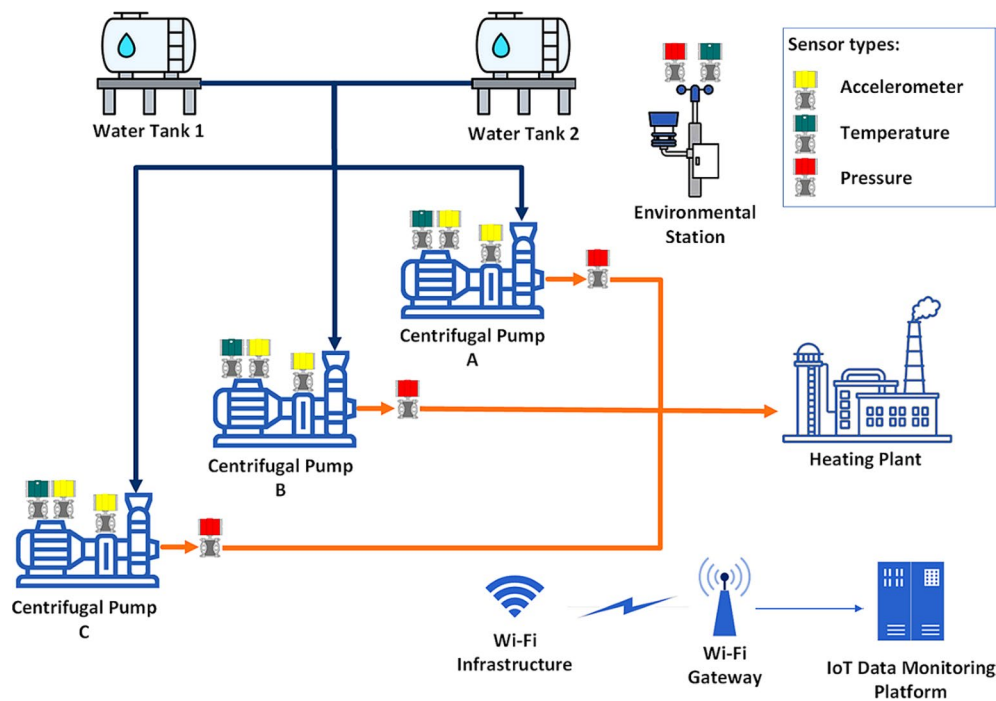


Fig. 3 Target plant and data acquisition system scheme.

Table 2 Description of the sensor data used in the analysis.

Sensor type	Variable label	Measured physical quantity	Unit
Accelerometer	$X_ACR_Mot.PV$	Vibrational velocity measured by the motor accelerometer	m/s
	$X_ACR_Mot.SV$	Peak value measured by the motor accelerometer	m/s ²
	$X_ACR_Mot.TV$	Contact temperature of the motor accelerometer	°C
Accelerometer	$X_ACR_Pmp.PV$	Vibrational velocity measured by the pump accelerometer	m/s
	$X_ACR_Pmp.SV$	Peak value measured by the pump accelerometer	m/s ²
	$X_ACR_Pmp.TV$	Contact temperature of the pump accelerometer	°C
Temperature	$X_Temp.PV$	Motor casing temperature	°C
Pressure	$X_Pres.PV$	Outlet fluid pressure from the pump	bar
Environmental	Barometer	Atmospheric pressure	mbar
	Temperature	Environmental temperature	°C

The variable prefix X indicates the centrifugal pump unit label, which can be A, B, or C.

Table 3 Data acquisition campaigns for centrifugal pumps A, B, and C.

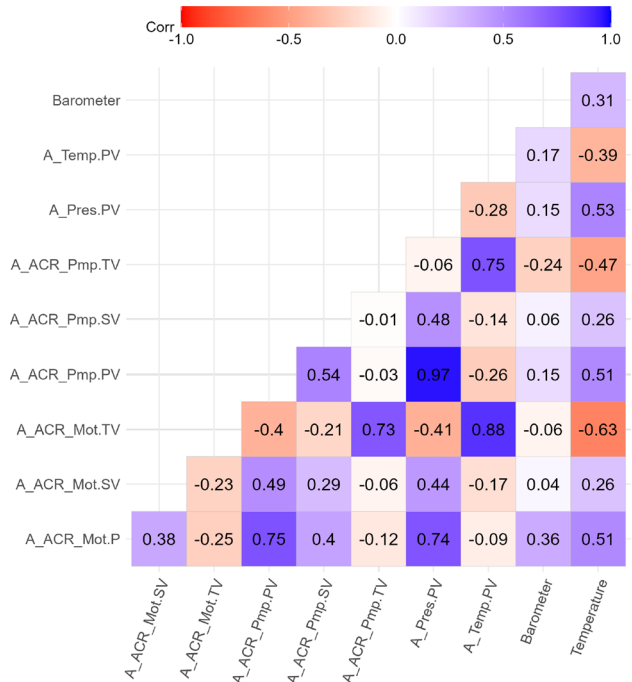
Centrifugal Pump unit	Operational date	Campaign duration	Analytical module
A	10/04/2024	00:20:58	Operational state analysis
B		00:04:59	
C		00:11:58	
A	11/06/2024	01:20:59	Environmental conditions impact analysis
B		05:59:25	
C		00:26:59	
A	30/10/2024	02:19:58	Interactions analysis
B		02:37:23	

Each acquisition date was associated with a different analytical module within the proposed framework. Data for centrifugal pump C are not available on 30/10/2024, as the pump was turned off.

Table 4 Missing values presence for variables of centrifugal pump A during the first campaign.

Variable	NAs	Variable	NAs
A_ACR_Mot.PV	13	A_ACR_Pmp.TV	33
A_ACR_Mot.SV	13	A_Pres.PV	18
A_ACR_Mot.TV	13	A_Temp.PV	75
A_ACR_Pmp.PV	33	Barometer	15
A_ACR_Pmp.SV	33	Temperature	14

Fig. 4 Pearson correlation matrix of the sensor variables recorded from centrifugal pump A during the March-April 2024 acquisition campaign.



remaining centrifugal pumps and campaigns, the resulting patterns were largely consistent with those observed in the representative case. For this reason, only an illustrative example is presented here, namely centrifugal pump A during the first acquisition campaign, as additional material would be redundant. The sensor variables exhibited heterogeneous statistical properties: temperature and pressure signals followed smooth trends, whereas displacement and acceleration showed higher variability and frequent peaks (Supplementary Fig. S1 online). Missing data accounted for 2% of total observations (Table 4). Temporal trends and heatmap of correlation matrices revealed consistent associations among groups of variables (thermal and mechanical), suggesting the presence of latent dependencies to be further investigated in subsequent phases of the analysis (Fig. 4). Following the descriptive analysis, all the data used for Operational state and Environmental conditions impact analyses were subjected to a standardized preprocessing pipeline: missing values were imputed via linear interpolation, and a third-degree Savitzky-Golay filter with an eleven-point window was applied to reduce noise while preserving signal features such as peaks^{32,33}. Finally, Z-score standardization was performed to homogenize variable magnitude prior to multivariate analyses. Instead, the data used for the Interactions analysis, after the missing values imputation by linear interpolation, each series was detrended using a Hodrick-Prescott filter ($\lambda = 1600$)³⁴, and the resulting cyclical component was subjected to successive Augmented Dickey-Fuller and KPSS tests. Whenever a series failed the stationarity criteria (ADF p -value ≥ 0.05 or KPSS p -value ≤ 0.05), it was differenced once and retested; this loop continued up to two differences until both tests confirmed level stationarity, without being white-noise. The final stationary series were then merged by timestamp and standardized via Z-score scaling.

