ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

CIVIL-426

MACHINE LEARNING FOR PREDICTIVE MAINTENANCE APPLICATIONS

Prof. Olga FINK

AXPO CHALLENGE

Shreyas Nara (415268)
Matas Jones (313222)
Benjamin Bahurel (326888)
Georg Schwabedal (328434)

18th December 2025

1 Introduction

In the context of the CIVIL-426 Machine Learning for Predictive Maintenance course, Axpo, the largest producer of renewable energy in Switzerland, challenged us to develop a method for detecting anomalies in the operation of their ball valves. This study focuses on the Kraftwerke Sarganserland AG (KSL) hydroelectric scheme located in the canton of St. Gallen. As illustrated in Figure 1, the scheme comprises a main reservoir, a balancing reservoir, and two power stations: Mapragg and Sarelli. Both facilities rely heavily on spherical valves, which are particularly well-suited to withstand the high pressures present in the penstocks.

These valves play a critical role in ensuring turbine inlet safety, as they must be capable of rapidly shutting off the water flow during emergency situations or maintenance operations. However, phenomena such as seal degradation, debris accumulation, or material ageing can cause the valves to leak thereby altering their normal operating dynamics and potentially compromising system safety.

Axpo tasked us with designing an unsupervised anomaly detection model that:

- Learns the normal valve opening and closing behaviour,
- Assigns an anomaly score to each opening and closing event based on deviations from this behaviour,
- Detects target deviations with minimal delay
- Generalizes well to unseen operating conditions.

In this report, we describe how these objectives were achieved. We begin with an exploratory data analysis, followed by the presentation of a model designed to detect anomalies in valve opening and closing durations. Finally, we introduce a second model capable of identifying outliers in the transient dynamics of ball valve closing events.

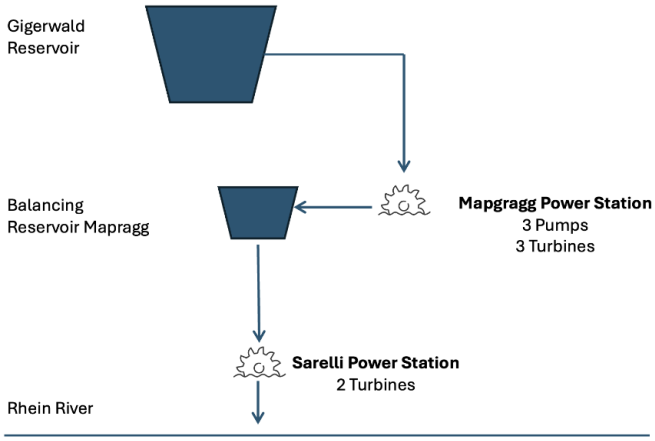


FIGURE 1
Mapgragg and Sarelli power stations schematic [source: EPFL CIVIL-426]

2 Exploratory Analysis

There are multiple metrics which are measured at the Mapragg and Sarelli stations:

- Active power [MW]
- Ball valve open status [-]
- Ball valve closed status [-]
- Guide vane position [%]
- Water pressure downstream [bar]
- Water pressure upstream [bar]

The active power represents the instantaneous power output of a turbine, measured in MW. The Mapragg station can operate in two modes:

- Turbine mode, in which water drives the turbines to generate electricity.
- Pump mode, in which water is pumped back up to the Gigerwald reservoir for later use.

In contrast, the Sarelli station can operate only in turbine mode. These operating modes can be identified by the sign of the active power: a positive value indicates power production, while a negative value indicates power consumption.

The ball valve open/closed status indicates whether the spherical ball valve is completely open or completely closed.

The guide vane position, expressed as a percentage, indicates how far the guide vane is opened: 0% corresponds to fully closed, while 100% corresponds to fully open. The guide vane regulates the flow of water entering the turbine for power production or during pumping.

The downstream water pressure refers to the pressure in the system after the ball valve, whereas the upstream pressure refers to the pressure before the ball valve.

The first step taken for the data exploration was to plot the raw data for each dataset. The figure below shows the data collected from the different sensors of the Mapragg MG1 training dataset but the analysis conclusions are valid for all the datasets, Mapragg and Sarelli, testing and training:

As one can observe from the figure 2, there are large gaps in the data where interruptions in the data sampling has occurred, likely due to the system being offline for maintenance or other related reasons. To solve this issue, our team decided to segment the data into smaller data frames in order to remove said large data gaps. Figure 3 displays the different signals acquired before the first large data gap (0s - 200'000s). As one can see, the data sampling is not the same for every signal. Indeed, the data sampling for the ball valve status is quite sparse whereas the upstream water pressure sampling rate is much greater. The table below summarizes the sampling rates of the different signals:

TABLE 1
Average sampling rates for Mapragg and Sarelli datasets

Signal	Mapragg [Hz]	Sarelli [Hz]
Ball valve closed	0.0001	>0.0000
Ball valve open	0.0001	>0.0000
Active power	0.0450	0.0067
Guide vane position	0.0479	0.0708
Water pressure downstream	0.0329	0.0199
Water pressure upstream	0.0815	0.0306

Please note that the table above only gives the average sampling rate for the different signals. In reality, the sampling rates vary with time but are close to the aforementioned average. As one can see, the sampling rate is vastly different between signals. Furthermore, the sampling rate between the Mapragg and Sarelli signals also differs.

In order to analyse the data further, data resampling and interpolation is required. To accomplish this, our team resampled each segment of the Mapragg and Sarelli signal segments individually with the fastest sampling rate of each segment. After this, we applied a forward interpolation. This method was preferred because it bases each new value only on past observations, which reduces the risk of introducing information from the future into the signal. Finally, EMA (exponential moving average) was applied to the resampled signals in order to smooth the signals.

There are three main states the Mapragg turbine can be in: turbine mode, pump mode and transient states. One can differentiate the modes from the transient states by observing a stable active power, positive or negative for the turbine and pump modes respectively. The transient state occurs during the closing or opening of the valves which may lead to hard-to-predict signal measurements. Figure **??** displays the different signals measured at the Mapragg power stations during a system start up and a system shutdown and during a steady-state power generation and pumping given sequence. The following statements can be observed: first during a state transient (start up or shutdown), the active power, guide vane position and downstream pressure all vary in the same direction whereas the upstream pressure varies in the opposite direction and is prone to oscillations in the signal, most likely due to hydraulic transient pressure

