Future Technology
Open Access Journal
ISSN 2832-0379

August 2026| Volume 05 | Issue 03 | Pages 150-163

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.3.14

Future Publishing LLC

Article

A data-driven multivariable framework for operational regime identification, product transition detection, and anomaly detection in industrial pumping systems using SCADA data

Johnatan Corrales-Bonilla*[1], William Hidalgo-Ozorio[1], Christian Corrales-Otáñez[2], Francisco Viteri[3]

[1]Universidad Técnica de Cotopaxi, Cotopaxi, Ecuador
[2]Independent Researcher, Ecuador
[3]Atlantis University, Miami, FL, USA

ARTICLE INFO

Article history:
Received 06 January 2026
Received in revised form
15 April 2026
Accepted 18 May 2026

Keywords:
SCADA data analytics, Multivariable analysis, Anomaly detection, Operational regimes, Centrifugal pumps, Condition monitoring

*Corresponding author
Email address:
johnatan.corrales5518@utc.edu.ec

DOI: 10.55670/fpll.futech.5.3.14

ABSTRACT

This study analyzes a centrifugal pumping system in an industrial facility using fifteen months of operational data collected from a Supervisory Control and Data Acquisition (SCADA) system. Applying a flow greater than zero criterion, 15,049 records corresponding to active operation were retained; after quality control and removal of incomplete and feature-inconsistent observations, 14,501 records were used for the multivariable analysis. Instead of analyzing variables independently, the study characterizes system behavior through the relationships among hydraulic, electrical, and fluid-related variables. Principal Component Analysis (PCA) is first applied, where the first two components explain 69.8% of the total variance. Based on this reduced representation, K-means clustering identifies two operational regimes, corresponding to dominant and low-load conditions. A Gaussian Mixture Model (GMM) applied to fluid density reveals two product regimes with mean values of 716.84 kg/m^3 and 830.35 kg/m^3. In addition, anomaly detection based on the Mahalanobis distance identifies 73 anomalous observations (0.5% of the dataset), associated with reduced discharge pressure, lower pressure differential, and decreased power consumption, indicating degraded operating conditions. The proposed framework provides a physically interpretable representation of system behavior, enabling the identification of operational regimes, product-related variations, and anomalous conditions within a unified analytical approach. This supports its application in industrial monitoring environments aligned with Industry 4.0 (I4.0) principles.

1. Introduction

Industrial pumping systems are essential components of fluid transport networks, particularly in the oil and gas, mining, power generation, and water distribution sectors. The hydraulic conditions under which fluid flows through a system are connected to the system's energy consumption and the properties of the transported fluid. Advanced Supervisory Control and Data Acquisition (SCADA) systems enable continuous data monitoring in industrial operations. In multiproduct pumping systems, variables such as flow rate, pressure, and power consumption interact to define system behavior. Due to their strong interdependence, these variables should be analyzed in conjunction rather than independently. While SCADA systems can capture high-frequency real-time data, industrial analysis is commonly performed using aggregated historical records, such as hourly reports stored in data historians. In this context, Leite et al. [1] highlight that the proliferation of industrial data opens up abundant possibilities for implementing data analytics-based monitoring to enhance system reliability and detect faults at an early stage in complex industrial infrastructures. Similarly, Das et al. [2] and Mokhtari et al. [3] identified anomalous patterns in operational datasets that predict failures in critical industrial systems. In modern industrial environments, advanced data analytics plays an important role in improving process monitoring, asset management, and

predictive maintenance strategies [4]. These environments integrate cyber-physical systems, smart sensing, and analytical platforms capable of collecting high-dimensional multivariable data streams that characterize the operational behavior of industrial processes. In this context, Du et al. [5] demonstrated that deep learning approaches for incorporating cyber-physical features yield more effective anomaly detection in industrial control systems. Similarly, Kim et al. [6] conducted a comparative evaluation and demonstrated that machine learning-based anomaly detection approaches outperform conventional threshold-based monitoring approaches in industrial control environments. Recent studies have also explored the use of generative models for industrial time-series data. However, most of these approaches focus on anomaly detection and do not explicitly address the joint interpretation of operational regimes, product transitions, and overall system behavior. For instance, Han and Gim [7] present a generative adversarial network-based approach for modeling industrial operational data, whereas Cho and Gong [8] construct a dynamic data abstraction framework that enables more efficient anomaly detection in complex industrial environments.

Industrial monitoring systems generally generate multivariable datasets that include operational variables such as flow rate, pressure, electrical power consumption, vibration, temperature, and the physical characteristics of the transported fluid. Due to the strong correlations between variables, classical univariate threshold-based monitoring strategies typically miss complex interactions between process variables. This limitation is particularly relevant in pumping systems, where hydraulic performance and energy consumption are directly coupled. To address this limitation, multivariable analytical methods that capture the dynamic characteristics of industrial processes have been investigated in several studies. For example, Liu et al. [9] proposed a time–frequency attention-based model for anomaly detection in complex systems, thereby highlighting the benefits of considering multivariable representations of the industrial process. Similarly, Aslam et al. [10] showed that integrated multivariable representations of operational data improve the detection of abnormal patterns in industrial control systems. Hybrid learning-oriented patterns were also proposed to improve anomaly detection performance. In particular, Pang et al. [11] proposed a hybrid-oriented monitoring framework using vector quantization and support vector machines for industrial anomaly detection. Additionally, ultralight anomaly detection models designed for industrial environments with limited computational resources have also been proposed [8].

Anomaly detection in industrial control systems is a major research area of statistical techniques, machine learning approaches, and deep learning frameworks. Aslam et al. [10] developed a transparent framework for detecting anomalies in an industrial SCADA environment. In contrast, Goetz and Humm [12] argued for dynamic graph-based models for cyber-physical production systems. Other studies include distributed learning approaches targeting anomalies in industrial use cases for anomaly detection. In this context, federated learning architectures have also been proposed to enable explainable anomaly detection in decentralized industrial environments [13]. Other time-series and deep learning frameworks have been applied in industrial control system anomaly detection [14]. Previous work has also shown that integrating multiple intrusion detection systems within a single system can improve anomaly detection across critical infrastructure [15]. Unsupervised techniques based on pattern analysis of communication data have also been effective in industrial systems [16]. Centrifugal pumping systems have attracted great attention in industrial contexts due to hydraulic phenomena such as cavitation, hydraulic imbalance, and mechanical component degradation. Recently, deep learning techniques have been used to identify mechanical faults in centrifugal pumps. For instance, Zaman et al. [17] proposed deep neural network-based architectures (VGG16, ResNet50) for fault detection in industrial pumping applications. Similarly, Sunal et al. [18] presented transfer learning strategies using convolutional neural networks to improve fault diagnosis accuracy in pumping applications. Zhao et al. [19] proved that multivariable analysis of operational data can expose anomalous behaviors in complex industrial systems that might otherwise go unnoticed with common monitoring techniques. Digital twin applications in pumping systems have been studied in recent literature [20,21], as have autoencoder- and ensemble-based methods for anomaly detection in industrial datasets [16,22]. Simultaneously, there has been significant development of deep learning- and graph-based approaches for industrial machinery condition monitoring [23-25].

In SCADA-based monitoring, work on anomaly detection using operational data and PLC–SCADA frameworks for industrial terminals has been reported in recent years [10,26]. Still, major limitations exist. Some recent works, including deep learning-based approaches such as DeepPipe [27], have exhibited good performance in anomaly detection for multiproduct pipeline systems. However, such models typically act as black-box solutions, restricting their interpretability and limiting their relevance in industrial decision-making systems. Moreover, sophisticated PLC–SCADA monitoring frameworks have been employed to optimize real-time data acquisition and control [26]. Nevertheless, they do not integrate multivariable statistical techniques to simultaneously monitor operational regimes, product transitions, and anomalous conditions. Three main gaps can be recognized. First, many studies investigate mechanical faults or sensor signal analysis while neglecting the interrelation of hydraulic features, energy consumption, and fluid characteristics. Second, most of these methods are validated on experimental or simulated data, with little application to real SCADA data from continuous industrial operation. Furthermore, in multiproduct pumping systems, where hydraulic performance is strongly affected by fluid density, the detection of product transitions and anomalies has been understudied. An integrated multivariable framework capable of addressing these phenomena simultaneously remains an open problem in industrial data analytics.

This work proposes a multivariable data analytics framework using SCADA data from an industrial centrifugal pumping system. This method uses historical operational data from two pumping units over approximately fifteen months of continuous industrial operation. In this work, we focus on data pre-processing and exploratory analysis, Principal Component Analysis (PCA), unsupervised clustering, Gaussian Mixture Model (GMM), and robust Mahalanobis distance-based anomaly detection. The study covers primary operating regimes, product transitions due to fluid density variations, and operational anomalies in pumping systems. A major contribution of this study is:

- Validation of a data-driven multivariable analytical framework based on real industrial SCADA data.