Methods

This section describes the analytical methods composing each module of the framework; the corresponding empirical findings are reported in the subsequent section.

Operational state analysis

The machine-level understanding can be addressed combining different methods. Clustering methods may be employed to detect natural groupings of operational profiles within the monitored data space. To enhance dimensional interpretability, facilitate visualization and extract operational rules associated with the identified patterns, techniques such as Principal Component Analysis and Decision Trees are used.

Density-based spatial clustering of applications with noise algorithm (DBSCAN)

DBSCAN³⁵ is a clustering algorithm designed to identify dense groups of data within a multidimensional space. It relies on two key parameters: Epsilon (ϵ), which defines the maximum distance between two points for them to be considered part of the same cluster, and MinPts, the minimum number of points needed in an area to form a cluster. DBSCAN is applied here under an independence assumption on the standardised, denoised sensor vectors: each observation is treated as an independent point $\mathbb{R}^{10}$, ignoring its temporal order during cluster formation. This is a deliberate modelling choice that prioritises shape-of-density detection over temporal contiguity; the resulting cluster labels are then re-ordered by timestamp and inspected ex-post to verify that the identified regimes are temporally sustained (Fig. 6). In this study, DBSCAN was chosen because data from centrifugal pumps are highly non-linear and may contain irregular regimes with different distributions. Unlike partition-based models, DBSCAN does not require the number of clusters to be specified a priori and can naturally handle noise or anomalous measurements. This property makes it particularly suitable for identifying recurring operational states under realistic industrial variability³⁶.

Principal component analysis (PCA)

For interpretative purposes, the clusters obtained from DBSCAN were projected into a reduced space using PCA³⁷. PCA was selected to summarize the multivariate structure of the signals and to visualize relationships among variables driving the identified operational states, and was preferred over non-linear embeddings such as t-SNE or autoencoders because its linear loadings remain directly interpretable in terms of the original sensor variables. This method is particularly suitable for industrial sensor data, where strong correlations among vibration, temperature, and pressure measurements can obscure the main sources of variability, as also highlighted through the correlation matrix displayed in the previous sections.

Decision trees (DT)

To extract explicit operating rules from the clusters identified by DBSCAN, a decision-tree classifier was trained for each centrifugal pump. DT³⁷ were selected for their ability to translate complex, multivariate sensor patterns into transparent, rule-based thresholds, directly interpretable by maintenance engineers; compared with post-hoc rule extraction from black-box classifiers, decision trees yield rules that are interpretable by construction. The dependent variable was the cluster label assigned to each observation, while all sensor variables were used as predictors.

Environmental conditions impact analysis

The context-level impact can be addressed through a two-stage analytical strategy. First, a statistical comparison test is used to detect significant differences in behavioural patterns across intervals. Once these differences are identified, they can then be further explored through variable importance analysis. A range of feature selection models³⁷ can be employed for this purpose, including regularized regression techniques (e.g., Lasso, Ridge), as well as tree-based methods such as Decision Trees and Random Forests.

Kolmogorov–Smirnov test (KS-test)

The KS-Test is a non-parametric and distribution free method for assessing whether two samples originate from the same distribution by comparing their empirical distribution functions (EDFs)³⁸. It is particularly suitable for industrial data, where the underlying distributions of sensor measurements are unknown or non-Gaussian, and parametric assumptions may not hold³⁹. In this study, the KS-Test was applied to compare the distributions of the monitored signals between two acquisition campaigns characterised by different ambient temperature and pressure levels, in order to identify which signals differ between the two periods. Between the first and second monitoring campaigns, all three centrifugal pumps exhibited an average external temperature increase of approximately 8 °C, accompanied by a corresponding shift in barometric pressure, providing a natural basis for comparative analysis.

Variable importance estimation and consensus ranking

After detecting distributional shifts, to investigate the influence of external conditions on machine behaviour, four complementary models were used to estimate variable importance: Lasso and Ridge Regression, Decision Trees (DT), and Random Forests (RF)^{37,40,41}. In all these models, the response variable was the indicator of the monitoring campaign, so that feature importance quantifies how well each predictor discriminates between the two environmental regimes. Lasso and Ridge are regularized linear models, with Lasso applying an ℓ_1 penalty that promotes sparse solutions and Ridge using an ℓ_2 penalty that distributes shrinkage more uniformly across correlated features. DT and RF, in contrast, are tree-based algorithms capable of capturing non-linear relations; their variable importance is derived from the total impurity reduction achieved by each feature across all splits. By combining linear and non-linear estimators, the analysis captures effects of environmental variables on machine response. Each model produced a ranked list of variables based on relative importance, and to integrate these heterogeneous rankings a consensus order was computed using the Branch & Bound (B&B) algorithm^{42,43}. This procedure yields the most representative global ranking while preserving the pairwise agreements among models.

Interactions analysis

For system-level assessment, the framework adopts a multi-method approach based on temporal and statistical association analysis. Specifically, techniques such as Granger test testing, regression-based modeling, and cross-correlation analysis can be employed complementary to investigate temporal precedence in prediction, short-lag co-variation, and lag-structure between sensors of concurrently operating machines. For the interaction analyses, two datasets from pumps A and B were used.

Granger test

The Granger test⁴⁴ evaluates whether past values of one time series X_t contain predictive information about another series Y_t . It compares two autoregressive models: one including only the lagged terms of Y_t , and another that also incorporates lagged terms of X_t . If the inclusion of X_t significantly reduces the prediction error for Y_t ,

Table 5 Optimal tuning parameters for the DBSCAN of pumps A, B, C.

	ϵ	MinPts	Average silhouette
A	0.792	3	0.610
B	0.844	4	0.474
C	0.526	3	0.524

Table 6 Optimal cross-validation tuning parameters for each model applied to the variable importance analysis of A, B, C.

	Lasso	Ridge	RF	DT
BestTune	λ	λ	mtry	cp
A	0.00035	0.02653	2	0.99
B	0.00011	0.00994	2	0.99
C	0.00051	0.04603	2	0.99

then X_t is said to Granger-cause Y_t in the statistical sense of the test, that is, X_t carries information about the future of Y_t not already contained in the past of Y_t . This is a statement about temporal precedence in prediction, not about physical causal influence: with shared inputs or latent common drivers, predictive-precedence relations may emerge without a direct mechanical link⁴⁵. The test is therefore informative as a screening tool for directional predictive relations between coupled sensor signals.