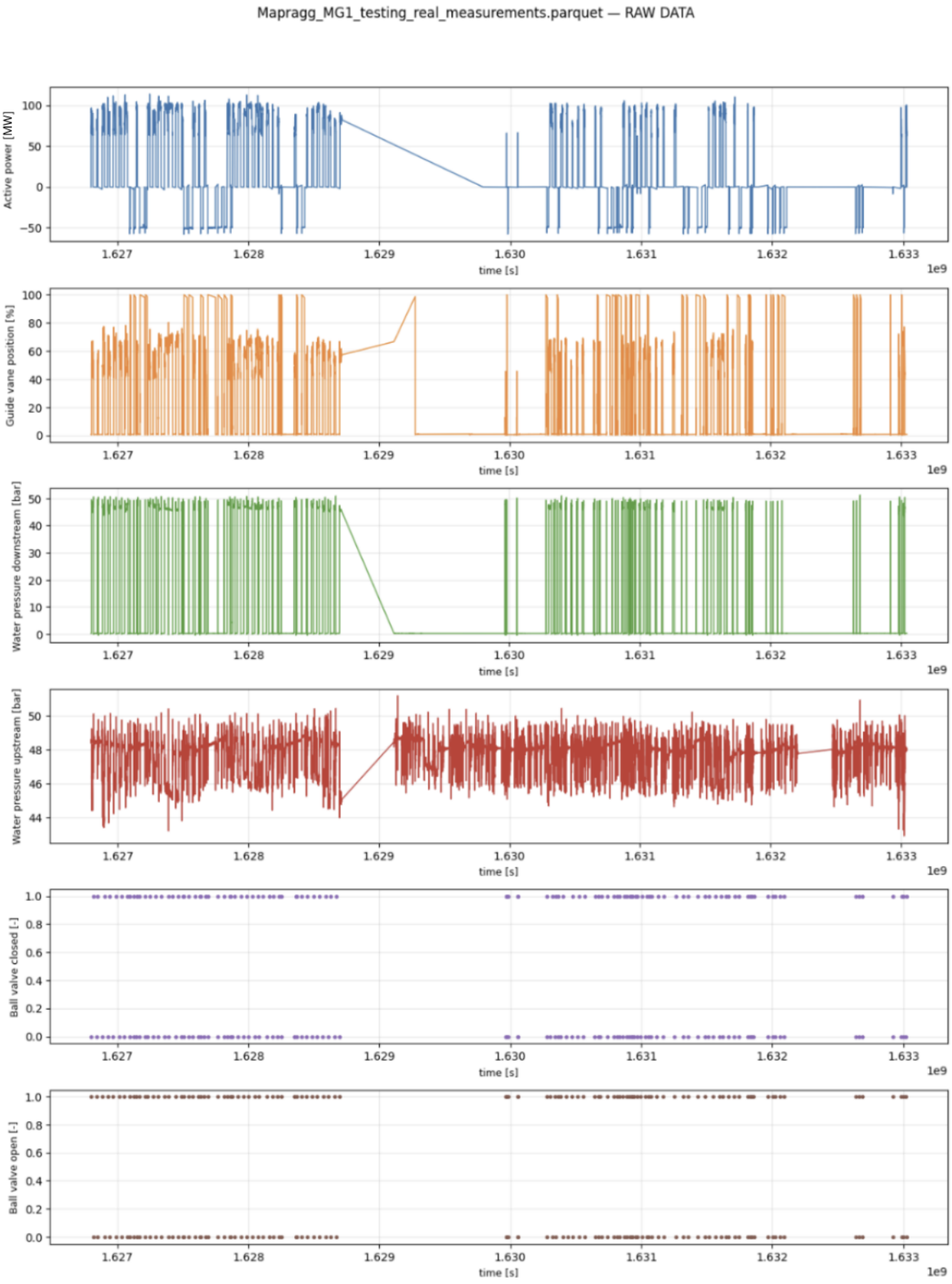


FIGURE 2
Different raw signals of the Mapragg MG1 training dataset

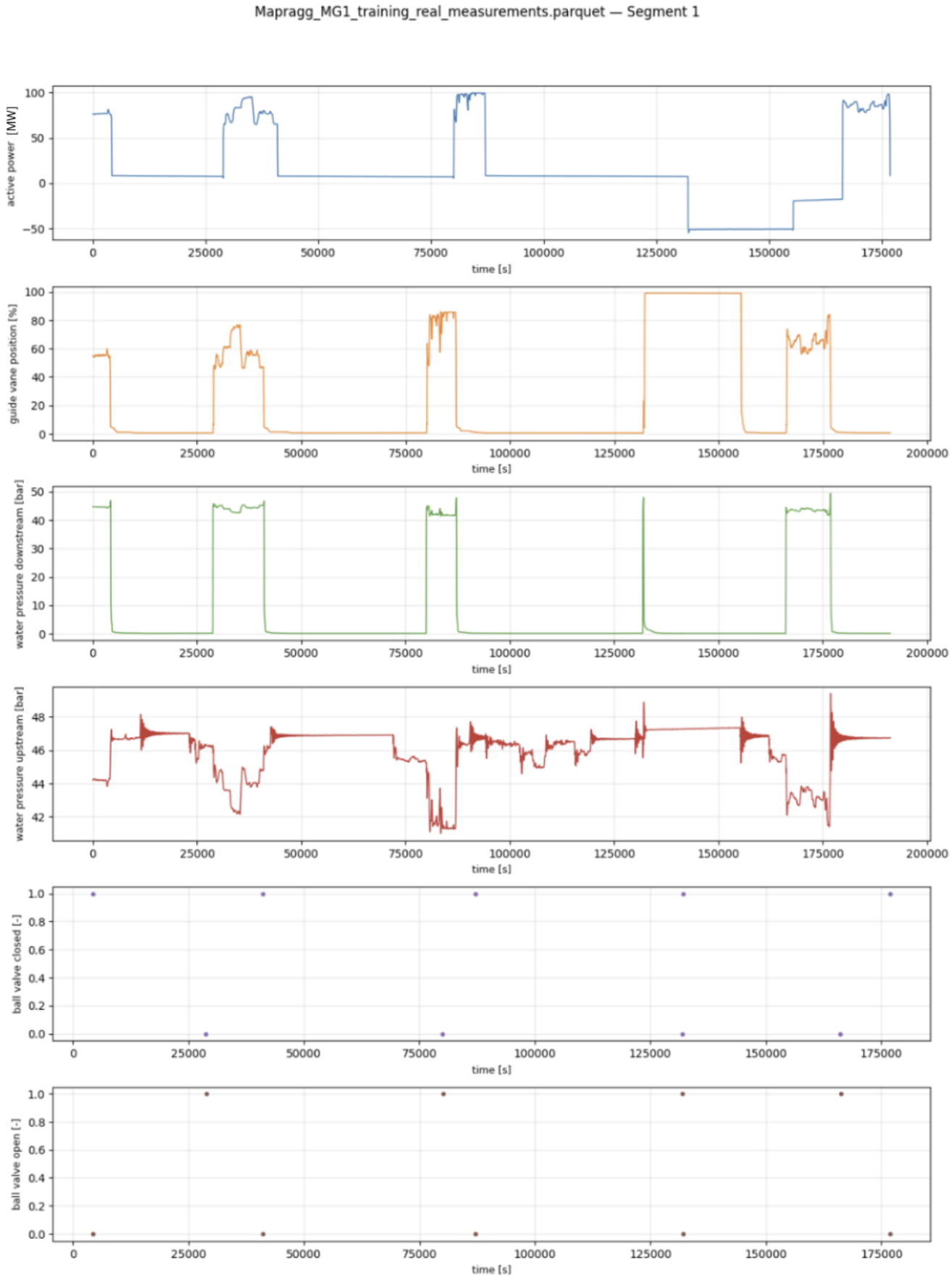


FIGURE 3
Segment 1 of the different signals of the Mapragg MG1 training dataset

waves caused by the fast changes in water flow when the unit switches between operating modes. In contrast, during the power generation and pumping steady states, the active power and guide vane position move in the same direction whereas the upstream and downstream pressure move together in the opposite direction. When it comes to the ball valve statuses, the ball valve is open during power generation and during start-up and closes during system shutdown. Unexpectedly, the ball valve appears to be mostly closed during pumping, on briefly opening before closing again. This was contrary to our initial belief that the ball valve needed to remain open for water to be able to flow back up into the reservoir for pumping. As of this moment, we were unable to find a reasoning to this phenomena.

2.1 Root Cause Analysis and Interpretation

To test the correlations identified in Figure **??**, our team separated the data into three operating categories: power generation, pumping, and state transitions. We then generated a correlation map for each category. Figure 4 presents the three corresponding correlation heat maps.

In the power-generation state (leftmost map), the active power and guide-vane position are strongly correlated, and both are inversely correlated with the upstream and downstream water pressures in the penstock, which themselves are strongly mutually correlated. This matches the expected physical behaviour: higher guide-vane opening increases flow and power while reducing penstock pressure due to increased water velocity.

In pumping mode, the patterns are more nuanced. There is an inverse correlation between active power and guide-vane position which is consistent with the guide vane remaining fully open while the turbine is pumping. A positive correlation appears between active power and the upstream water pressure which is also expected: when the unit pumps water upward, the upstream pressure increases. Other correlations are relatively small, likely because the pumping cycles are short, with rapid activation and deactivation, providing little stable data for stronger statistical relationships.

Finally, the transient-state correlation matrix shows that active power, guide-vane position, and downstream pressure tend to vary together while all three are inversely correlated with the upstream pressure. This reflects the hydraulic oscillations and control adjustments occurring during state changes.