- Measurement of operational regimes with PCA and unsupervised clustering for dimensionality reduction.
- Product regimes recognition through GMM based on fluid density.
- Anomaly detection based on robust Mahalanobis distance with Minimum Covariance Determinant estimation.

This study provides a hybrid method integrating PCA-based dimensionality reduction, K-means operational regime identification, GMM-based product regime characterization, and robust Mahalanobis anomaly detection, building on previously developed works and approaches within an integrated, physically interpretable SCADA analytics workflow, compared with isolated anomaly detection, single-method clustering, or black-box predictive models. This integration is the study's primary methodological novelty, as it enables the joint examination of operational regimes, product-related density transitions, and anomalous multivariable deviations by combining these methods with real industrial SCADA records without relying on labeled failure data. The paper is structured as follows: First, Section 2 presents an overview of the industrial system and the SCADA dataset. The proposed multivariable methodology is outlined in Section 3. In Section 4, we report the findings from the studies. In turn, these results are discussed below, and their limitations are outlined in Section 5. Section 6 ends with reflections and recommendations for further research.

2. Industrial system and dataset description

This paper presents an operational study of an industrial SCADA-based fluid transportation system with centrifugal pumping units. Sensors and control devices are incorporated to register and log all operational factors in real time. Such information can be used to perform advanced analyses of system behavior and to enable data-driven monitoring. The apparatus comprises two centrifugal pumping devices situated at a multiproduct transport terminal. They are called Pump 1 and Pump 2. The instruments for recording hydraulic conditions, electrical parameters, and fluid characteristics are provided to each unit. All measurements were stored in a historical database for long-term system performance data. The dataset covers January 2024 to April 2025, a period of 15 months (uninterrupted operation), and contains 20,710 hourly entries. It consists of 35 variables (hydraulic parameters, electromechanical metrics, and fluid properties). Since centrifugal pumps can continuously operate under off-design conditions, we performed a preliminary preprocessing step to remove nonrepresentative operational records.

Specifically, observations of a zero volumetric flow rate were eliminated (Flow = 0). At this stage, a total of 15,049 operational records were retained, corresponding to 72.67% of the original dataset. This value was reduced to 14,611 after quality control, and further to 14,501 after removing incomplete and feature-inconsistent observations. Temporal resolution of the dataset remains a vital factor for meaningful interpretation. Even though the SCADA system measures data at a higher frequency, hourly data from daily operational reports were used for this analysis, meaning that the study was not designed to capture rapid dynamic events such as early-stage cavitation. However, the primary goal of this analysis is to look at operating patterns, steady-state anomalies, and the system's long-term behavior. As a result, the remaining dataset comprises 7,137 observations for Pump 1 and 7,364 for Pump 2 (the full set comprises 14,501 observations for a final multivariable analysis). The missing data accounted for less than 1% and were from sparse-sensor observations. Forward-fill interpolation was performed for such gaps to maintain temporality and, therefore, not have a strong impact on the system dynamics. We've chosen them because they represent hydraulic performance, energy use, and fluid properties of the pumping system. These include volumetric flow rate, suction pressure, discharge pressure, differential pressure (ΔP), electrical power, and fluid density. It explains the interaction between hydraulic behavior and energy demand under certain operating conditions. Table 1 summarizes the operational factors analyzed. Flow rate, pressure, electrical power consumption, and fluid density are important variables that provide insight into the system's behavior under different operating conditions. Together, they allow simultaneous identification of operational regimes and reveal product transitions and anomalous behaviors that reflect interactions among the hydraulic, electrical, and fluid parameters that single-variable methods cannot uncover. Vibration level was recorded from the original dataset, but was excluded from multivariable analysis due to an excessive number of missing values (>80%), which precluded its reliable use in the analytical framework

Table 1. Main recorded and selected operational variables considered during dataset screening

Category	Variable	Unit	Description
Hydraulic	Flow	BPH	Volumetric flow rate transported by the pumping system
Hydraulic	Suction Pressure	psi	Pressure at pump inlet
Hydraulic	Discharge Pressure	psi	Pressure at pump outlet
Derived	Differential Pressure (ΔP)	psi	Difference between discharge and suction pressure
Electrical	Power	kW	Electrical power consumed by the pump motor
Derived	Energy Index	(kW/BPH)	Energy consumption per unit of transported fluid
Fluid Property	Fluid Density	kg/m^3	Density of transported product
Mechanical	Vibration level	mm/s	Mechanical vibration of the pump system

3. Methodology

In this study, a multivariable data analytics approach is applied to analyze operational data obtained from an industrial SCADA monitoring system. The motivation for the approach is to uncover the dominant operating regimes, detect potential changes in the transported fluid, and identify anomalous behavior in the pumping system. The general analytical workflow in this study is shown in Figure 1, which summarizes the overall data processing and modeling. The analytical workflows comprise data preprocessing, feature engineering, exploratory statistical analysis, dimensionality reduction, unsupervised clustering for regime identification, probabilistic modeling for product regime detection, and robust anomaly detection.

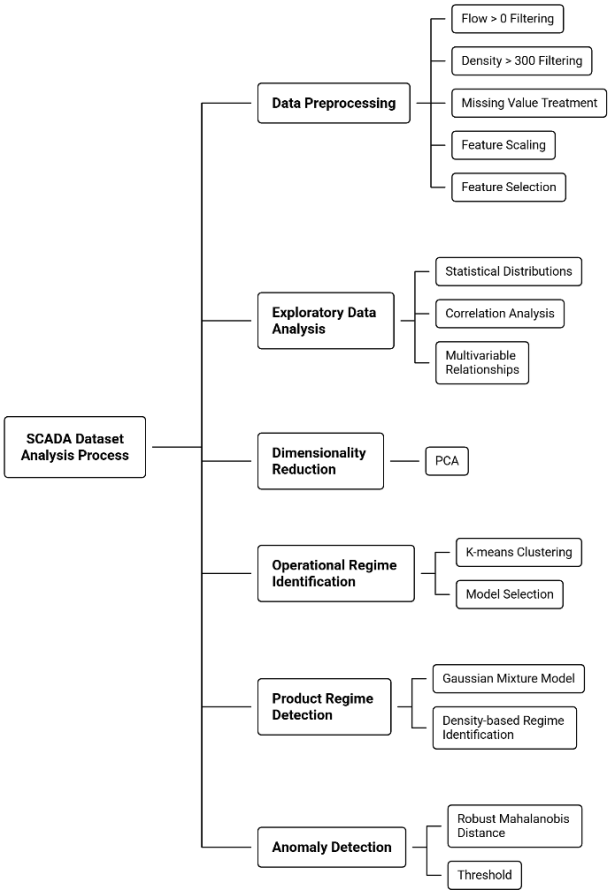

Figure 1. Methodological pipeline for the multivariable analysis of SCADA operational data

3.1 Data preprocessing

Data was analyzed using Python in Google Colaboratory with the Pandas, NumPy, SciPy, and Scikit-learn libraries. To maintain only representative operational records, the raw SCADA data were preprocessed and filtered according to the physical operation of the multiproduct pumping system. The defining operational periods were based on the Flow > 0 criterion- positive flow represents active fluid movement through the pumping system, and zero-flow records represent idle, shutdown, or non-transport conditions that do not represent hydraulic energy transfer. The data-quality threshold of Density > 300 kg/m^3 is used only as a conservative lower bound, but not as a product-classification criterion. A cursory analysis of the operational records revealed that the only numbers below this threshold were two isolated readings of 73 kg/m^3, occurring in both pumping units simultaneously, as well as records near the same product of nearly 722 kg/m^3. These isolated under-range values were consistent with SCADA acquisition artifacts, sensor-detected noise, scaling errors, or temporary under-range conditions at the transmitter. Not finding valid observations between 300 kg/m^3 and the lowest retained density value of 699.56 kg/m^3, the threshold removed only non-physical readings while remaining sufficiently distant from the valid product-density range. Therefore, this threshold did not distort the valid density distribution or perturb the subsequent characterization of the GMM-based product regime.

Almost no missing data were found in the records for the majority of the selected variables, with rates below 1%. Forward-fill temporal interpolation was thus applied when missing observations were very limited and temporally isolated in the hourly SCADA sequence. This method allows the story to continue and creates the chronological continuity, without introducing artificial trends or smoothing short-term operational drift. Normalization of Z-scores was then applied to numeric characteristics using Eq. (1), thereby enabling analysis of variables with different units and magnitudes in a unified feature space. Since this study is based on an unsupervised exploratory framework rather than a supervised predictive train-test procedure, normalization was applied to the final complete-case dataset for PCA, K-means, GMM, and robust Mahalanobis anomaly detection.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

where x represents the initial value, μ denotes the average, and σ signifies the standard deviation.