Cross correlation function (CCF)

CCF⁴⁶ was used to characterise the lag structure of candidate cross-pump links suggested by Granger and/or regression evidence, and to flag long-lag peaks that are not physically plausible under the observed operating overlap. The CCF measures the degree of similarity between two signals as a function of a time shift τ , revealing whether changes in one series precede, follow, or coincide with those in another. In practical terms, a positive lag ($\tau > 0$) indicates that variations in X_t anticipate those in Y_t , whereas a negative lag ($\tau < 0$) implies the opposite, and a peak at $\tau = 0$ denotes instantaneous coupling.

Parameter tuning used for the analysis

DBSCAN parameters were tuned through a two-step heuristic approach. The optimal ϵ was identified using the k -dist plot⁴⁷, which involved computing and sorting the k -nearest neighbour distances; the elbow in the resulting curve indicated the transition to dense clusters. With ϵ fixed, the MinPts parameter was optimized by evaluating silhouette scores⁴⁸ across values ranging from 2 to 10 and selecting the one yielding the highest score. Euclidean distance was used as the similarity metric. The results of the tuning process are summarized in Table 5.

In addition to the parameters specific to DBSCAN, several other model-specific hyperparameters were optimized. These include the regularization parameter λ for both Ridge and Lasso regression, which controls the strength of the penalty term, the complexity parameter for DT, which determines the degree of pruning, the number of trees in the ensemble for RF, and the number of variables considered at each split (mtry). All hyperparameters were selected using a 10-fold cross-validation approach. Furthermore, for both Lasso and Ridge regression models, the objective function minimized during training was the binomial deviance³⁷. The results of the hyperparameters tuning for each model are summarized in Table 6.

The B&B algorithm was used for consensus ranking, the Gini index for DT splitting. Granger predictive-precedence tests were carried out by fitting each pair of stationary variables to a VAR model with lag order p minimising AIC for $p \in [1, 20]$.

Table 7 DBSCAN clustering results for centrifugal pumps A, B, C.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Noise Points
A	340	369	4222	–	49
B	850	24	290	3792	24
C	1267	698	7	2984	49

Table 8 Cumulative variance percentage explained by the first three principal components for centrifugal pumps A, B, and C.

	Dim.1	Dim.2	Dim.3
A	46.517	68.704	81.063
B	47.384	69.317	86.552
C	55.514	80.514	92.749

Table 9 Summary of loadings, contributions, and $\cos^2$ for the first two PCA components for each variable in A.

	I Component			II Component		
	Loadings	Contributions	$\cos^2$	Loadings	Contributions	$\cos^2$
A_ACR_Mot.PV	0.772	12.809	0.596	0.387	6.744	0.150
A_ACR_Mot.SV	0.746	11.981	0.557	0.272	3.338	0.074
A_ACR_Mot.TV	–0.712	10.904	0.507	0.635	18.162	0.403
A_ACR_Pmp.PV	0.876	16.502	0.768	0.386	6.701	0.149
A_ACR_Pmp.SV	0.573	7.071	0.329	0.314	4.443	0.099
A_ACR_Pmp.TV	–0.439	4.143	0.193	0.794	28.429	0.631
A_Pres.PV	0.873	16.400	0.763	0.353	5.627	0.125
A_Temp.PV	–0.558	6.687	0.311	0.735	24.351	0.540
Barometer	0.259	1.441	0.067	0.107	0.512	0.011
Temperature	0.749	12.061	0.561	–0.194	1.693	0.038

Results

This section presents the results obtained by applying each analytical module to the real-world industrial dataset previously described. To ensure the robustness and consistency of the entire analytical process, parameter tuning was carried out across all methodological stages. The specific strategies and selected values are described at the end of this section.

Operational state analysis

The empirical findings of the machine-level module are reported below.

Operational states (DBSCAN)

Using the optimized parameters (ϵ and MinPts) derived from silhouette and k -distance analyses, centrifugal pump A was segmented into three clusters, while pumps B and C exhibited four clusters each. The number of data points assigned to each cluster, along with those classified as noise, are summarized in Table 7.

Principal component structure

The analysis revealed that, for all three pumps, the first two principal components accounted for 70–80% of the total variance (Table 8), highlighting that most of the variability in vibration, temperature, and pressure signals can be represented in a compact, physically interpretable two-dimensional space. For the description of the dimensions factor loadings, contributions and $\cos^2$ were interpreted (Tables 9, 10, 11). For all three centrifugal pumps, the variables with the highest contributions and loadings on the first dimension are those related to the

Table 10 Summary of loadings, contributions, and $\cos^2$ for the first two PCA components for each variable in B.

	I Component			II Component		
	Loadings	Contributions	$\cos^2$	Loadings	Contributions	$\cos^2$
B_ACR_Mot.PV	0.720	10.935	0.518	0.157	1.130	0.025
B_ACR_Mot.SV	0.944	18.810	0.891	0.181	1.489	0.033
B_ACR_Mot.TV	−0.527	5.860	0.278	0.705	22.691	0.498
B_ACR_Pmp.PV	0.932	18.350	0.870	0.202	1.856	0.041
B_ACR_Pmp.SV	0.944	18.829	0.892	0.186	1.581	0.035
B_ACR_Pmp.TV	−0.463	4.532	0.215	0.830	31.434	0.689
B_Pres.PV	0.916	17.707	0.839	0.191	1.657	0.036
B_Temp.PV	−0.423	3.775	0.179	0.299	4.089	0.090
Barometer	−0.237	1.186	0.056	−0.159	1.160	0.025
Temperature	−0.027	0.015	0.001	−0.850	32.913	0.722

Table 11 Summary of loadings, contributions, and $\cos^2$ for the first two PCA components for each variable in C.

	I Component			II Component		
	Loadings	Contributions	$\cos^2$	Loadings	Contributions	$\cos^2$
C_ACR_Mot.PV	0.880	13.944	0.774	0.319	4.076	0.102
C_ACR_Mot.SV	0.922	15.304	0.850	0.323	4.186	0.104
C_ACR_Mot.TV	−0.664	7.934	0.440	0.692	19.175	0.479
C_ACR_Pmp.PV	0.948	16.203	0.899	0.264	2.788	0.070
C_ACR_Pmp.SV	0.945	16.081	0.893	0.295	3.489	0.087
C_ACR_Pmp.TV	−0.573	5.906	0.328	0.808	26.101	0.653
C_Pres.PV	0.940	15.916	0.884	0.298	3.559	0.089
C_Temp.PV	−0.588	6.233	0.346	0.794	25.224	0.630
Barometer	−0.342	2.106	0.117	0.206	1.701	0.042
Temperature	−0.144	0.373	0.021	−0.492	9.699	0.242

displacement and acceleration of both the pump and the motor. On the other hand, the variables that contribute most to the second dimension are those related to the internal (of pump and motor) and external temperature. Consequently, it can be inferred that the first axis represents the increase in acceleration/displacement of the machinery, while the second axis captures the increase in the temperature of the components, which may occur during system ignition. Based on this interpretation, the first quadrant of the plan could represent the steady-state phase of the three centrifugal pumps, while the fourth quadrant could represent the shutdown phase. After interpreting the principal axes, the joint visualization of variables and observations was obtained through a biplot⁴⁹, which simultaneously displays both in the same coordinate space (Fig. 5). Across all centrifugal pumps, clusters occupy distinct regions of the biplot, corresponding to specific operational regimes distributed across the principal-component plane. The upper-right region (positive PC1 and PC2) mainly represents steady-state operation, characterized by high displacement and temperature; the lower-right region reflects shutdown or pre-ignition conditions with low displacement and moderate temperature; while intermediate clusters located near the origin indicate transitional phases such as early ignition or cooling-down. This consistent spatial arrangement across pumps confirms that the clustering structure reflects physically meaningful transitions in machine behaviour.