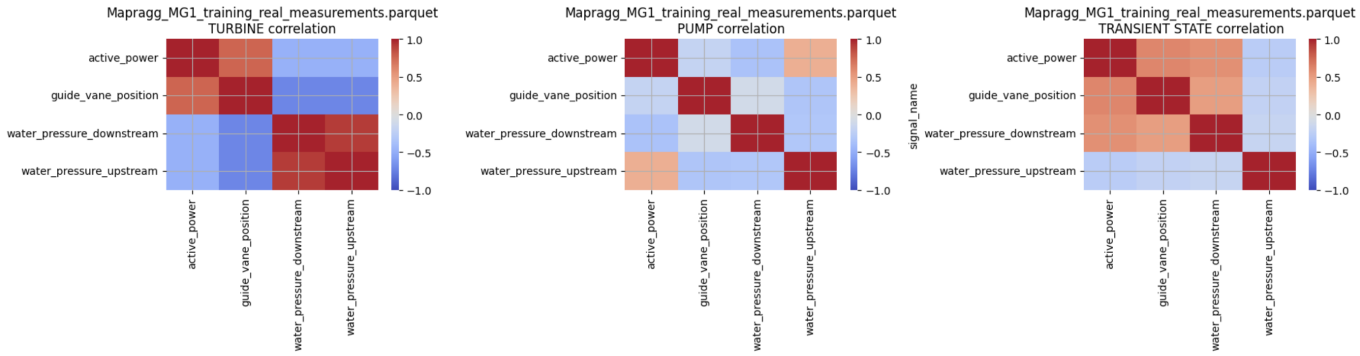
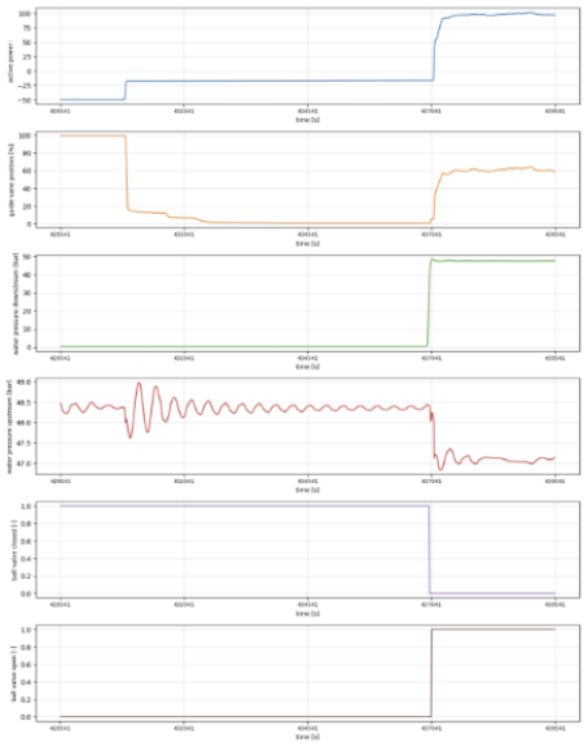
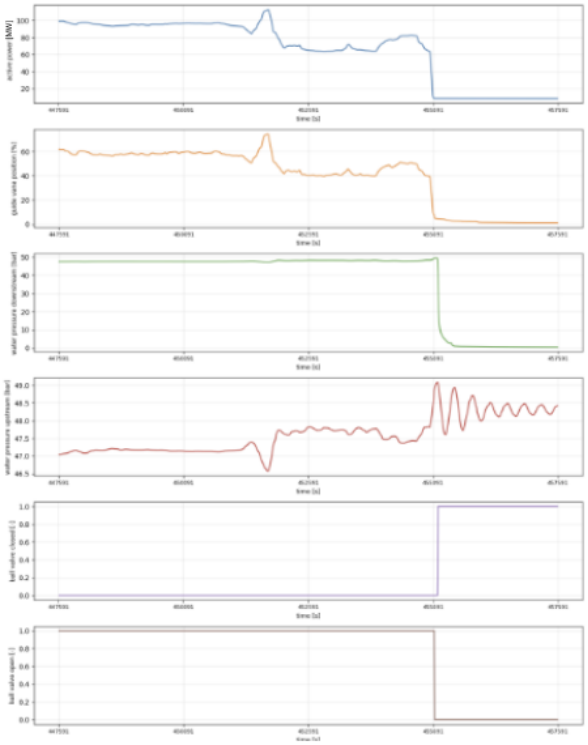


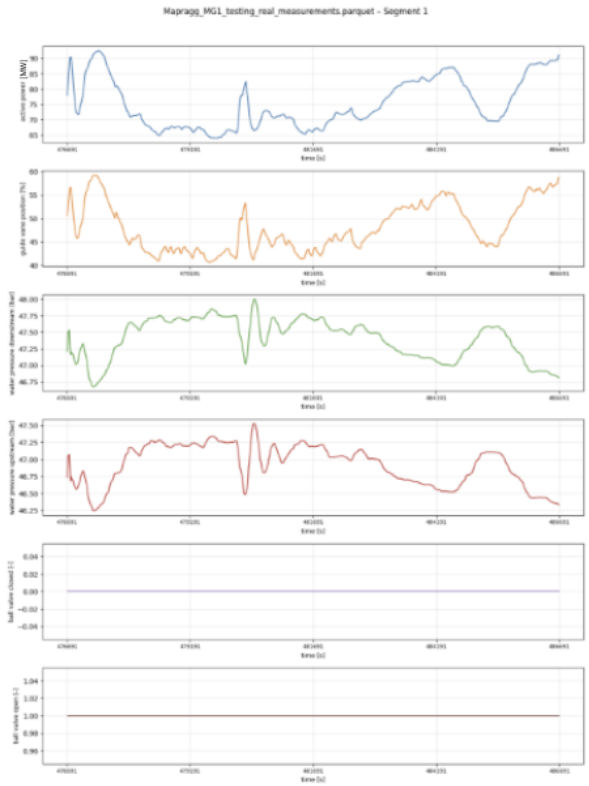
FIGURE 4
Correlation heat maps for the Mapgragg MG1 training set on power generation, pumping and transient states



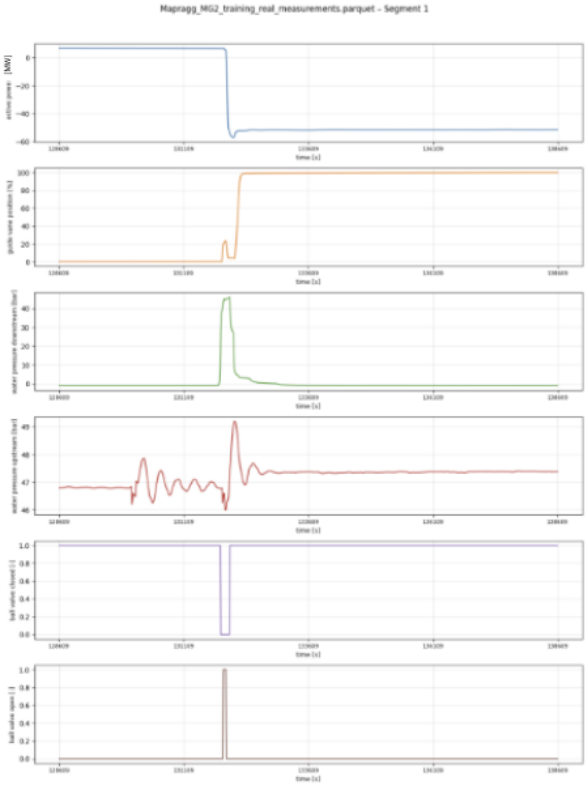
((A)) Mapragg start up sequence signals



((B)) Mapragg shutdown sequence signals



((C)) Mapragg power generation signals



((D)) Mapragg pumping signals

FIGURE 5
Different signals of the Mapgragg MG1 operating in turbine, pumping and transient modes

3 Task 1

In task 1, we are challenged with determining the transient opening and closing times for the different ball valves and then performing a statistical method, such as the Z-score, to detect any anomalies. It is important to note that we are not allowed to use the Boolean ball valve status signals.

Our team achieved this goal in the following steps:

1. Opening/Closing event detection and segregation
    - Event flagging based on rolling average of 4 allowed signals
    - Event verification using wider rolling average on downstream pressure

    $\rightarrow$ If both flag an event, there is an event
2. Isolate and extract neighbourhood around the event region
3. Train TCN to generate opening/closing event probabilities on said event regions
4. Perform linear regression to determine the event time based on the opening/closing probabilities
5. Extract transient time and perform statistical analysis.

3.1 Event detection

In the data analysis section, we observed that during a state transient, during an opening or a closing, the four allowed signals (active power, guide vane position, downstream pressure and upstream pressure) all fluctuated greatly. Thus, we applied a rolling average which flagged a potential event if said average fluctuated above a certain threshold.