3.2 Feature engineering

Two derived variables were computed to enhance the physical interpretation of the system:

$$\Delta P = P_{\text{discharge}} - P_{\text{suction}} \quad (2)$$

$$EnergyIndex = \frac{Power}{Flow} \quad (3)$$

The pressure differential determines the hydraulic load, and the energy index is defined as energy use per unit of transported fluid. The obtained inputs account for the influence of hydraulic load on energy efficiency over different operating conditions. Other hydraulic parameters, including the Reynolds number and pump hydraulic efficiency, were not calculated because the SCADA dataset lacked the necessary physical characteristics, including fluid viscosity, pipe diameter, pump curves, and complete head-efficiency data. Consequently, feature engineering was limited to variables that could be consistently and reproducibly derived from available operational records.

3.3 Exploratory data analysis (EDA)

EDA was conducted to examine the statistical structure and relationships of the selected variables. The data were analyzed with univariate distributions, correlation matrices, and the most common correlations between hydraulic, electrical, and fluid variables to understand the data. Particular attention was paid to fluid density as a critical issue in multiproduct transport systems. Density contrasts indicate transitions between transported products and directly affect the hydraulic response and energy consumption of the pumping system.

3.4 Dimensionality reduction

PCA was used to reduce the dataset's dimensionality while retaining most of the variance. It maps high-dimensional correlated data to a lower-dimensional space of interpretable components and extracts the principal operating patterns. It also removes redundant information among correlated variables, facilitating clustering methods that fit the new, reduced feature space more efficiently. Since this model captures the major variance structure adequately in high-dimensional systems of industrial analysis for dimensionality reduction and pattern extraction, it has been prevalent in industrial monitoring [28]. The standardized data covariance matrix is as follows:

$$\Sigma = \frac{1}{n-1} X^T X \quad (4)$$

Eigenvalue decomposition was then performed:

$$\Sigma v_i = \lambda_i v_i \quad (5)$$

where λ_i describes the eigenvalues, and v_i stands for the related eigenvectors.

Principal components were chosen based on their cumulative explained variance. For the purposes of the present investigation, we retained the initial components that explained most of the variance to represent the system's multivariable operational structure more directly. The streamlined representation was then used for clustering and anomaly detection. The mathematical structures employed in this work are also based on common definitions reported commonly in multivariate statistical analysis and machine learning literature [28,29].

3.5 Identification of operational regimes (K-means)

The operational regimes were determined by K-means clustering in a PCA-transformed environment. K-means was also selected because of its ability to discover compact, distinct clusters in low-dimensional feature spaces. Both the silhouette score and the Davies–Bouldin index were used to assess cluster quality and determine the optimal number of clusters. This procedure allows the identification of distinct operational regimes based on the data's multivariable nature, without relying on preset thresholds or arbitrarily defined operating conditions. Also, K-means focuses on minimizing the within-cluster sum of squares (WCSS):

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \quad (6)$$

where C_i shows cluster i and μ_i represents its centroid.

At k=2, the optimal solution is two dominant operational regimes with different combinations of hydraulic variables, energy consumption, and fluid properties.

3.6 Product regime modeling (GMM)

Using fluid density in a GMM, product regimes have been identified. It was selected as a probabilistic approach that is flexible for modeling multimodal distributions; thus, it is also applicable to different product regimes in multiproduct systems. This approach characterizes the underlying density distribution rather than relying on fixed threshold values, making it a useful method for detecting product transitions. In contrast to deterministic clustering methods, GMM accounts for membership uncertainty during class classification and is particularly suitable for industrial processes that undergo gradual changes [29]. The model assumes that the data are generated from a mixture of Gaussian distributions:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \quad (7)$$

where π_k are the mixture weights, μ_k the means, and Σ_k the covariance matrices.

One to five Gaussian components GMM configurations were assessed by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Models were fitted with the Expectation-Maximization (EM) algorithm with full covariance matrices, and convergence was verified before model selection. Although additional components were evaluated, the two-component GMM was chosen because it yielded the most physically interpretable description of the monitored multiproduct system consistent with the two dominant density regimes observed in the operational data. This probabilistic formulation enabled density-based product regimes and potential product-transition zones to be defined without relying on fixed product thresholds.

3.7 Anomaly detection (Robust Mahalanobis distance)

Anomaly detection using a complete Mahalanobis distance formulation was applied to the full dataset for the entire structure across the variables. The distance approach and covariance-aware approach are typically adopted for data-driven process monitoring and fault diagnosis, as they identify joint deviations among closely connected variables [28]. Mahalanobis distance is:

$$D_M(\mathrm{x}) = \sqrt{(\mathrm{x} - \mu)^T \Sigma^{-1} (\mathrm{x} - \mu)} \quad (8)$$

where x represents the observation vector, μ the mean vector, and Σ the covariance matrix of the dataset.

Our estimated covariance matrix Σ for outlier-contamination resilience was obtained using the Minimum Covariance Determinant (MCD) estimator, an effective covariance estimator for improving robustness to outlier contamination. The anomaly threshold, defined as the 99.5th percentile of the robust Mahalanobis distance distribution, represents the upper 0.5% tail of multivariable deviations. A conservative percentile-based threshold was chosen to detect only extreme deviations from the normal covariance structure, with the intention of avoiding excessive false positives in an unlabeled industrial dataset. Observations above this threshold were classified as anomalous candidates.

4. Results

4.1 Operational dataset characterization and quality filtering

The examined dataset includes 20,710 SCADA records, of which 15,049 correspond to active operational conditions (72.67%), enabling comprehensive analysis under real industrial conditions (Table 2).

Table 2. Dataset overview

Characteristic	Value
Total records	20,710
Operational records	15,049
Operational percentage (%)	72.67
Monitoring duration (months)	15
Temporal resolution	1 hour
Number of variables	35
Pumping units	2

A quality control (QC) operation was then carried out to verify the reliability of the data and incomplete observations were removed to ensure consistency for the data-driven analysis. Therefore, the remaining dataset for PCA, clustering and anomaly detection was 14,501 records in total. The temporal distribution of operational records (Figure 2) corroborates a high coverage of the monitoring during the study period, with no significant gaps resulting in a bias to the analysis. While we found missing variables, most highly relevant variables with missing values were still <1%, an acceptable value for industrial SCADA systems. Variables with excessive amount of missing data (>80%) were removed as they did not contribute to the operational dynamics of interest.

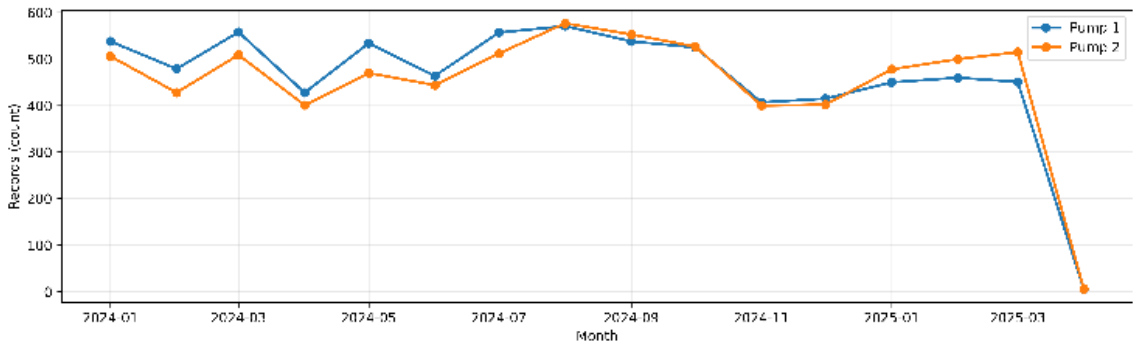

Figure 2. Temporal distribution of operational records across the monitoring period