Decision rules

The resulting trees (Supplementary Fig. S2 online), were expressed as concise *if-then* rules describing operational states in terms of thermal and mechanical thresholds. Representative rules are reported below for each pump; intermediate branches follow analogous structures, combining the same variables with slightly adjusted thresholds.

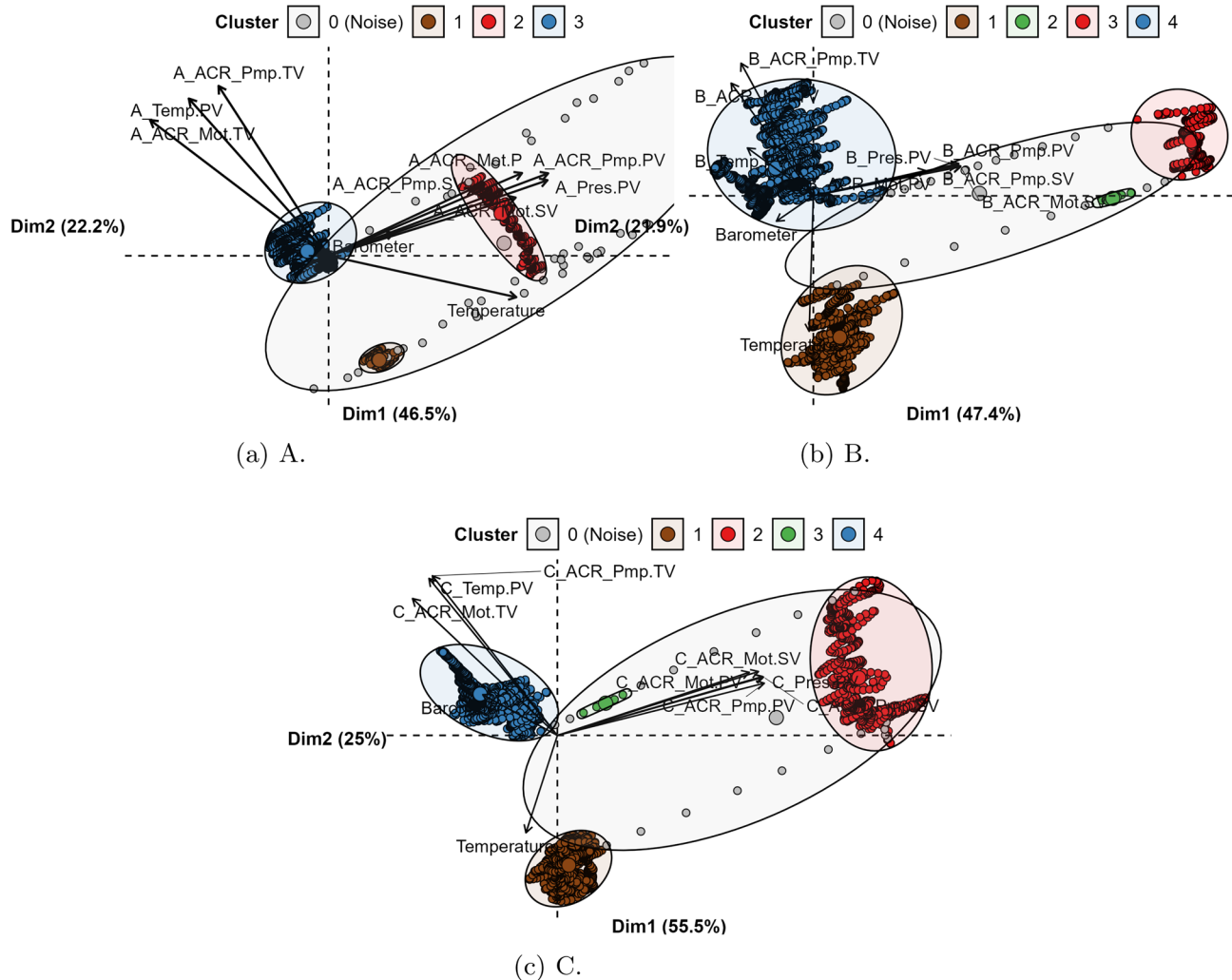


Fig. 5 Biplots visualization of the relationship between grouped observations and variables of centrifugal pumps A, B, C.

- Pump A: if motor temperature < 26 °C and acceleration < 0.7 m/s² → cluster 1; if motor temperature > 26 °C → cluster 3.
- Pump B: if pump temperature < 20 °C and motor acceleration < 2.8 m/s² → cluster 1; if pump temperature > 20 °C → cluster 4.
- Pump C: if pump temperature < 22 °C and pump acceleration < 2 m/s² → cluster 1; if motor temperature > 22 °C → cluster 4.

These rules confirm that thermal and mechanical thresholds dominate the separation of operational regimes, while pressure acts mainly as a secondary variable refining boundary conditions. Evaluated against the DBSCAN labels, the decision trees reproduce the cluster assignment with high fidelity, reaching an overall classification accuracy of 99.82% for pump A, 99.84% for pump B and 99.38% for pump C.

The rules assign each observation to a single regime at a fixed value of one variable. To quantify the uncertainty of this assignment, the DBSCAN labels were also fitted with an ante-hoc interpretable model, a network with an input-gating layer⁵⁰, which returns a membership probability for each regime. The two assignments agree for 99.8% (A), 99.8% (B) and 99.4% (C) of the observations, and the probability of the assigned regime is below 0.90 for less than 1.1% of them. Intermediate probabilities occur only close to the thresholds, where the probability of the assigned regime rises from 0.1 to 0.9 over 0.95 °C (A), 1.23 °C (B) and 0.47 °C (C); this

interval contains 2-3% of the observations, with a mean probability of 0.95-0.96. The probabilistic assignment therefore reproduces the regimes obtained from the rules, and the thresholds are indeterminate only within about one degree. Supplementary Table S4 online reports the full results.

Figure 6 illustrates the temporal evolution of cluster membership for each pump, showing stable, recurrent transitions. Despite the independence assumption used during clustering, each cluster identified by DBSCAN forms temporally contiguous runs and the transitions between clusters are sparse and structured, supporting the interpretation of the identified groups as sustained operating regimes rather than as artefacts of momentary multivariate similarity. To make this interpretation explicit in physical terms, Table 12 reports, for each cluster, the mean outlet pressure, pump peak acceleration and motor casing temperature in their original units. The cluster signatures therefore match the operating phase expected for those physical values: as no independent on/off operating log is available, these recorded process variables (most directly the outlet pressure) serve as the reference operating state, so this agreement provides an internal validation of the assigned cluster labels.

Environmental conditions impact analysis

The empirical findings of the context-level module are reported below.