This method worked but it lead to many false positives. To improve the situation, a second filter was introduced. It consisted on a second rolling average solely on the downstream pressure which we found was a more robust and consistent metric to determine the eventuality of an event. If both averages signalled the presence of an event, the region around said event was isolated from the overall data.

Figure 6 shows an overview of the detected opening and closing events from the Mapragg MG1 testing set. The red lines spanning the plots indicates the presence of an event. It isn't too difficult to determine the presence of an event. However, it is more difficult to determine the the start and the end of an event. If one focused on a singular event such as in the case of figure 7, one would see that there aren't many indications among the four analysed signals showing a current opening or closing. In any case, we extracted the neighbourhood and moved on to the next step.

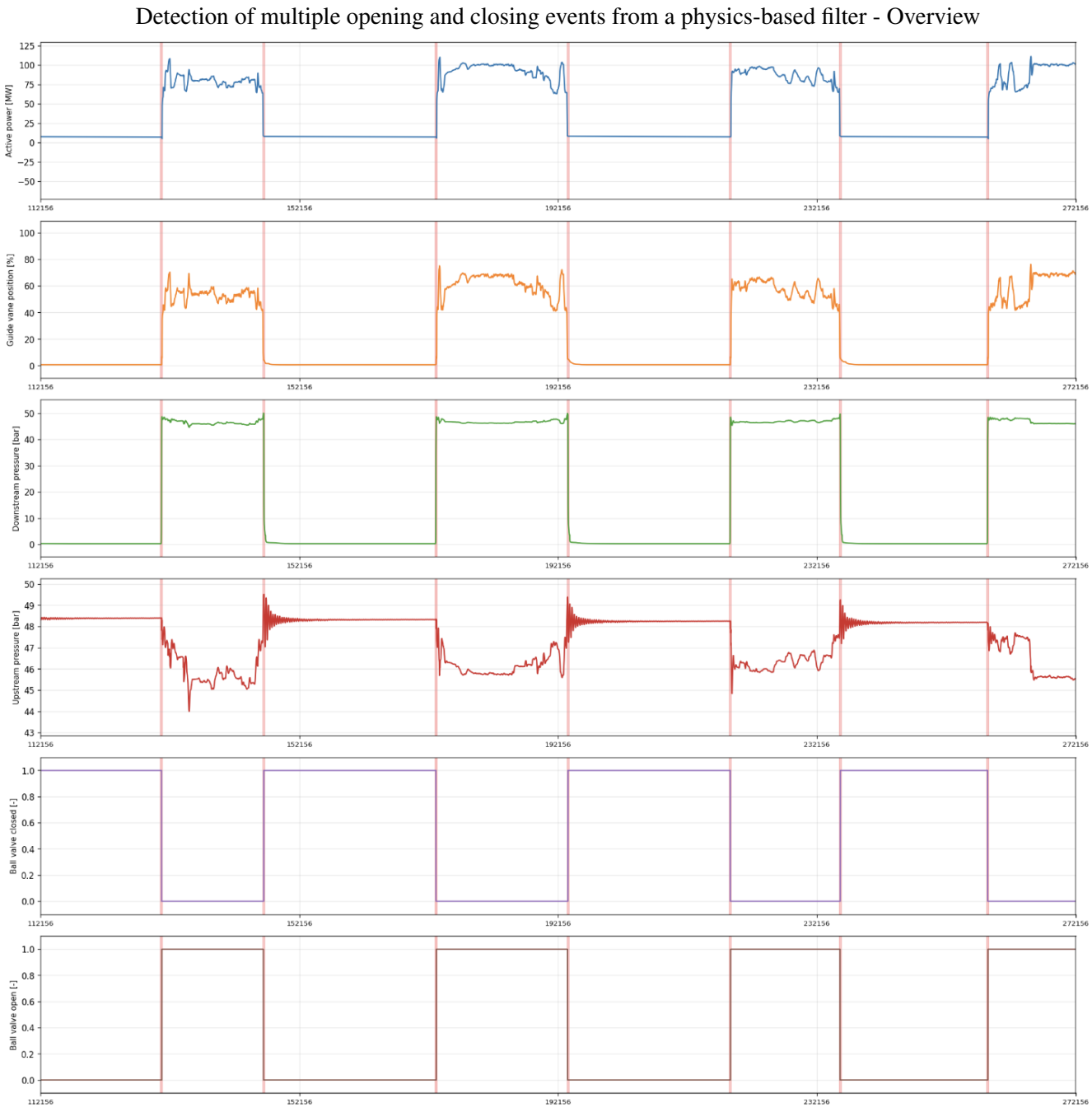


FIGURE 6
Event detection using rolling average on 4 signals

Detection of multiple opening and closing event from a physics-based filter - Close up

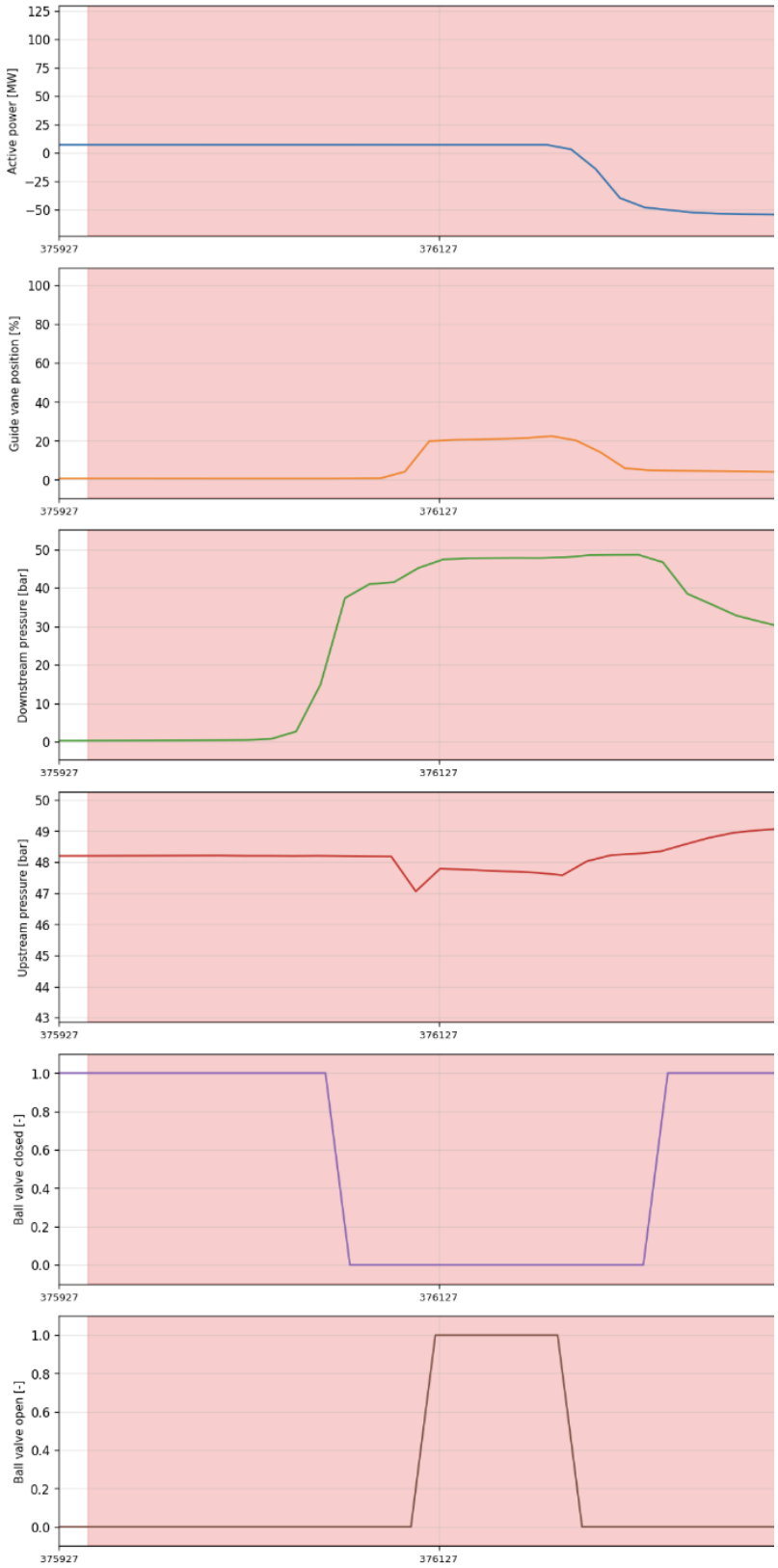


FIGURE 7
Event detection using rolling average on 4 signals, close up

3.2 Event timestamp determination

After extracting the event neighbourhoods, the next step was to determine the precise start and end of each valve event. Our initial approach relied on applying physically motivated filters based on signal derivatives, but this produced limited success. To address this, we deployed a Temporal Convolutional Network (TCN) that takes four input channels and outputs two probability channels that represent the likelihood of the ball valve being open or closed. A TCN was selected because it captures temporal dependencies in sequential data, which fits the structure of our signals.