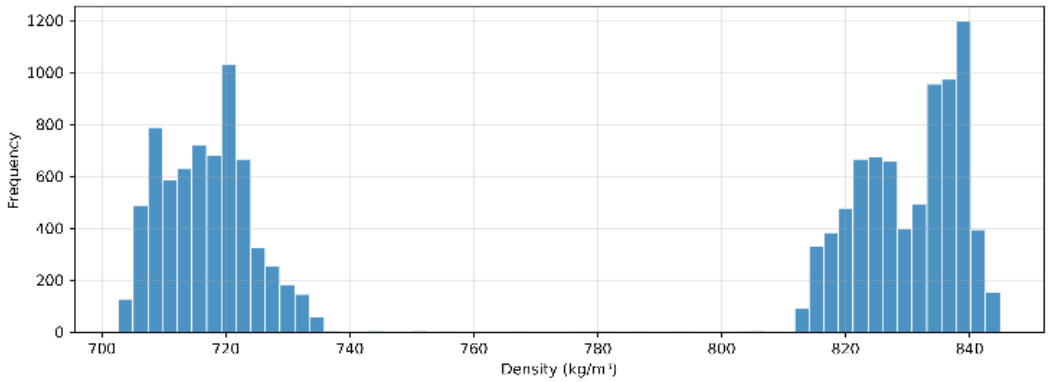

Figure 3. Fluid density distribution during operational periods, showing a bimodal pattern associated with low- and high-density transported product regimes

A stepwise filtering and quality-control procedure was applied to ensure the consistency and reproducibility of the dataset used for multivariable analysis. Table 3 summarizes the complete data flow from the original SCADA records to the final complete-case dataset used for PCA, K-means clustering, GMM-based product regime characterization, and robust Mahalanobis anomaly detection. Thus, 14,611 records correspond to the QC-passed operational dataset, whereas 14,501 complete observations were retained as the final multivariable dataset used for PCA, K-means clustering, GMM-based product regime characterization, and robust Mahalanobis anomaly detection. The density distribution of the final dataset shows a bimodal structure, suggesting the presence of low- and high-density regimes of transported product (Figure 3).

A descriptive statistical analysis was then performed to characterize the distributional behavior of the main operational variables retained for multivariable modeling. Table 4 summarizes central tendency, dispersion, extreme values, skewness, and kurtosis through the mean, median, standard deviation, minimum, and maximum of the principal hydraulic, electrical, and fluid-related variables. The high skewness and kurtosis values for flow, discharge pressure, and power suggest highly asymmetric distributions with extreme operational values, as is commonly found in industrial SCADA datasets. The difference between mean and median for flow, discharge pressure, and power additionally indicates the effect of extreme values on central tendency. Fluid density differs from this figure, with low skewness and negative kurtosis, consistent with a bounded bimodal distribution associated with transported product regimes.

Such distributional characteristics support the use of robust multivariable methods for operational regime identification and anomaly detection.

Table 3. Dataset filtering and quality-control flow

Stage	Criterion applied	Records retained	Records removed	Purpose
Raw SCADA dataset	Original hourly records	20,710	—	Complete historical dataset
Operational filtering	Flow > 0	15,049	5,661	Removal of idle or non-pumping states
Quality control filtering	Physical and percentile-based filters	14,611	438	Removal of non-physical or unreliable records
Complete-case multivariable subset	Removal of incomplete and feature-inconsistent observations	14,501	110	Final dataset used for PCA, K-means, GMM, and anomaly detection

Table 4. Descriptive statistics of main operational variables

Variable	Mean	Median	Std	Min	Max	Snew.	Kurt.
Flow (BPH)	883.95	921.83	212.78	50.00	11007.00	18.22	832.13
Discharge pressure (psi)	1584.19	1620.52	326.80	4.75	17399.00	13.03	734.52
Density (kg/m^3)	778.43	817.42	57.12	699.56	848.17	-0.16	-1.90
Power (kW)	496.35	504.36	116.59	-33.40	4752.00	3.87	215.44

4.2 Multivariable structure of the system (PCA)

Figure 4 presents a correlation between the variables that indicates strong correlation between both hydraulic and energy quantities (discharge pressure, pressure differential, and power consumption). This phenomenon is in agreement with the physical behavior of centrifugal pumping systems. In a quantitative manner, the highest positive associations detected were the discharge pressure and pressure differential (r = 0.98), the power consumption and pressure differential (r = 0.88), and the power consumption and discharge pressure (r = 0.86). Moreover, flow rate and energy index are negatively related (r = −0.50), as predicted on the operating curve of centrifugal pumps, where energy consumption per unit volume decreases as flow rate increases. Such results confirmed the co-modelled hydraulic-energy properties of the pumping system; that is, increased hydraulic load is accompanied by increased electrical power demand.

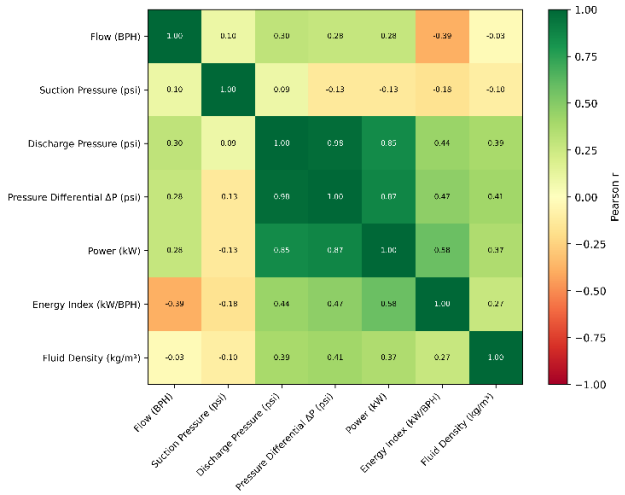

Figure 4. Correlation matrix of operational variables

PCA applied to the final complete-case multivariable dataset (14,501 observations) showed that the system can be characterized efficiently in a much smaller dimensional space, as summarized in Table 5. The first two principal components (PC1 and PC2) account for 69.80% of the total variance, while the inclusion of PC3 raises the cumulative explained variance to 83.32%. To enhance the physical interpretation of the PCA space, the loading matrix of the first three principal components was analyzed. The contribution of each original and constructed factor to the principal components is depicted in Table 6, such that the reduced physical space is related to hydraulic, energetic, and fluid phenomena.

Table 5. PCA explained variance

Component	Variance (%)	Cumulative (%)
PC1	48.81	48.81
PC2	20.99	69.80
PC3	13.52	83.32
PC4	11.09	94.41
PC5	4.01	98.43
PC6	1.58	100.00

Table 6. PCA loading matrix for the first three principal components

Variable	PC1 loading	PC2 loading	PC3 loading
Flow (BPH)	0.123	0.717	-0.336
Suction Pressure (psi)	-0.070	0.378	0.899
Discharge Pressure (psi)	0.508	0.178	0.149
Pressure Differential ΔP (psi)	0.521	0.094	-0.049
Power (kW)	0.508	0.041	-0.043
Energy Index (kW/BPH)	0.333	-0.523	0.228
Fluid Density (kg/m^3)	0.286	-0.165	0.002

The loading order reveals that PC1 is primarily related to hydraulic load and energy demand, as discharge pressure, pressure differential, and power are the most positive contributions. Thus, PC1 can be seen as the hydraulic or energy axis that captures the pumping mode. In comparison, PC2 is primarily affected by flow and the energy index on the opposite side, indicating that this measure captures differences between transport flow and energy usage for a given unit volume. PC3 primarily relates to the suction pressure, which denotes the inlet-side pressure variation. These findings support the view that PCA is not only a dimensionality reduction but also an interpretable simulation of the coupled hydraulic/energetic behavior of the pumping system. The first two principal components explain 69.80% of the total variance, suggesting that the dominant multivariable structure of the pumping system can be captured in 2D. This representation is physically interpretable since the loading matrix relates PC1 primarily to hydraulic load and energy demand via discharge pressure, pressure differential, and

power, while PC2 relates mainly to the correlation between transported flow and energy consumption for each volume. While the other components have secondary variability to preserve, the PC1–PC2 plane offers the operational space to show dominant regimes and peripheral conditions. In the cumulative variance curve shown in Figure 5, the first three components increase the explained variance to 83.32%, whereas the first four components increase the explained variance to 94.41%. As a result, the residual variance was considered complementary rather than disregarded, and the PC1–PC2 projection was kept for visualization and clustering due to its clearer physical interpretation and operational separation. A PCA representation of this closely packed superimposing operational core, along with the peripheral conditions appropriate to atypical measurements, illustrates this reduced representation (Figure 6). The projection of Figure 6 illustrates the separation of the dominant operational core and peripheral atypical conditions in the 2D PCA space.