Distributional shift

Across pumps and variables, the KS-Test rejects the null hypothesis of distributional equality between the April and June campaigns for all ten sensor signals on each of the three pumps, indicating campaign-level drift in pressure, vibration, and temperature measurements. Two cautions are important for interpretation. First, with large sample sizes even small distributional changes can yield very small p -values; second, sensor time-series are not

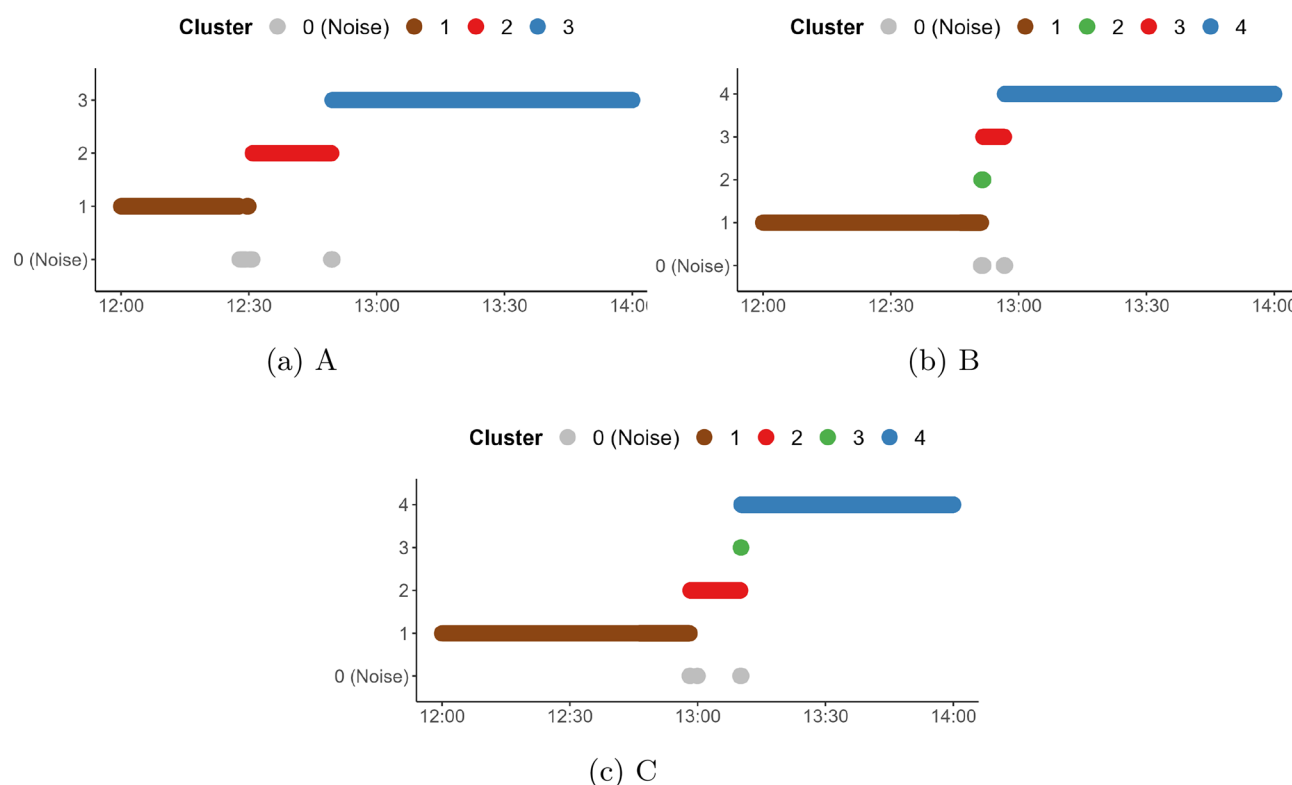


Fig. 6 Temporal evolution of operational clusters for centrifugal pumps A, B, and C. Each colour segment represents a time window assigned to a specific cluster, capturing dynamic transitions in machine behaviour over the monitoring period.

Table 12 Physics-aware characterisation of the DBSCAN clusters: per-cluster mean of outlet pressure, pump peak acceleration and motor casing temperature in their original physical units (noise points excluded).

Pump	Cluster	Outlet pressure [bar]	Pump peak acceleration [m/s^2]	Motor casing temperature [$^{\circ}\text{C}$]	Interpreted operational phase
A	1	0.45	0.6	18.1	Standby/shut-down
A	2	41.8	21.5	24.2	Steady-state pumping
A	3	0.47	0.5	28.0	Post-operation cooling
B	1	0.48	0.4	18.6	Standby/shut-down
B	2	0.48	22.0	18.7	Start-up/no-discharge transient
B	3	43.2	23.3	18.8	Steady-state pumping
B	4	0.49	0.5	19.3	Standby/shut-down
C	1	0.45	0.4	19.3	Standby/shut-down
C	2	43.3	28.5	21.2	Steady-state pumping
C	3	43.2	4.4	23.6	Steady-state pumping (low vibration)
C	4	0.45	0.4	26.0	Post-operation cooling

Clusters with near-zero outlet pressure and vibration correspond to standby/shut-down states, clusters with high pressure and vibration correspond to active steady-state pumping, and elevated casing temperature in otherwise inactive clusters identifies post-operation cooling, consistently with the operational phase expected for those physical values.

i.i.d. and can reflect changes in operating regime and scheduling in addition to ambient conditions. Accordingly, the KS-Test is used here as a screening step to detect campaign-level shift, while the subsequent variable-importance analysis helps determine which measurements (including the external sensors) most strongly discriminate between the two regimes.

Variable importance and consensus ranking

Supplementary Fig. S3 online illustrates the variable-importance profiles for each algorithm across the three pumps, while Supplementary Table S1 online reports the final consensus ordering. Finally, the consensus ranking was estimated. The consensus ranking consistently places environmental temperature and barometric pressure as the top two discriminators between the April and June campaigns across all three pumps; internal temperatures (ACR_Pmp.TV, Temp.PV) follow. Importantly, because the response is the campaign indicator, these importance scores quantify *discriminability* rather than causal effect: operating regime, demand, and maintenance actions may co-vary with the campaign and partially explain the observed drift. In this work we therefore interpret the context module as a drift screening step; the campaign indicator aggregates ambient variability and operational/scheduling variability, and the procedure does not separate the two causally

Interactions analysis

The empirical findings of the system-level module are reported below.

Granger and regression evidence

Pairwise Granger tests were applied to all variable combinations between the two pumps to identify directional predictive relations, that is, cases in which past values of a signal from one pump statistically precede future values of a signal from the other pump. These relations are not interpreted as causal. To ensure robustness, all tested relations are reported for transparency; in the discussion we emphasise those supported by joint evidence (Granger $p \leq 0.05$ and regression $p \leq 0.01$) and physically plausible short lags. Subsequently, linear regression models were fitted to each significant pair to verify whether the dependencies exhibited a consistent and immediate trend. For the associated t -tests on regression coefficients are reported; we treat $p \leq 0.01$ as strong evidence supporting the sign and magnitude of the association. This two-stage filtering ensured that only strong, reproducible directional relationships were preserved for interpretation in the interaction analysis.

Table 13 Qualitative taxonomy of system-level directional predictive relations between pumps A and B.