The implementation uses only the measured ball valve statuses as labels. The dataset was split into 80 percent for training and 20 percent for evaluation. After training, the model was applied to the full dataset to produce estimated probability traces for the valve-open and valve-closed states. Figure 8 shows an example of these predicted probabilities. On their own, the probability curves are not very informative. However, visual inspection shows that they match the measured valve statuses reasonably well, which indicates that they contain the transitions we want to detect.

Our goal was to identify when the probabilities entered a transient phase, meaning when they stopped being constant and began decreasing toward 0 or increasing toward 1. To do this, we applied a slope-based detection method to locate these transients and selected a representative point near the midpoint of each transition, approximately where the probability crosses 0.5. A linear regression was then fitted along each transient to estimate the actual beginning and end of each opening or closing event. These intervals are displayed as coloured segments in the figure. Once these boundaries are known, the total opening or closing time can be computed as the time difference between the inferred open and closed status updates.

This method performs well in practice and usually agrees with the measured valve statuses, although not perfectly. It is not clear whether the discrepancies are caused by limitations in the probability estimates, inaccuracies in the measurements, or effects of resampling combined with the relatively low sampling rate of the original valve status signal. Despite this uncertainty, the TCN-based method provides a reliable and interpretable way to extract event boundaries from multichannel sequential data.

TCN valve opening and closing probability plot

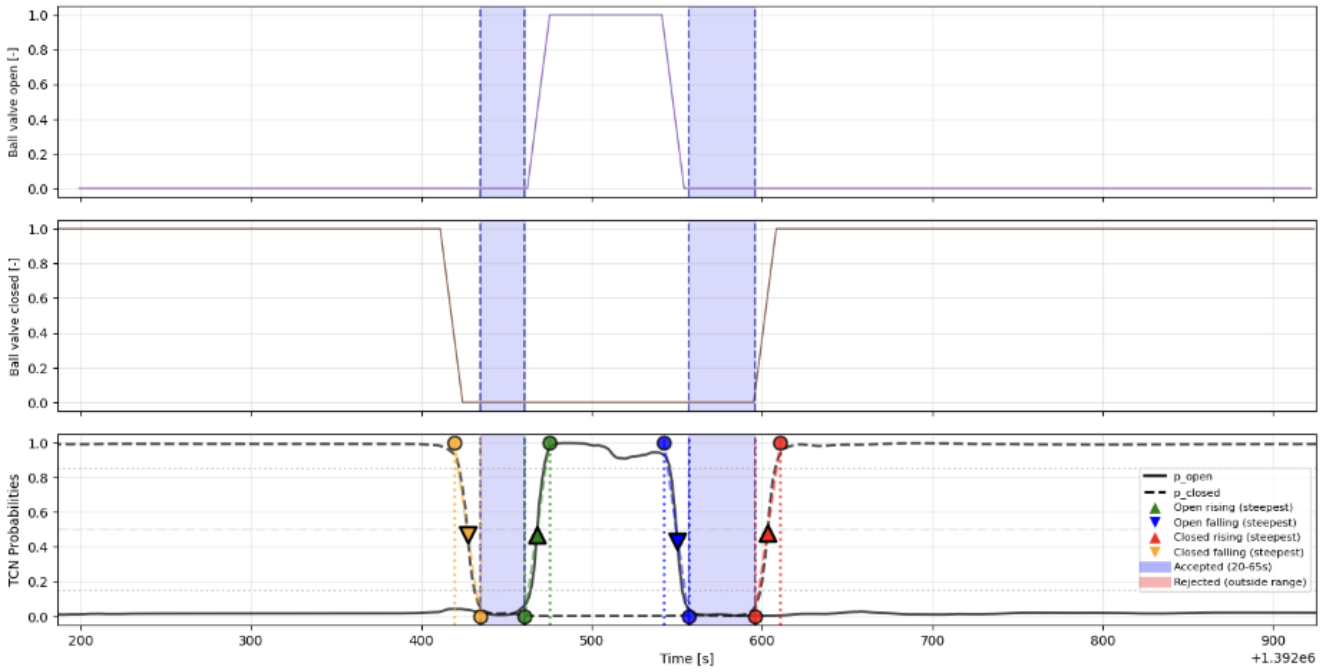


FIGURE 8
Valve status probability generated by TCN

We needed to see how accurate the developed model was. To this end, we used the measured valve open/close statuses to determine the duration of a ball valve event. The same logic was used as for the probable opening/closing duration estimations but this time using the measured data. An example of the results is given in Figure 9.

Table 2 summarizes the duration estimation accuracy of the proposed method. The table reports the error between event

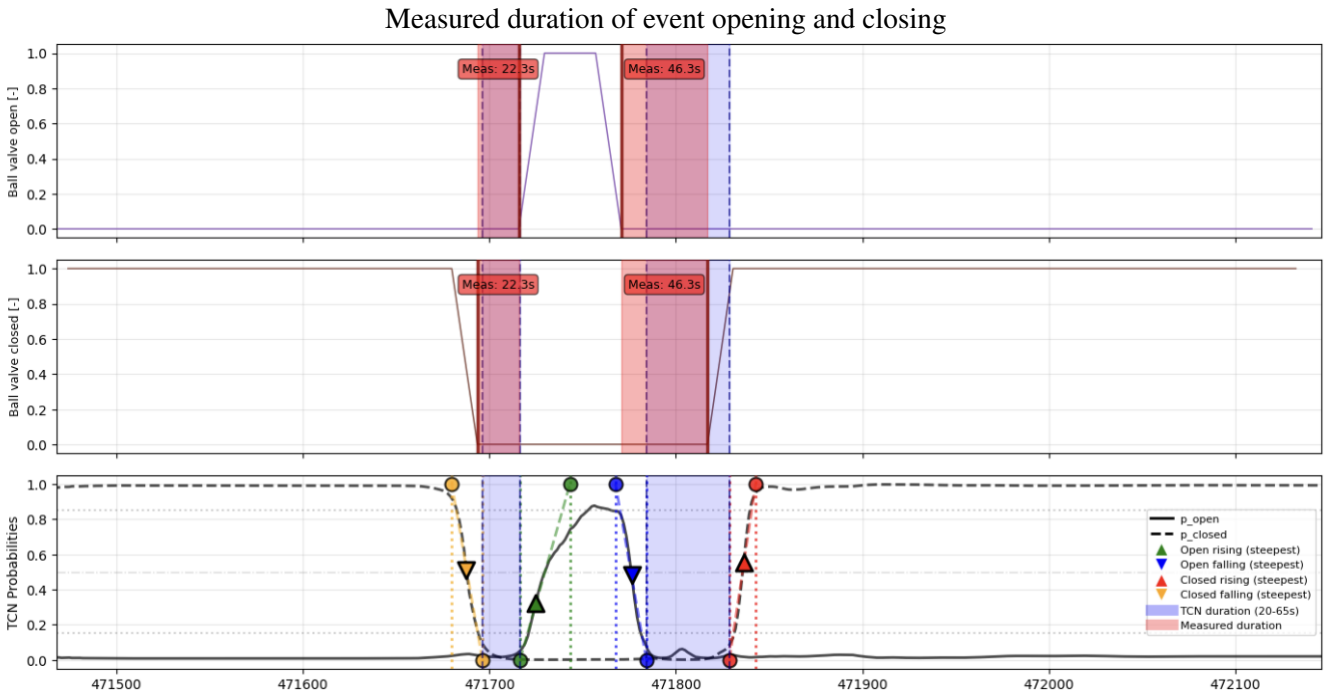


FIGURE 9
Duration accuracy determination

durations predicted by the TCN combined with a linear regressor and the corresponding measured durations. Overall, the results show a strong agreement between predicted and measured values. The error distribution is centred close to zero, with a mean error of -0.35 seconds and a median error of -0.01 seconds, indicating only a slight systematic underestimation by the model. The mean absolute error remains low at 1.58 seconds, while the standard deviation of the error is 2.84 seconds, confirming limited dispersion around the true durations.