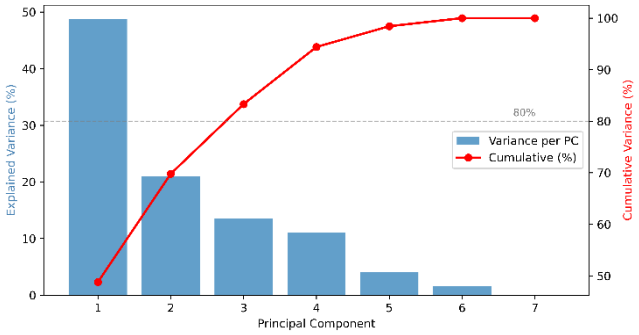

Figure 5. Explained variance of principal components

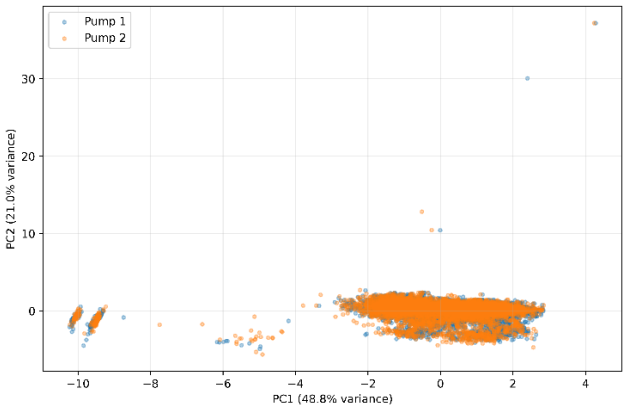

Figure 6. Projection of operational data in PCA space

4.3 Identification of operational regimes

Clustering analysis with K-means revealed two distinct operational conditions: a dominant regime and a low-load condition (Table 7). Silhouette analysis was conducted to assess the separation quality between the operational clusters in PCA space. Figure 7 presents the highest silhouette score for k = 2 (0.7909), indicating the best separation and cohesion among all the evaluated configurations. This finding indicates that the pumping system may be occupied by two dominant operational regimes. An elbow approach was also used to analyze the WCSS behavior in different clustering configurations. A pronounced decrease in inertia is observed up to k = 2, followed by a more gradual decrease for larger values of k (Figure 8). This shows that two clusters strike a balance between the model's simplicity and the operational variability of clustering.

Table 7. Cluster validation metrics

Metric	Value
Optimal number of clusters	2
Silhouette score	0.7909
Davies–Bouldin index	0.2943
Calinski–Harabasz index	9,425.45

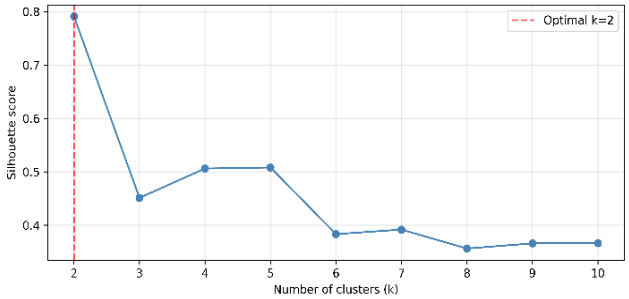

Figure 7. Silhouette score for K-means clustering in PCA space

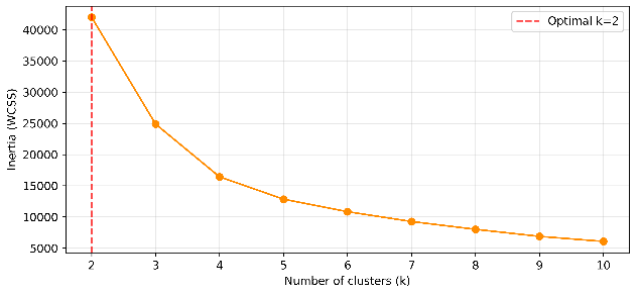

Figure 8. K-means clustering elbow curve

The silhouette score, Davies–Bouldin index, and elbow criterion collectively support the choice of k = 2, suggesting a tight, well-separated, and physically interpretable partition of the PCA-constrained operational space. This matching between the validation scores minimizes the burden of the clustering decision on a single internal measure. To further strengthen the robustness, K-means was compared with hierarchical clustering and DBSCAN, within the same PCA-reduced operating space. Two clusters were similarly shaped by K-means and hierarchical clustering, with silhouette coefficients of 0.7909 and 0.7903, and Davies–Bouldin indices of 0.2943 and 0.3003, respectively, indicating the stability of a two-regime partition. DBSCAN also identified two clusters, yielding a slightly higher silhouette score of 0.8032 and a lower Davies–Bouldin index of 0.2086, but assigned 76 observations to noise, suggesting its greater sensitivity to density and parameter selection. Consequently, K-means persisted as the primary clustering strategy because it produced a complete two-regime partition, consistent with hierarchical clustering and directly interpretable from the hydraulic and energetic cluster profiles. A high silhouette value indicates strong separation between clusters, and a low Davies–Bouldin index suggests compact group structures.

The distribution of observations across clusters (Table 8) is highly unbalanced. These clusters confirm the presence of a dominant operational regime and a secondary low-load condition (Table 9). As shown in Figure 9, Cluster 0 represents the dominant pumping regime (97.75%, n = 14,175), characterized by high flow rates, elevated pressure differential (mean ΔP = 1,271.63 psi), and substantial energy consumption, corresponding to normal operating conditions. In contrast, Cluster 1 accounts for only 2.25% of observations (n = 326) and exhibits an almost negligible pressure differential (ΔP = 2.66 psi versus 1,271.63 psi in the dominant regime) and substantially reduced power consumption, indicative of a low-load or near-idle operating state. Some observations within this cluster exhibit significant flow fluctuations and physically inconsistent measurements, suggesting both real operational disturbances and potential sensor artifacts.

Table 8. Cluster distribution

Cluster	Records	Percentage (%)
0	14,175	97.75
1	326	2.25

Table 9. Mean profile clusters

Cluster	Flow	ΔP	Power	Energy Index	Density
0	898.14	1271.63	509.26	0.597	778.02
1	577.05	2.66	60.70	0.107	781.74

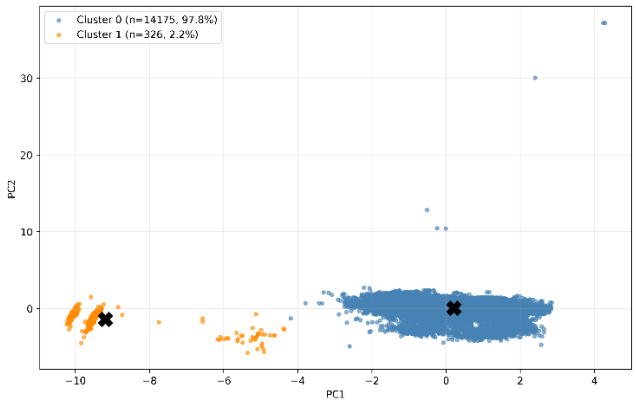

Figure 9. K-means projection of operational data in PCA space

4.4 Product regime characterization

A Gaussian Mixture Model (GMM) was applied to the fluid density variable to identify dominant product-related operational regimes within the multiproduct pumping system. To determine the optimal number of Gaussian components, model selection was performed using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). As shown in Figure 10, both criteria exhibited a substantial reduction when moving from one to two Gaussian components, indicating that the density distribution is best represented by two dominant probabilistic regimes. Additional components produced only marginal improvements, suggesting that a two-component GMM provides an adequate and physically interpretable representation of the transported products within the operational dataset. The numerical AIC/BIC evaluation for one to five Gaussian components showed the largest improvement when moving from one to two components. Although AIC and BIC continued decreasing for models with additional components, the two-component configuration was retained because it provided the most physically interpretable representation of the monitored multiproduct system, corresponding to the two dominant density regimes observed in the operational data. Additional components mainly subdivided the same physical product-density regimes and therefore did not improve the system's operational interpretation. A GMM for fluid density revealed two dominant regimes:

- Low-density regime: 716.84 kg/m^3
- High-density regime: 830.35 kg/m^3

Both density regimes exhibit well-separated central tendencies and ranges, with no significant overlap in their mean, median, or interquartile ranges (Table 10). These results show that fluid density is a reliable discriminator for product regime classification in this system.