Interaction types are interpretative labels guided by variable semantics, plant connectivity, and lag plausibility; they describe directional predictive relations compatible with shared-circuit mechanisms.

Interaction type (qualitative)	Representative variable pair(s)	Interpretative rationale
Hydraulic continuity	A_ACR_Mot.SV → B_Pres.PV A_Pres.PV → B_Pres.PV	Near-zero/short-lag associations involving pressure are consistent with hydraulic continuity in a shared circuit
Dynamic compensation	B_ACR_Pmp.SV → A_ACR_Mot.PV B_ACR_Pmp.SV → A_ACR_Mot.SV	Short-lag co-variation between mechanical proxies across pumps may reflect coordinated regulation (e.g., inverter/control actions) to balance load and maintain total discharge
Thermal propagation	B_ACR_Pmp.TV → A_Temp.PV A_Temp.PV → B_ACR_Pmp.TV	Short-lag links involving temperature variables are consistent with heat propagation through the shared fluid/housing.

Cross-correlation and interaction types

Since pump B was active from 09:45 to 16:10 and pump A operated from 12:04 to 14:22, any pronounced peaks at $\tau < -2$ min are more plausibly explained by different activation windows and initial thermal conditions (pre-heating/scheduling) than by real-time propagation during simultaneous operation. The interaction analysis shows that The relations supported by joint Granger and regression evidence are concentrated at $|\tau| \leq 1$ min, with the cross correlation peak located at $\tau = 0$ for the majority of these relations. Predictive precedence is therefore present in the conditional autoregressive structure, but the strongest association between sensor pairs is mostly contemporaneous rather than occurring with a measurable time delay. Representative examples of the combined Granger, regression, and CCF results are reported in Supplementary Table S2 online for the direction A→B and in Supplementary Table S3 online for B→A. Table 13 then groups the retained short-lag links into qualitative interaction types. Across the retained short-lag links, four dominant evidence patterns emerged:

- *Hydraulic continuity* Short-lag associations involving pressure are consistent with coupling through a shared circuit (e.g., A_Pres.PV → B_Pres.PV, with the CCF peak at $\tau \approx 0$). This is interpreted as directional predictive relation consistent with shared-circuit coupling rather than proof of physical causality.
- *Dynamic compensation* Near-zero-lag co-variation between mechanical proxies across pumps (e.g., B_ACR_Pmp.SV → A_ACR_Mot.PV, with a strong instantaneous CCF) may reflect coordinated regulation (e.g., inverter/control actions) to balance load and maintain discharge; alternative common-drive explanations cannot be excluded from observational data.
- *Thermal propagation* Short-lag links involving temperature-related variables are consistent with heat propagation through the shared fluid/housing (e.g., B_ACR_Pmp.TV → A_Temp.PV, with the strongest association at $|\tau| \leq 1$ min), and are interpreted only within physically plausible lag windows

Discussion

The three modules yield a descriptive characterisation of pump behaviour at the machine, context, and system levels. At the machine level, DBSCAN and PCA separate startup, steady-state, and shutdown regimes, approximated by decision-tree rules on pressure, temperature, and vibration thresholds. At the context level, Kolmogorov-Smirnov tests detect distributional shifts between the April and June campaigns, and the consensus ranking across Lasso, Ridge, decision trees, and random forests places environmental temperature and barometric pressure as the top two discriminators. At the system level, Granger predictive-precedence tests and cross-correlation analysis identify directional predictive relations between pumps A and B at $|\tau| \leq 1$ min, with peaks predominantly at $\tau = 0$, grouped into three interaction types: hydraulic continuity, dynamic compensation, and thermal propagation. The machine-level results are consistent with clustering-based analyses of rotating equipment^{16,17}; the incremental

step here is the rule-based translation via decision trees and the integration into a multi-level scheme. The context-level results are consistent with reports that ambient temperature and pressure influence signal characteristics of rotating equipment⁹; the added element is the consensus ranking across four heterogeneous estimators. The system-level results are consistent with plant-wide analyses of interaction propagation^{10–12}; the interaction taxonomy adds a qualitative interpretation layer over the statistical output.

Several considerations should be made when interpreting these findings. First, at the machine level no on/off operating log is available, so cluster labels are validated internally against the physical values of the variables, consistently with the interpretability-by-construction principle of the pipeline. Second, the context-level analysis does not isolate ambient effects, since the campaign indicator jointly encodes ambient, operational, and scheduling variability, which cannot be separated without a controlled experiment in a production plant; the module is accordingly positioned as a drift-diagnostic step, with the consensus ranking reported as an indication of which measurements most strongly discriminate the campaigns. Third, the system-level analysis is descriptive rather than causal: Granger tests identify temporal precedence in prediction, and under shared inputs or latent common drivers such precedence may arise without a direct mechanical link⁴⁵; the Granger step is therefore combined with cross-correlation and only associations at short lags ($|\tau| \leq 1$ min) are retained. The tests employed at this level are moreover linear in structure, so dependencies expressed through nonlinear lead-lag mechanisms are not captured; continuous-time architectures with input-dependent time constants⁵¹ are a candidate direction for modelling such dependencies, at the cost of the transparency that motivates the present pipeline. Fourth, the empirical demonstration is confined to three pumps of a single plant across three acquisition campaigns of heterogeneous duration; the plant was selected for jointly exposing multiple machines, seasonal variability, and shared hydraulic and thermal paths within a single public dataset, and the modular structure of the framework supports transfer to other plants and equipment classes as further public datasets become available.

Open research questions

Several open research questions follow from the study and from the reviewed literature.

The first concerns the extension of the framework to other classes of rotating and multi-stage industrial equipment (turbines, compressors, production lines with sequential stages), whose dynamics differ from those of centrifugal pumps in the coupling between mechanical, thermal, and process variables. A candidate direction is the adaptation of the three modules to equipment-specific sensing schemas and operating regimes, and the assessment of how the interaction taxonomy generalises to architectures in which the shared paths are not primarily hydraulic. Where the acquisition history available for such equipment is limited, data-efficient formulations of the individual modules, of the kind surveyed for data-scarce industrial settings⁵², would be required.

The second concerns the transition from an offline pipeline operating on archived acquisition windows to a streaming implementation for continuous monitoring. A candidate direction is the development of online counterparts of the three modules with mechanisms distinguishing transient anomalies from persistent regime changes.

The third concerns the integration of the framework with physics-based models of the plant. Purely data-driven analyses of coupled industrial machines remain ambiguous under latent common drivers⁴⁵, whereas hybrid pipelines that combine data-driven components with domain knowledge have been advocated in the recent literature^{15,53}. A candidate direction is the incorporation of physics-informed priors into the interaction module, so as to constrain the space of admissible dependencies and to move from directional predictive relations toward physically interpretable ones.

The fourth concerns the validation of the interaction taxonomy against physics-based fault-propagation paths. The present study is one component of a broader effort in which the analysis of real-plant data is integrated with dedicated software development toward a digital twin of the plant; within that roadmap, the qualitative interaction types derived here are to be validated against fault-propagation paths reproduced by the twin, providing a stringent test of their engineering plausibility. The same direction applies at the context level: physical constraints

linking ambient variables to machine response would allow the drift between acquisition campaigns to be attributed at comparable operating conditions, rather than screened as an aggregate difference.