Despite this overall good performance, the maximum observed error reaches 10.00 seconds. Such large deviations, although infrequent, may adversely affect the reliability of the anomaly detection, as discussed in the following section.

TABLE 2
Duration estimation accuracy and statistics

TCN Duration Statistics	
Count	5433
Mean duration [s]	40.96
Median duration [s]	38.71
Standard deviation [s]	9.70
Range [s]	$[20.22,\ 65.00]$
Measured Duration Statistics	
Count	5731
Mean duration [s]	42.25
Median duration [s]	40.91
Standard deviation [s]	10.16
Range [s]	$[21.28,\ 64.25]$
Accuracy Metrics	
Mean error (TCN – Measured) [s]	-0.35
Median error [s]	-0.01
Mean absolute error (MAE) [s]	1.58
Standard deviation [s]	2.84
Maximum error [s]	10.00

3.3 Statistical measurements

Finally, we address the task of detecting anomalies using statistical methods. To this end, we first analysed the distributions of opening and closing event durations for both the training and testing data sets. These distributions are shown as histograms in Figures 10 and 11. Two distinct behavioural patterns can be observed: mono-modal and bimodal distributions, both of which exhibit outliers around their main modes. Due to the presence of multimodality, conventional outlier detection techniques such as z-score thresholding or IQR methods are ill-suited for this problem, as they rely on unimodal or approximately Gaussian assumptions. Instead, we adopted a density-based statistical approach, in which data points lying outside the high-density regions (highlighted in green) are classified as anomalous.

The results of this anomaly detection process are summarized in Table 3, which reports the number and proportion of detected outliers across the different operating conditions.

TABLE 3
Cluster-Based Anomaly Detection Results

	Opening Phase						Closing Phase					
	Training		Testing		Total		Training		Testing		Total	
	n	Out.	n	Out.	n	Rate (%)	n	Out.	n	Out.	n	Rate (%)
Mapragg MG1	600	11	99	3	699	2.0	587	15	94	3	681	2.6
Mapragg MG2	694	11	85	3	779	1.8	640	7	81	7	721	1.9
Mapragg MG3	473	7	137	1	610	1.3	460	8	144	4	604	2.0
Sarelli MG1	283	7	66	3	349	2.9	254	8	60	2	314	3.2
Sarelli MG2	288	16	74	8	362	6.6	260	2	62	5	322	2.2

As can be observed, the detected outliers comprise a mixture of both training and testing samples. As discussed in the previous subsection, this behavior is largely attributable to imperfections in the event duration estimation method, which is likely responsible for a portion of the false positives. This is particularly evident in the training data set, where no anomalies are expected by design.

Although the absolute number of outliers is higher in the training set than in the testing set, this discrepancy is primarily due to the larger size of the training data set. When the number of outliers is normalized by the respective data set sizes, the proportion of testing outliers exceeds that of the training outliers. This observation is encouraging, as it suggests that the detection method is able to identify deviations more frequently in unseen data.

Overall, these results indicate that the proposed approach has strong potential for detecting anomalies in spherical valve opening and closing operations. With targeted refinements, notably improvements to the event duration estimation algorithm, the method could become a reliable tool for opening and closing anomaly detection.

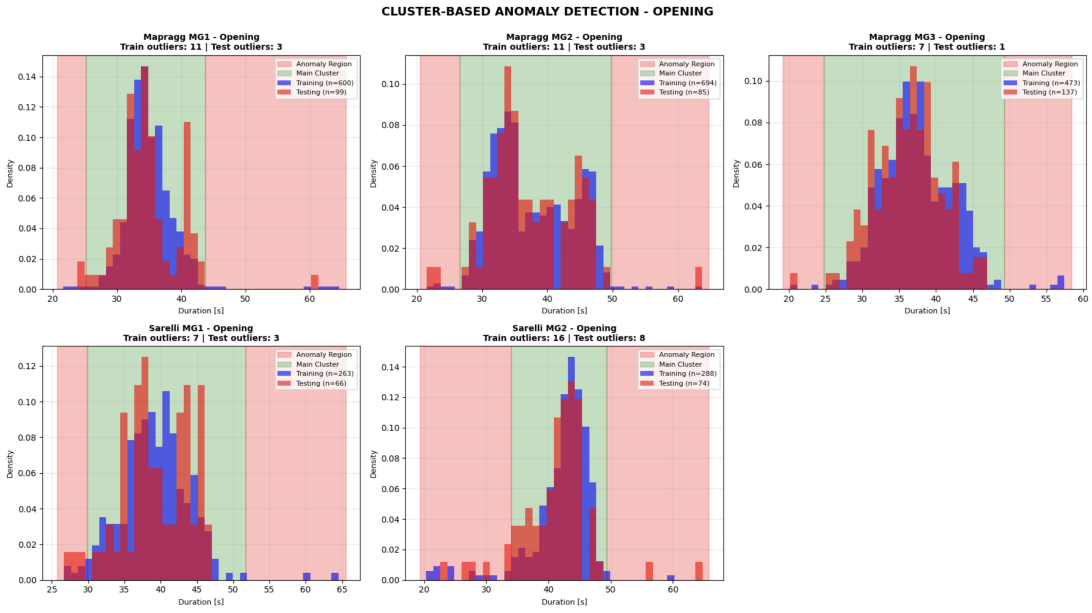


FIGURE 10
Opening duration anomaly detection based on cluster density analysis

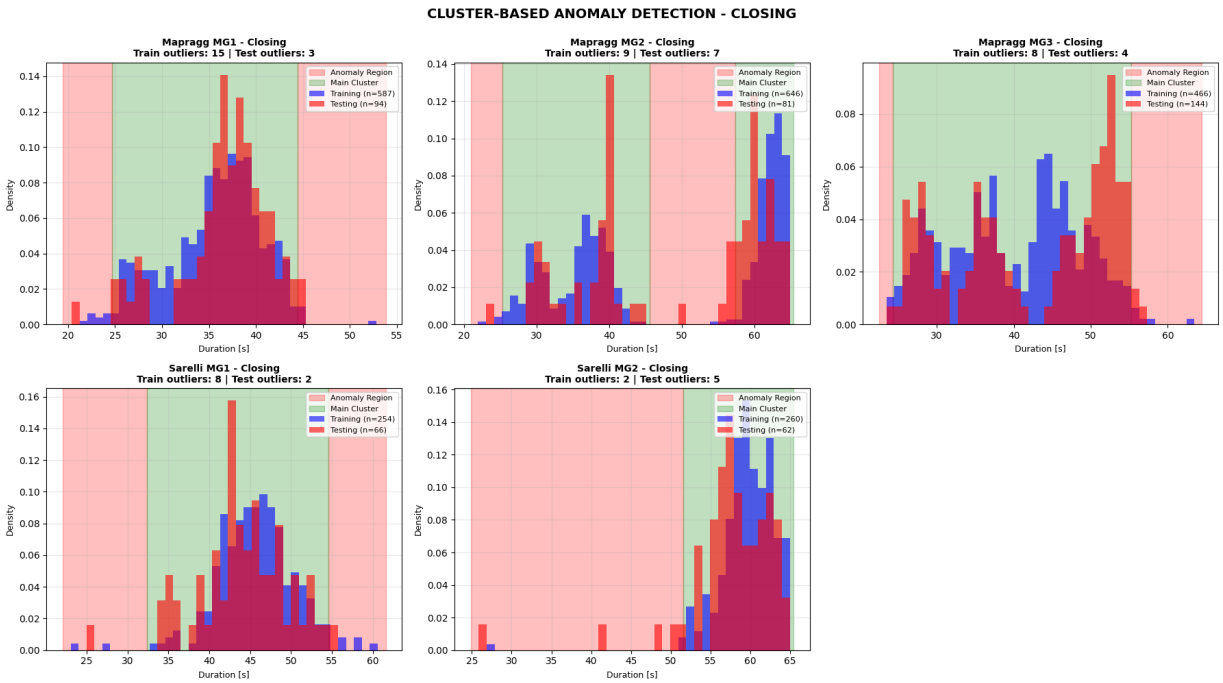


FIGURE 11
Closing duration anomaly detection based on cluster density analysis

4 Task 2

4.1 Data Preprocessing

The time-series data were first temporally standardized to ensure consistent alignment across all signals. All timestamps were converted to a unified datetime format, made timezone-naive, and rounded to the nearest second. The data were then resampled onto a fixed 1 Hz time grid, creating a uniform, continuous temporal axis from the beginning to the end of each recording. When multiple measurements occur within the same second, their values were aggregated using the mean.