Table 10. Density regime summary

Metric	Low-density regime	High-density regime
Number of observations	6,670	7,831
Mean density (kg/m^3)	716.84	830.35
Standard deviation (kg/m^3)	7.68	8.62
Minimum (kg/m^3)	699.56	772.26
25th percentile (kg/m^3)	710.72	823.79
Median (kg/m^3)	716.74	832.41
75th percentile (kg/m^3)	721.50	837.59
Maximum (kg/m^3)	770.00	848.17

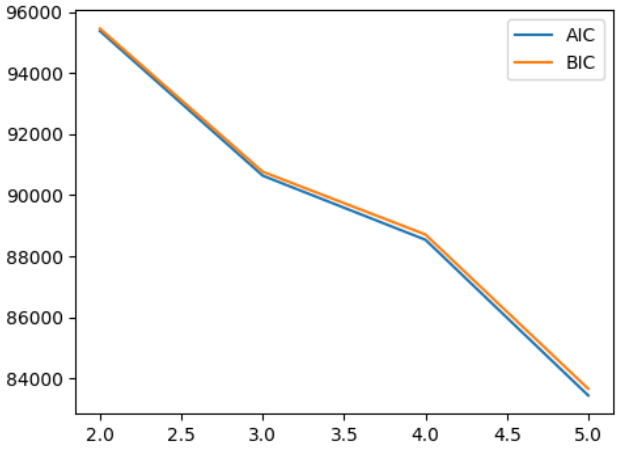

Figure 10. AIC and BIC scores for Gaussian Mixture Model selection

The density distribution in Figure 11 is consistent with this, with a distinct bimodal pattern and minimal overlap across regimes. This establishes two distinct product states, each with specific density ranges. Such behavior is consistent with multiproduct transport systems, which show that different fluids have distinct physical properties.

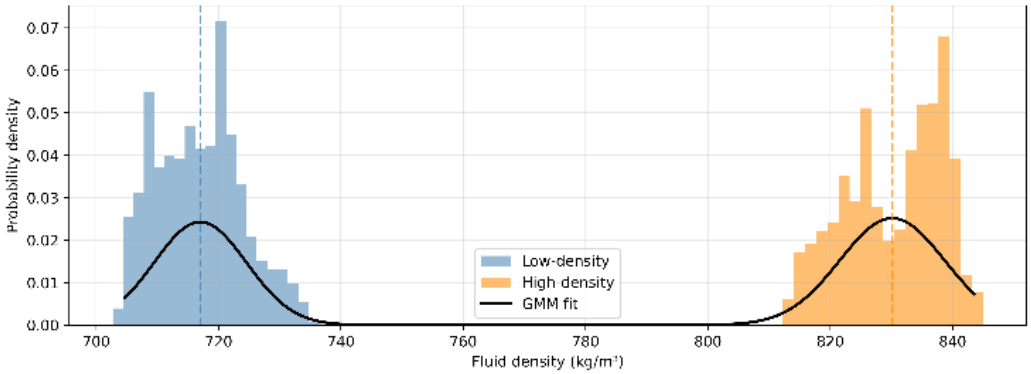

Figure 11. Fluid density distribution by product regime (GMM)

This operational characterization directly informs the system's internal behavior. The density classification successfully identifies two stable operating states, in which fluid density reliably indicates the type of transported product.

4.5 Detection of product interfaces

Using a multivariable interface detection method for the six operational variables, we identified 65 transition events across both pumping units with a detection threshold set at the 99.5th percentile of the interface score distribution (Table 11). The observed interface events are seen spatially separated from the normal operational cloud in PCA space, located largely on the periphery of both density regimes (Figure 12). This spatial separation verifies that product transitions are multivariable deviations from steady-state operating conditions rather than simple independent variations in a single variable. The similar numbers of monitoring event occurrences (31 and 34) in both pumps indicate that, during the observed period for the same pipeline system, they experienced roughly equal numbers of product transitions.

Table 11. Interface detection summary

Pump	Records	Events	Threshold
Pump 1	7,137	31	10.55
Pump 2	7,364	34	10.07

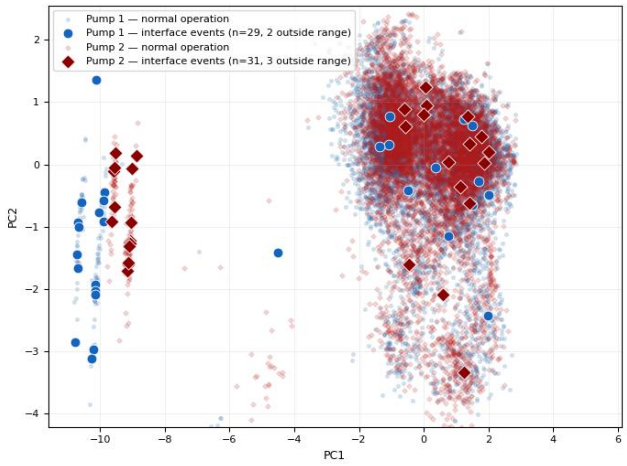

Figure 12. Interface events identified in PCA space- Pump 1 and Pump 2

4.6 Detection of operational anomalies

A multivariable anomaly detection algorithm based on robust Mahalanobis distance was developed to identify strong deviations from expected system behavior. The analysis was performed using the 14,501 valid operational records from the SCADA dataset. For the per-pump anomaly summary, observations were classified as anomalous whenever their robust Mahalanobis score exceeded the 99.5th percentile of the corresponding pump-specific score distribution. The number of anomalies detected for each pumping unit is summarized in Table 12. Seventy-three abnormal observations were noticed, representing about 0.5% of the operational dataset. This low share would characterize continuous system operations over the majority of the monitoring time. However, these anomalies are not randomly distributed. They instead display specific operational patterns. Table 13 shows a representative comparison of normal and anomalous conditions.

Table 12. Per-pump anomaly detection summary using robust Mahalanobis distance

Pump	Records	Anomalies	Threshold
Pump 1	7,137	37	88.10
Pump 2	7,364	36	97.11

Table 13. Mean feature shift between normal operation and detected anomalies

Variable	Mean (Normal)	Mean (Anomalies)	Shift (Anom – Normal)
Flow (BPH)	890.55	964.30	+73.75
Suction Pressure (psi)	357.86	312.34	−45.52
Discharge Pressure (psi)	1604.36	882.46	−721.90
Pressure Differential ΔP (psi)	1246.51	570.12	−676.39
Power (kW)	500.42	254.06	−246.36
Energy Index (kW/BPH)	0.5700	0.5300	−0.0400
Fluid Density (kg/m^3)	777.98	802.56	+24.58

The anomalies detected show the following multivariable pattern (Table 13):

- sudden drops in discharge pressure (−721.90 psi).
- considerable decreases in pressure differential (−676.39 psi).
- a large reduction in power consumption (−246.36 kW).
- moderate variation of the flow rate.
- significant changes in fluid density.

Although detailed maintenance or operational event logs were not available to confirm each detected event, the anomaly pattern is physically consistent with degraded hydraulic operating conditions. The simultaneous reduction in discharge pressure, pressure differential, and power consumption indicates that the observed deviations are coordinated multivariable rather than isolated single-sensor fluctuations. Therefore, the operational validation of these anomalies is based on hydraulic consistency and cross-method agreement with Isolation Forest and LOF, rather than direct event labeling. Accordingly, the detected observations should be interpreted as physically consistent multivariable anomaly candidates rather than confirmed failure events. Mahalanobis scores were further analyzed to assess the reliability of this recommended technique (Figure 13), where the scores are highly right-skewed and bunched at lower values, with anomalies leading to a long tail extending to greater distances.

The per-pump thresholds in Table 12 (88.10 for Pump 1; 97.11 for Pump 2) reflect the individual score distributions, whereas Table 14 reports a global sensitivity analysis using the pooled distribution across all 14,501 records. At the 99.0th percentile, 145 anomalies were detected (1.000%); at the selected 99.5th percentile, 73 observations were identified (0.503%); at the 99.7th and 99.9th percentiles, detections dropped to 44 and 15, respectively. This monotonic reduction confirms that the 99.5th percentile provides a conservative balance between capturing relevant multivariable deviations and avoiding excessive false positives in an unlabeled SCADA dataset.

As an additional robustness check, the robust Mahalanobis results were compared with Isolation Forest and Local Outlier Factor (LOF). The three unsupervised methods identified the same number of anomalous observations, 73 records, supporting the stability of the detection results under different algorithmic assumptions. However, the robust Mahalanobis approach was retained as the main anomaly detection method because it explicitly exploits the covariance structure among hydraulic, electrical, and fluid-related variables. This provides a physically interpretable anomaly score that can be directly related to coordinated multivariable deviations in industrial monitoring. The primary quantitative measurements extracted at each level of analysis are presented in Table 15.