Conclusions

This work has presented a modular, interpretable, three-level framework for the descriptive analysis of industrial centrifugal pumps, combining DBSCAN and PCA at the machine level, non-parametric distributional testing with a consensus variable-importance ranking at the context level, and Granger predictive-precedence testing with cross-correlation analysis at the system level. Applied to a public real-plant dataset, it produces regime baselines, drift diagnostics, and an interaction taxonomy, each reported in a form open to inspection by a plant engineer. Within the limitations discussed above, the outputs provide a reproducible baseline for extensions to other equipment classes, larger interconnected systems, and streaming implementations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-71432-2>.

Acknowledgements The CIRA group acknowledge the SACIP (SAlus CIRA Plant) Project, funded by the Italian PRORA, under which the data were collected, the framework was designed and developed, and the data analysis was performed.

Author contributions This paper is a joint cooperation among authors. G.Z. original key idea proposal. Al.DA. main author of the original draft of the paper and the technical implementation, including the overall statistical framework and the data mining contribution. An.DA. and R.S. validation of the statistical learning approaches. A.M. provision of the dataset description and domain-specific insights. M.F. and G.R. definition of the domain objectives. All authors revised the final version of the manuscript.

Data availability Data are publicly available on Zenodo <https://doi.org/10.5281/zenodo.18479728>, accessed on May 15, 2026.

Declarations

Competing interest The authors declare no competing interests.

Use of generative AI The authors declare that no generative artificial-intelligence tools were used to create, generate, or alter any of the figures or images in this manuscript; all figures, including the schematic illustrations of the framework and of the data-acquisition system, are original and were produced by the authors. No generative AI system, including MS Copilot, was used for the study design, the literature review, data collection, data analysis, the interpretation of the results, or the generation of the scientific content of the manuscript. The authors take full responsibility for the content of the publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Nemer, M. A., Azar, J., Makhoul, A. & Bourgeois, J. A review of early time series classification methods on machinery dataset. In *2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics*, 347–352. <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics62450.2024.00074> (2024).
2. Tsallis, C., Papageorgas, P., Piromalis, D. & Munteanu, R. A. Application-wise review of machine learning-based predictive maintenance: Trends, challenges, and future directions. *Appl. Sci.* <https://doi.org/10.3390/app15094898> (2025).
3. Manco, G. et al. Fault detection and explanation through big data analysis on sensor streams. *Expert Syst. Appl.* **87**, 141–156. <https://doi.org/10.1016/j.eswa.2017.05.079> (2017).
4. Gawde, S. et al. Multi-fault diagnosis of industrial rotating machines using data-driven approach: A review of two decades of research. *Eng. Appl. Artif. Intell.* **123**, 106139. <https://doi.org/10.1016/j.engappai.2023.106139> (2023).
5. Jiang, X., Tang, W., Zhao, H. & Cheng, X. Toward smart condition monitoring of rotatory machines: An optimized probabilistic signal reconstruction methodology for fault prediction with multisource uncertainties. *IEEE Access* **10**, 60862–60875. <https://doi.org/10.1109/ACCESS.2022.3180888> (2022).
6. Diaz, F. H., Borrás, C. & García Cena, C. E. Fault detection in axial piston hydraulic pumps: Integrating principal component analysis with Silhouette-based cluster evaluation. *Enfoque UTE* **16**, 1–9. <https://doi.org/10.29019/enfoqueut.e.1120> (2025).
7. Wang, Q., Liu, J., Wei, B., Chen, W. & Xu, S. Investigating the construction, training, and verification methods of k-means clustering fault recognition model for rotating machinery. *IEEE Access* **8**, 196515–196528. <https://doi.org/10.1109/ACCESS.2020.3028146> (2020).
8. Brusa, E., Cibrario, L., Delprete, C. & Di Maggio, L. G. Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring. *Appl. Sci.* <https://doi.org/10.3390/app13042038> (2023).
9. Ji, Y., Giangrande, P. & Zhao, W. Effect of environmental and operating conditions on partial discharge activity in electrical machine insulation: A comprehensive review. *Energies* <https://doi.org/10.3390/en17163980> (2024).
10. Hua, D., Dong, J., Xiang Peng, K. & Simani, S. An integrated distributed fault diagnosis framework for large-scale industrial processes based on spatiotemporal causal analysis. *IEEE Trans. Ind. Inform.* <https://doi.org/10.1109/TII.2025.3567381> (2025).
11. Wu, P. et al. Multichannel dynamic graph convolutional network-based fault diagnosis and its application in blast furnace ironmaking process. *IEEE Sens. J.* <https://doi.org/10.1109/JSEN.2023.3325353> (2023).
12. Chen, X., Yang, Y. & Wang, J. Plant-wide processes monitoring and fault tracing based on causal graphical model. *IET Control Theory Appl.* **18**, 2322–2334. <https://doi.org/10.1049/cth2.12499> (2024).
13. de Campos Souza, P. V., Lughofer, E. & Guimaraes, A. J. An interpretable evolving fuzzy neural network based on self-organized direction-aware data partitioning and fuzzy logic neurons. *Appl. Soft Comput.* **112**, 107829. <https://doi.org/10.1016/j.asoc.2021.107829> (2021).
14. Liu, Y., Wang, T. & Chu, F. Hybrid machine condition monitoring based on interpretable dual tree methods using Wasserstein metrics. *Expert Syst. Appl.* **235**, 121104. <https://doi.org/10.1016/j.eswa.2023.121104> (2024).
15. Feng, X., Wang, D., Hou, B. & Yan, T. Interpretable federated learning for machine condition monitoring: Interpretable average global model as a fault feature library. *Eng. Appl. Artif. Intell.* **124**, 106632. <https://doi.org/10.1016/j.engappai.2023.106632> (2023).
16. Oliosi, E., Calzavara, G. & Ferrari, G. On sensor data clustering for machine status monitoring and its application to predictive maintenance. *IEEE Sens. J.* **23**, 9620–9639. <https://doi.org/10.1109/jsen.2023.3260314> (2023).
17. Bekar, E. T., Nyqvist, P. & Skoogh, A. An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. *Adv. Mech. Eng.* <https://doi.org/10.1177/1687814020919207> (2014).
18. Nadim, T., Ragab, M. & Andersen, B. Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining. *J. Intell. Manuf.* **33**, 1221–1237. <https://doi.org/10.1007/s10845-021-01903-y> (2022).
19. Mazzoleni, M. et al. A fuzzy logic-based approach for fault diagnosis and condition monitoring of industry 4.0 manufacturing processes. *Eng. Appl. Artif. Intell.* **115**, 105317. <https://doi.org/10.1016/j.engappai.2022.105317> (2022).
20. Chatterjee, J. & Dethlefs, N. Temporal causal inference in wind turbine Scada data using deep learning for explainable AI. *J. Phys. Conf. Ser.* **1618**, 022022. <https://doi.org/10.1088/1742-6596/1618/2/022022> (2020).
21. Dang, X.-H., Shah, S. Y. & Zerfos, P. seq2graph: Discovering dynamic non-linear dependencies from multivariate time series. In *2019 IEEE International Conference on Big Data (Big Data)*, 1774–1783. <https://doi.org/10.1109/BigData47090.2019.9006103> (2019).
22. Calabrese, F., Regattieri, A., Bortolini, M. & Galizia, F. G. Data-driven fault detection and diagnosis: Challenges and opportunities in real-world scenarios. *Appl. Sci.* <https://doi.org/10.3390/app12189212> (2022).