Resampling introduces missing values in periods where no data was originally recorded. These missing segments were handled using a gap-limited forward-fill strategy, where short gaps of up to 300 seconds are filled forward, while longer gaps remain missing. This approach preserves short-term continuity in the presence of brief signal interruptions but avoids introducing artificial behavior across extended periods of inactivity, such as multi-hour gaps.

To reduce high-frequency noise while preserving underlying dynamics, exponential moving average smoothing was applied to continuous-valued signals using a smoothing factor of 0.5. Discrete signals were excluded from this operation to maintain their binary nature. After temporal alignment, missing-data handling, and smoothing, the processed time series is suitable for event detection and window-based sample construction.

In the next step, the continuous time series was segmented into fixed-length temporal windows of 360 seconds. The *ball_valve_closed* signal was used as a ground-truth indicator to locate valve-closing events, and each window was centered on the corresponding transition, spanning 180 seconds before and 180 seconds after closure. This duration was selected based on domain expert input (Romande Energie - hydraulic maintenance division lead: Rodolphe Richard), who confirmed that a 360-second window is both necessary and sufficient to capture the full mechanical, hydraulic, and electrical response associated with a closing sequence.

4.2 Modelling Methodology

To model the complex system behavior during valve transitions, we adopted a joint distribution strategy that captures the physical coupling between guide vane position, pressure variations, and active power. Instead of treating signals in isolation, this approach stacks and flattens six synchronized sensor streams into a 2160-dimensional input vector (6 signals $\times$ 360 time steps), allowing a single model to learn the holistic dynamics of a normal closing event.

The core of our pipeline is a deep fully connected autoencoder designed to compress these high-dimensional windows into a 128-unit latent bottleneck. The encoder utilizes successive nonlinear transformations through layers of 512 and 256 neurons, while the decoder mirrors this structure to reconstruct the original signals. To ensure the model accurately captures the physics of the valve movement, we utilized a hybrid training objective: a standard Mean Squared Error (MSE) complemented by a Gaussian-weighted MSE centered on the transition point.

For unsupervised anomaly characterization, we apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) directly in the raw 128-dimensional latent space. Retaining the full latent representation preserves cross-signal correlations learned by the encoder that would be distorted by further dimensionality reduction. After anomalies are flagged using the OR-based Moving Window MSE threshold (97.5th percentile), their latent positions are compared to synthetic anomaly seeds using a k-Nearest Neighbors (k=5) inference scheme. This procedure enables attribution of anomalies to known failure archetypes, while remaining fully unsupervised and independent of real fault labels. Figure 12 illustrates the end-to-end architecture from joint signal input to latent clustering.

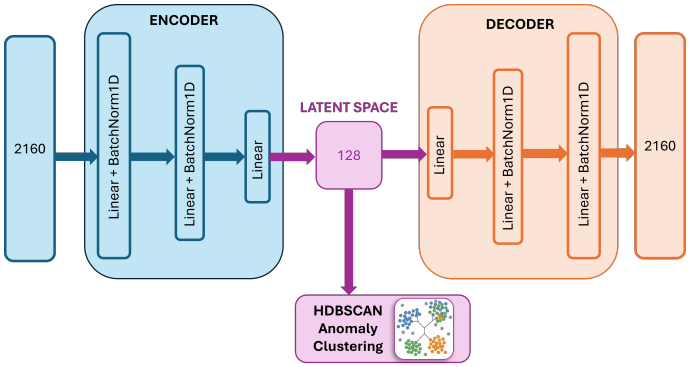


FIGURE 12
Model Architecture

Training is performed using a hybrid reconstruction-based objective, which combines a uniform mean squared error (MSE) with a Gaussian-weighted MSE (amplitude = 4.0) centered on the valve transition point to prioritize reconstruction accuracy during critical closing phases. Prior to feature extraction, all analog signals undergo Exponential Moving Average (EMA) smoothing ($\alpha = 0.5$) to reduce noise. Features are then normalized using a MinMaxScaler. The model architecture consists of a deep autoencoder with a 2160-dimensional input (6 signals over a 360s window) compressed into a 128-unit latent bottleneck. Data is split into training and validation subsets (80/20 ratio), with the turbine and pump regimes differentiated by calculating the mean active_power during the first 180 seconds of each window. Optimization is carried out using the Adam optimizer (learning rate = 0.001) paired with a ReduceLROnPlateau scheduler (Why the Optuna value wasn't used directly). A batch size of 32 was used over 100 epochs, utilizing early stopping based on validation loss; these hyperparameters were optimized via a grid search with Optuna. Table 4 summarizes the experimental setup.

TABLE 4
Experimental Setup for Autoencoder Training

Parameter	Value
Loss Function	Hybrid (Uniform MSE + Transition-Weighted MSE)
Smoothing	EMA ($\alpha = 0.5$)
Architecture	$2160 \rightarrow 512 \rightarrow 256 \rightarrow 128$
Optimizer	Adam (LR = 0.001, with scheduler)
Data Split	80% training / 20% validation
Regime Logic	Mean P_{active} of first 180s
Data Normalization	MinMaxScaler
Hyperparameter Search	Optuna

The optimization process was conducted independently for each operational regime, utilizing a total of 1,146 turbine samples and 776 pump samples. The Turbine model achieved a best validation loss of 0.002578, reaching near to end

stopping at epoch 99, while the Pump model converged more rapidly, triggering early stopping at epoch 78 with a best validation loss of 0.003029. This discrepancy in convergence rates reflects the distinct physical dynamics and signal variance present in the two regimes, pumping working against gravity and therefore presenting signals harder to represent paired with less occurrences. As shown in Figure 13, the training and validation loss curves exhibit a consistent downward trend without significant divergence, indicating that the models successfully generalized the normal valve closing behavior without overfitting to the specific training windows.

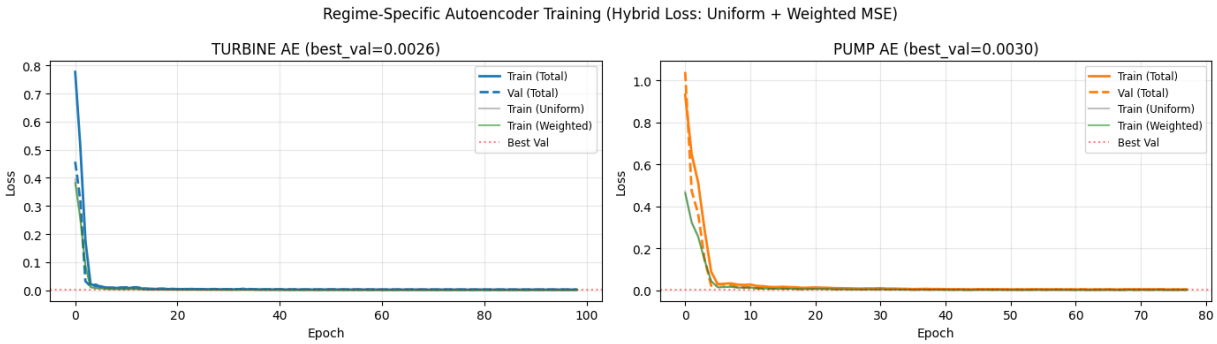


FIGURE 13
Training curves

4.3 Synthetic anomaly generation

Since labeled fault data for hydropower valve operations is not available, anomaly types had to be defined synthetically in order to evaluate the proposed unsupervised framework. The project requirement of assigning probabilities to specific anomaly types therefore necessitated the construction of a controlled synthetic anomaly generation pipeline. Given that the autoencoder detects deviations from a learned nominal manifold using a per-signal moving window mean squared error (MW-MSE), the introduction of synthetic perturbations provides the only practical reference for assessing detection sensitivity and downstream anomaly characterization. Injecting these predefined deviations into otherwise healthy valve-closing transitions establishes a consistent baseline against which reconstruction error behavior and latent-space structure can be analyzed.