Table 14. Global sensitivity analysis of robust Mahalanobis anomaly thresholds

Percentile threshold	Mahalanobis threshold	Detected anomalies	Anomaly ratio (%)
99.0	88.014	145	1.000
99.5	93.935	73	0.503
99.7	96.990	44	0.303
99.9	104.078	15	0.103

Table 15. Quantitative summary of key results

Indicator	Value	Unit
Optimal number of clusters	2	clusters
Silhouette score	0.7909	-
Davies–Bouldin index	0.2943	-
Total anomalies	73	records
Anomaly ratio	0.503	%

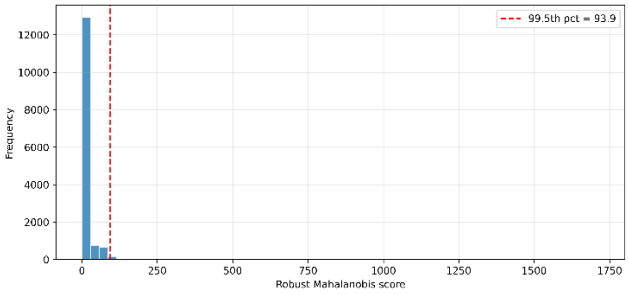

Figure 13. Distribution of robust Mahalanobis distance scores for anomaly detection

Overall, these quantitative results show that the proposed framework consistently identifies operational regimes, product-related density states, interface events, and a small subset of physically interpretable anomaly candidates within the final multivariable dataset. These findings indicate that the dataset is well-organized, covers different operating regimes, and has an almost negligible rate of anomalies. These results confirm that the proposed framework reliably identifies meaningful multivariable deviations in industrial pumping systems, providing a physically interpretable and reproducible basis for condition monitoring.

5. Discussion

The findings presented in Section 4 show that the proposed framework provides an integrated representation of the pumping system's operational structure. By jointly analyzing hydraulic, electrical, and fluid-related variables, the framework identifies dominant operating regimes, density-driven product regimes, product-interface events, and anomalous multivariable deviations that would be difficult to detect using isolated single-variable thresholds. Most important among the results from the work presented in this report is the multimodal distribution of fluid density, reflecting two predominant product-state characteristics across the multiproducts. The GMM detected density centers at 716.84 kg/m^3 and 830.35 kg/m^3, confirming the presence of two well-separated transported-product regimes. For multiproduct pipeline systems, the variations in density are commonly related to batch transfers or changes in the transported fluids, therefore density can provide information about the product changes and alterations in the operating regime. Similarly, in previous industrial process monitoring studies, results on fluid properties and multivariable interactions were used to identify latent process states [9,11]. Moreover, Gaussian-mixture based modelling is appropriate for modelling multimode behavior and latent operating states in cases where regime changes may be gradual or partially overlapping [29].

PCA indicates substantial covariance among hydraulic parameters, electricity use, and fluid characteristics. The results show that the first and second principal components account for 69.8% of the total variance, indicating a strong physical interaction between hydraulic loads and energy consumption in centrifugal pumping systems. PCA is widely used for dimensionality reduction in process monitoring, as it finds the main variation structure and identifies the most prevalent operating patterns. For example, Liu et al. [9] have demonstrated that multivariable feature representations can indeed reflect the industrial system's operational structure, leading to improved anomaly detection performance. Similarly, Du et al. [5] demonstrated that feature extraction and dimensionality reduction approaches offer significant improvements in data-driven monitoring over traditional solutions used in Industrial Cyber-Physical Systems.

K-means clustering applied in the PCA-reduced space yielded two well-separated clusters, Silhouette = 0.7909, Davies–Bouldin Index = 0.2943. The resulting clusters correspond to unique combinations of hydraulic parameters, energy supply, and fluid composition, enabling a clear separation between the dominant operating regime and a secondary low-load condition. It is important to identify such a structure for condition monitoring, as it serves as a baseline against which deviations and discrepancies can be established. Unsupervised clustering algorithms have been effectively used to discover operational states and deviations from normality in data-driven analysis. Mokhtari et al. [3] show that it is feasible to identify operation regimes from industrial data through clustering analysis, unless the labeled entity is not yet present. Similarly, Kim et al. [6] showed that machine learning methods are superior to threshold-based classical monitoring methods for complex industrial systems that have many interacting states.

The comparison between hierarchical clustering and DBSCAN reinforces the choice of K-means as the primary regime-identification method. Hierarchical clustering produced a nearly equivalent two-regime structure, and DBSCAN yielded slightly improved internal separation metrics but classified 76 observations as noise. Thus, K-means was kept because it yielded a complete, stable, and physically interpretable two-regime partition that closely aligned with the hydraulic and energetic cluster profiles. Robust Mahalanobis distance-based anomaly detection performed on the recorded data identified 73 anomalous observations, equivalent to about 0.5% of the data. Substantial reductions of discharge pressure (−721.90 psi), pressure differential (−676.39 psi), and power consumption (−246.36 kW) were observed, consistent with hydraulic load loss conditions associated with such events. Distance-based detection methods are best applied to unlabelled datasets that lack an a priori signature of failure, as they leverage the full covariance structure across variables to identify joint deviations that univariate thresholds cannot detect. Seo et al. [30] showed that interpretable anomaly-detection techniques can be implemented in SCADA systems where operational transparency and explainability are essential for industrial decision-making. Likewise, Goetz and Humm [12] also showed that data-driven anomaly detection methods can robustly identify anomalous behavior in cyber-physical industrial systems.

Overall, the proposed framework is a useful addition to industrial monitoring, as it allows performing an operational dynamics, product cycle, and anomalous behavior analysis in SCADA data together in the joint operation in an interpretable multivariable and unsupervised manner. This is consistent with the Industry 4.0 requirements, as it facilitates reliability enhancement, predictive maintenance, and transparent decision-making in industrial infrastructures [1]. However, this study has limitations. The results of the study were obtained for only one industrial installation, which limits generalization to other pumping stations, pipeline configurations, and processing environments. Subsequently, cross-site validation is mandatory for future activities to assess the proposed framework's stability across various operating scenarios. The second is that while the framework can pinpoint operating regimes and anomalous conditions, it doesn't directly diagnose root causes. For root cause analysis, it is necessary to include analysis of other vibration and temperature signals to improve fault mode analysis.

From an implementation point of view, the computational complexity is low due to the incorporation of PCA, K-means, GMM, and distance-based anomaly scoring in the proposed platform, and the proposed method can be optimized on historical SCADA data and/or incorporated into periodic monitoring operations. Therefore, the proposed methodology can be integrated into industrial SCADA historians or predictive maintenance dashboards as a decision-support layer within the framework. On the ground, the framework could also periodically reconfigure operational regimes, detect density-related product transitions, and detect multivariable deviations for engineering review. As such, the framework can be deployed as an offline or near-real-time monitoring layer without labeled failure records or computationally-heavy deep learning infrastructure. Although a percentile-threshold sensitivity check was incorporated and the robust Mahalanobis approach was supported by agreement with Isolation Forest and LOF, future work should evaluate the sensitivity of anomaly detection results to the contamination parameter used in robust covariance estimation and compare the robust Mahalanobis distance with the classical covariance-based Mahalanobis formulation.

Another limitation is the hourly sampling resolution of historical SCADA records used for analysis. While this temporal resolution is sufficient to detect long-term operational regimes, density-related product transitions, and steady-state anomalous conditions, it is insufficient for rapid transient phenomena such as incipient cavitation, short-duration hydraulic instabilities, or fast sensor disturbances. In the future, we propose to include higher-frequency SCADA data and temporal modeling extensions, including, but not limited to, Hidden Markov Models, LSTM-based architectures, or transformer-based models [4] to capture sequential transitions and transient operating patterns. These extensions would help assess both the framework's generalizability and its ability to better differentiate among product transitions, operational disturbances, and early-stage degradation mechanisms.

6. Conclusions

This paper introduces a multivariable analytical framework based on SCADA data for analysis of an industrial centrifugal pumping system. Using statistical and unsupervised learning methods under real industrial conditions, the proposed approach can identify operational regimes, detect product transitions, and recognize anomalous conditions. The results showed that the system's behavior can be quantified using a relatively small set of parameters that capture the interplay among hydraulic, electrical, and fluid properties. K-means clustering demonstrated two prevailing operating regimes as indicated by a Silhouette score of 0.7909

and Davies–Bouldin index of 0.2943. Regarding fluid density, the GMM identified two product regimes, with mean values of 716.84 kg/m^3 and 830.35 kg/m^3, consistent with the system's multiproduct character. Furthermore, the Mahalanobis distance-based anomaly detection returned 73 anomalous observations (0.5% of the dataset), with significant reductions of discharge pressure (−721.90 psi), pressure differential (−676.39 psi), and power consumption (−246.36 kW), indicating degraded operating conditions. The robustness of the proposed approach was further substantiated by comparison with the Isolation Forest and LOF methods, which yielded consistent anomaly detection results across all three techniques. These findings confirm that the framework reliably identifies meaningful deviations in complex industrial systems. Collectively, the proposed methodology offers a reliable and interpretable approach to data-driven monitoring of complex pumping systems using SCADA data, consistent with Industry 4.0 (I4.0) principles. The primary limitations of this study are the use of data from a single industrial installation and the hourly sampling resolution of the historical SCADA records. Therefore, future work should validate the framework in other pumping stations and incorporate higher-frequency data to improve the detection of transient operating events.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] D. Leite, E. Andrade, D. Rativa, and A. M. A. Maciel, “Fault Detection and Diagnosis in Industry 4.0: A Review on Challenges and Opportunities,” Sensors, vol. 25, no. 1, 2025, doi: 10.3390/s25010060.