23. Marcinkevičs, R. & Vogt, J. E. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* <https://doi.org/10.1002/widm.1493> (2023).
24. Martone, A. & Zazzaro, G. Centrifugal pump dataset. *Zenodo* <https://doi.org/10.5281/zenodo.18479728> (2026).
25. CIRA—Italian Aerospace Research Centre. Official website of CIRA (2026). Accessed May 2026.
26. Sunal, C. E., Dyo, V. & Velisavljevic, V. Review of machine learning based fault detection for centrifugal pump induction motors. *IEEE Access* **10**, 71344–71355. <https://doi.org/10.1109/ACCESS.2022.3187718> (2022).
27. Chen, L., Li, Z., Shi, W. & Li, W. Research on fault detection and automatic diagnosis technology of water hammer in centrifugal pump. *Appl. Sci.* **14**, 5606. <https://doi.org/10.3390/app14135606> (2024).
28. Sánchez, R.-V. et al. Evaluation of hand-crafted feature extraction for fault diagnosis in rotating machinery: A survey. *Sensors* **24**, 5400. <https://doi.org/10.3390/s24165400> (2024).
29. Stavropoulos, P., Papacharalampopoulos, A. & Souflas, T. Indirect online tool wear monitoring and model-based identification of process-related signal. *Adv. Mech. Eng.* **12**, 1687814020919209. <https://doi.org/10.1177/1687814020919209> (2020).
30. Li, Z., He, Q. & Li, J. A survey of deep learning-driven architecture for predictive maintenance. *Eng. Appl. Artif. Intell.* **133**, 108285. <https://doi.org/10.1016/j.engappai.2024.108285> (2024).
31. Martone, A. et al. Sensor-based monitoring data from an industrial system of centrifugal pumps. *Data*. <https://doi.org/10.3390/data10060091> (2025).
32. Liu, Y., Dang, B., Li, Y., Lin, H. & Ma, H. Applications of Savitzky–Golay filter for seismic random noise reduction. *Acta Geophysica* **64**, 101–124. <https://doi.org/10.1515/acgeo-2015-0062> (2016).
33. Lu, J., Ran, Q., Wang, H., Tan, K. & Pei, Z. Research on mems gyroscope motion signal processing method based on stochastic resonance and optimized mode decomposition. *Meas. Sci. Technol.* **35**, 126103. <https://doi.org/10.1088/1361-6501/ad727f> (2024).
34. Hodrick, R. J. & Prescott, E. C. Postwar U.S. business cycles: An empirical investigation. *J. Money Credit Bank.* **29**, 1–16. <https://doi.org/10.2307/2953682> (1997).
35. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* <https://doi.org/10.1145/3068335> (2017).
36. Bagherzade Ghazvini, M., Sánchez-Marré, M., Bahilo, E. & Angulo, C. Operational modes detection in industrial gas turbines using an ensemble of clustering methods. *Sensors* <https://doi.org/10.3390/s21238047> (2021).
37. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 2nd edn. (Springer, 2021).
38. Rojas-Lima, J. E., Domínguez-Pacheco, F. A., Hernández-Aguilar, C., Hernández-Simón, L. M. & Cruz-Orea, A. Kolmogorov–Smirnov test for statistical characterization of photopyroelectric signals obtained from maize seeds. *Int. J. Thermophys.* **40**, 4. <https://doi.org/10.1007/s10765-018-2462-4> (2018).
39. Kini, K. R., Madakyaru, M., Harrou, F., Menon, M. K. & Sun, Y. Improved fault detection in chemical engineering processes via non-parametric Kolmogorov–Smirnov-based monitoring strategy. *ChemEngineering* <https://doi.org/10.3390/chemengineering8010001> (2024).
40. Laufer, B., Docherty, P. D., Murray, R., Krueger-Ziolek, S. & Martens, S. Sensor selection for tidal volume determination via linear regression-impact of lasso versus ridge regression. *Sensors* **23**, 7407. <https://doi.org/10.3390/s23177407> (2023).
41. Boulesteix, A., Janitza, S., Kruppa, J. & König, I. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2**, 493–507. <https://doi.org/10.1002/widm.1072> (2012).
42. Marden, J. I. *Analyzing and Modeling Rank Data* (Chapman & Hall, 1996).
43. Nefedov, V. N. & Osipova, V. A. Consolidation of individual rankings by the method of branches and bounds. *J. Comput. Syst. Sci. Int.* **61**, 981–989. <https://doi.org/10.1134/S1064230722060144> (2022).
44. Shojaie, A. & Fox, E. B. Granger causality: A review and recent advances. *Annu. Rev. Stat. Appl.* **9**, 289–319. <https://doi.org/10.1146/annurev-statistics-040120-01s0930> (2022).
45. Eichler, M. Causal inference with multiple time series: Principles and problems. *Philos. Trans. R. Soc. A* **371**, 20110613. <https://doi.org/10.1098/rsta.2011.0613> (2013).
46. Box, G. E. P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics, 5th edn (Wiley, 2015). ISBN 978-1-118-67502-1.
47. Bethaz, P. & Cerquitelli, T. Enhancing the friendliness of data analytics tasks: An automated methodology. In *Proceedings of the EDBT/ICDT Workshops*, 1–6. <https://doi.org/10.1145/3437984.3438085> (2021).
48. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) (1987).
49. Greenacre, M. J. *Biplots in Practice* (Fundación BBVA, 2010).
50. Zhou, L. et al. Towards trustworthy AI in industry 5.0: Ante-hoc interpretability with deep learning. *Comput. Ind. Eng.* **213**, 111805. <https://doi.org/10.1016/j.cie.2026.111805> (2026).
51. Leng, J. et al. Liquid neural network in smart manufacturing: A new opportunity. *Adv. Eng. Inform.* **70**, 104222. <https://doi.org/10.1016/j.aei.2025.104222> (2026).

52. Leng, J. et al. Untrained neural networks in data-scarce smart manufacturing: A paradigm for data-efficient industrial intelligence. *Comput. Ind. Eng.* **213**, 111744. <https://doi.org/10.1016/j.cie.2025.111744> (2026).
53. Leng, J. et al. Physics-informed machine learning in intelligent manufacturing: A review. *J. Intell. Manuf.* **37**, 2215–2257. <https://doi.org/10.1007/s10845-025-02641-1> (2026).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alessia D'Ambrosio¹ · Antonio D'Ambrosio¹ · Roberta Siciliano¹ · Angelo Martone² · Michele Ferrucci² · Gianpaolo Romano² · Gaetano Zazzaro² 

✉ Gaetano Zazzaro
g.zazzaro@cira.it

¹ University of Naples Federico II, Naples, Italy

² CIRA - Italian Aerospace Research Centre, Capua, Italy