[2] T. K. Das, S. Adepu, and J. Zhou, “Anomaly Detection in Industrial Control Systems Using Logical Analysis of Data,” Computers & Security, vol. 96, 2020, doi: 10.1016/j.cose.2020.101935.

[3] S. Mokhtari, A. Abbaspour, K. K. Yen, and A. Sargolzaei, “A Machine Learning Approach for Anomaly Detection in Industrial Control Systems Based on Measurement Data,” Electronics, vol. 10, no. 4, 2021, doi: 10.3390/electronics10040407.

[4] E. Dong, X. Zhan, H. Yan, Y. Bai, R. Wang, and Z. Cheng, “A data-driven intelligent predictive maintenance decision framework for mechanical systems integrating transformer and kernel density estimation,” Computers & Industrial Engineering, vol. 201, 2025, doi: 10.1016/j.cie.2025.110868.

[5] Y. Du, Y. Huang, G. Wan, and P. He, “Deep Learning-Based Cyber–Physical Feature Fusion for Anomaly Detection in Industrial Control Systems,” Mathematics, vol. 10, no. 22, 2022, doi: 10.3390/math10224373.

[6] M.-C. Kim, J.-H. Lee, D.-H. Wang, and I.-S. Lee, “Induction Motor Fault Diagnosis Using Support Vector Machine, Neural Networks, and Boosting Methods,” Sensors, vol. 23, no. 5, 2023, doi: 10.3390/s23052585.

[7] C. Han and G. Gim, “Time-Series-Based Anomaly Detection in Industrial Control Systems Using Generative Adversarial Networks,” Processes, vol. 13, no. 9, 2025, doi: 10.3390/pr13092885.

[8] J. Cho and S. Gong, “Dynamic data abstraction-based anomaly detection for industrial control systems,” Electronics, vol. 13, no. 1, p. 158, 2024, doi: 10.3390/electronics13010158.

[9] J. Liu, Y. Sha, W. Zhang, Y. Yan, and X. Liu, “Anomaly Detection Method for Industrial Control System Operation Data Based on Time–Frequency Fusion Feature Attention Encoding,” Sensors, vol. 24, no. 18, 2024, doi: 10.3390/s24186131.

[10] M. M. Aslam, L. C. D. Silva, R. A. A. H. M. Apong, and A. Tufail, “An Optimized Anomaly Detection Framework in Industrial Control Systems Through Grey Wolf Optimizer and Autoencoder Integration,” Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-12775-0.

[11] J. Pang, X. Pu, and C. Li, “A Hybrid Algorithm Incorporating Vector Quantization and One-Class Support Vector Machine for Industrial Anomaly Detection,” IEEE Transactions on Industrial Informatics, vol. 18, no. 12, pp. 8786-8796., 2022, doi: 10.1109/TII.2022.3145834.

[12] C. Goetz and B. G. Humm, “Process Anomaly Detection in Cyber–Physical Production Systems Based on Conditional Discrete-Time Dynamic Graphs,” Applied Sciences, vol. 15, no. 21, 2025, doi: 10.3390/app152111354.

[13] L. Xu, K. Shang, X. Zhang, C. Zheng, and L. Pan, “Multi-Scale Feature Fusion-Based Real-Time Anomaly Detection in Industrial Control Systems,” Electronics, vol. 14, no. 8, 2025, doi: 10.3390/electronics14081645.

[14] B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and H. Kim, “A Comparative Study of Time Series Anomaly Detection Models for Industrial Control Systems,” Sensors, vol. 23, no. 3, 2023, doi: 10.3390/s23031310.

[15] S. Kim et al., “Two-Phase Industrial Control System Anomaly Detection Using Communication Patterns and Deep Learning,” Electronics, vol. 13, no. 8, 2024, doi: 10.3390/electronics13081520.

[16] A. L. Alfeo, M. G. C. A. Cimino, G. Manco, E. Ritacco, and G. Vaglini, “Using an Autoencoder in the Design of an Anomaly Detector for Smart Manufacturing,” Pattern Recognition Letters, vol. 136, 2020, doi: 10.1016/j.patrec.2020.06.008.

[17] W. Zaman, M. F. Siddique, S. Ullah, F. Saleem, and J.-M. Kim, “Hybrid Deep Learning Model for Fault Diagnosis in Centrifugal Pumps,” Machines, vol. 12, no. 12, 2024, doi: 10.3390/machines12120905.

[18] C. E. Sunal, V. Velisavljevic, V. Dyo, B. Newton, and J. Newton, "Centrifugal Pump Fault Detection with Convolutional Neural Network Transfer Learning," Sensors, vol. 24, no. 8, 2024, doi: 10.3390/s24082442.

[19] Z. Zhao et al., "Deep Learning Algorithms for Rotating Machinery Intelligent Diagnosis: An Open Source Benchmark Study," ISA Transactions, vol. 107, pp. 224-255., 2020, doi: 10.1016/j.isatra.2020.08.010.

[20] Z. Xu et al., "A Digital Twin System for Centrifugal Pump Fault Diagnosis Driven by Transfer Learning Based on Graph Convolutional Neural Networks," Computers in Industry, vol. 163, 2024, doi: 10.1016/j.compind.2024.104155.

[21] A. Kumar, R. Kumar, J. Xiang, Z. Qiao, Y. Zhou, and H. Shao, "Digital Twin-Assisted AI Framework for Bearing Defect Diagnosis in Centrifugal Pump," Measurement, vol. 235, 2024, doi: 10.1016/j.measurement.2024.115013.

[22] C. V. Prasshanth, S. N. Venkatesh, T. K. Mahanta, N. R. Sakthivel, and V. Sugumaran, "Fault Diagnosis of Monoblock Centrifugal Pumps Using Pre-Trained Deep Learning Models," Engineering Applications of Artificial Intelligence, vol. 136, 2024, doi: 10.1016/j.engappai.2024.109022.

[23] M. A. B. Syed, M. R. Hasan, N. I. Chowdhury, M. H. Rahman, and I. Ahmed, "A Systematic Review of Time Series Algorithms and Analytics in Predictive Maintenance," Decision Analytics Journal, vol. 15, 2025, doi: 10.1016/j.dajour.2025.100573.

[24] H. Chen, J. Li, X.-B. Wang, L.-Q. Yu, and Z.-X. Yang, "Review of Intelligent Fault Diagnosis for Rotating Machinery under Imperfect Data Conditions," Expert Systems with Applications, vol. 285, 2025, doi: 10.1016/j.eswa.2025.127726.

[25] Z. Li, H. Jiang, and Y. Dong, "A Convolutional-Transformer Reinforcement Learning Agent for Rotating Machinery Fault Diagnosis," Expert Systems with Applications, vol. 198, 2025, doi: 10.1016/j.eswa.2025.126669.

[26] O. Rashad, O. Attallah, and I. Morsi, "A smart PLC-SCADA framework for monitoring petroleum products terminals in industry 4.0 via machine learning," Measurement and Control, vol. 55, no. 5–6, pp. 1–14, 2022, doi: 10.1177/00202940221103305.

[27] J. Zheng, C. Wang, Y. Liang, Q. Liao, Z. Li, and B. Wang, "Deeppipe: A deep-learning method for anomaly detection of multi-product pipelines," Energy, vol. 252, p. 125025, 2022, doi: 10.1016/j.energy.2022.125025.

[28] A. Melo, M. Melo, and J. Pinto, "Data-Driven Process Monitoring and Fault Diagnosis: A Comprehensive Survey," Processes, vol. 12, no. 2, p. 251, 2024, doi: 10.3390/pr12020251.

[29] Y. Cui, W. Fan, and Y. Zhou, "Dimensionality reducing Gaussian mixture-based reconstruction for fault detection in multimode processes," The Canadian Journal of Chemical Engineering, vol. 102, no. 12, pp. 4267–4280, 2024, doi: 10.1002/cjce.25308.

[30] J. K. Seo, J. Lee, B. Kim, W. Shim, and J. T. Seo, "AI-Based Anomaly Detection in Industrial Control and Cyber–Physical Systems," Electronics, vol. 15, no. 1, 2026, doi: 10.3390/electronics15010020.