To preserve interpretability and avoid confounding effects, a strict single-anomaly-per-window constraint was enforced. Each anomalous window contains at most one injected perturbation affecting a single signal, preventing overlap between multiple failure mechanisms within the same sample. This constraint ensures that changes in reconstruction error and latent representation can be attributed to a single, isolated deviation rather than to interactions between concurrent anomalies. In addition, anomaly assignments were distributed uniformly across anomaly types and injectable signals, preventing systematic bias toward any particular fault mechanism during evaluation.

Anomaly injection was implemented using a controlled random assignment procedure. For each operational regime (pump and turbine), windows were shuffled using a fixed random generator to ensure reproducibility, and a fixed fraction (20%) was selected for anomaly injection. Each anomaly-signal combination was enforced to appear at least once before repetition. For each selected window, the target signal was MinMax-normalized to the interval $[0, 1]$, perturbed using parameters sampled from fixed ranges specified in the implementation, and then denormalized back to its original scale. This procedure ensures that injected anomalies span the full range of configured parameter values while remaining homogeneous across the dataset, allowing consistent use of synthetic anomalies as reference points for latent-space analysis and type inference.

The following table summarizes the mathematical perturbations used and their corresponding physical justifications within the context of a high-pressure hydropower plant.

These anomalies are specifically restricted to the closing sequence (indices [180, 360)) of the 360-second windows. This ensures that the model is tested specifically on its ability to distinguish anomalous transients from the expected physical dynamics of the valve during its most critical operational phase.

TABLE 5
Physical Justification of Synthetic Anomaly Types

Anomaly Type	Mathematical Logic	Physical Interpretation
Spikes	Short-duration Gaussian bursts	**Electronic Interference:** High-frequency sensor noise or electromagnetic interference (EMI) from turbine generators affecting PLC inputs.
Level Shift	Constant bias shift	**Sensor Miscalibration:** Abrupt offset in pressure transducers or encoder mounting instability.
Linear Drift	Gradual additive ramp	**Progressive Wear:** Slow hydraulic fluid leaks or gradual seal degradation causing steady pressure loss.
Variance Change	Scaled signal noise floor	**Mechanical Looseness:** High-frequency vibration or fluttering in guide vanes due to bearing wear.
Sinusoidal	Periodic oscillatory signal	**Hydraulic Instability:** Pressure pulsations or surging within the penstock during valve transitions.
Delayed Closure	Temporal shift of transition	**Mechanical Friction:** Debris accumulation or silt in the valve housing slowing the mechanical sequence.
Water Hammer	Intensity-scaled pressure peak	**Transient Overpressure:** Dangerous pressure surge caused by an excessively rapid flow interruption.
Signal Dropout	Segment forced to zero	**Connectivity Loss:** Intermittent cable faults or complete loss of power to specific sensor loops.

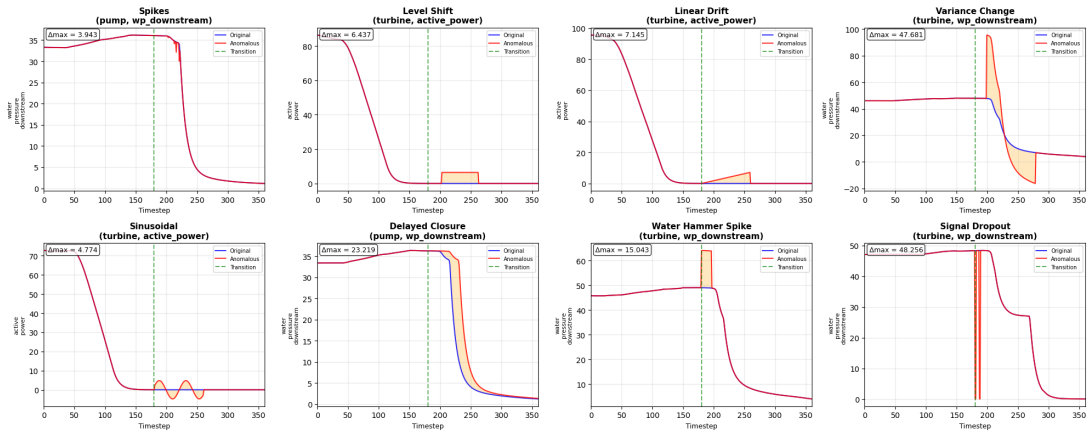


FIGURE 14
Examples of generated synthetic anomalies

4.4 Anomaly Scoring Logic

The objective of the scoring phase is to convert the autoencoder reconstruction error into a reliable decision metric. Instead of relying on a single global error over the full 360 s window, we use a Moving Window Mean Squared Error (MW-MSE) to capture localized temporal deviations during valve transitions that would be attenuated by global averaging.

Moving Window Mean Squared Error (MW-MSE). For each of the six signals, the mean squared error is computed over a sliding window of size $W = 30\,\text{s}$. At each time step t, the windowed error is defined as

$$\text{MSE}_{\text{sig}}(t) = \frac{1}{W} \sum_{i=t}^{t+W-1} \left(x_{\text{sig},i} - \hat{x}_{\text{sig},i}\right)^2. \tag{1}$$

The final score for a given signal is obtained by averaging these windowed errors over the full 360 s window. This preserves sensitivity to transient effects such as spikes or drifts that are typical of mechanical valve faults.

Per-signal thresholding (OR-basis logic). Because the six physical signals operate on different scales and exhibit distinct noise characteristics, a single global threshold is not appropriate. We therefore define a threshold for each signal as the 97.5th percentile of its MW-MSE distribution on healthy training data.

An event is flagged as anomalous using an OR-basis rule: a window is classified as anomalous if the error of any signal exceeds its corresponding threshold,

$$\text{Flag} = \begin{cases} 1, & \text{if } \exists\, \text{sig} \in \{1, \ldots, 6\} \text{ such that } \text{MSE}_{\text{sig}} > \text{Threshold}_{\text{sig}}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This rule maximizes sensitivity by ensuring that failures affecting a single component, such as a localized pressure surge or a delayed guide vane response, are not masked by the stability of the remaining signals.

4.5 Anomaly Classification and Type Inference

Once a window is flagged as anomalous by the OR-basis MW-MSE logic, its associated 128-dimensional latent vector is added to a classification dataset. This dataset combines unlabeled real test samples with labeled synthetic anomaly seeds. We use the full 128 dimensions before applying any dimensionality reductions to not loose valuable information.

Hierarchical clustering (HDBSCAN). A small minimum cluster size (`min_cluster_size` $= 3$) is selected to accommodate the limited number of real anomalous samples and to avoid discarding rare but repeatable synthetic failure patterns. Given the scarcity of labeled industrial faults, this setting represents a pragmatic trade-off between cluster stability and sensitivity.

Rationale for sub-clustering. Primary HDBSCAN clusters are further refined through sub-clustering for three main reasons. First, a coarse cluster may represent a broad class of anomalies, such as pressure-related faults, while sub-clustering separates distinct mechanisms (e.g., water hammer spikes versus linear pressure drift). Second, compressing multiple signals into a single latent vector can cause different faults to partially overlap in reconstruction error, and sub-clustering helps isolate signal-specific structure captured by the encoder. Third, smaller and more homogeneous groups improve the reliability of subsequent inference, as clusters dominated by a single synthetic type form clearer diagnostic signatures.

k-NN type inference. Final anomaly classification is obtained using a k-Nearest Neighbors algorithm with $k = 5$ applied within the clusters/(sub)clusters. For each flagged real test sample, the five closest synthetic neighbors in latent space are selected. The anomaly type is assigned by majority vote, and a confidence score is derived from the proportion of neighbors sharing the same label. This step converts unsupervised detection into a probabilistic attribution to candidate failure archetypes.

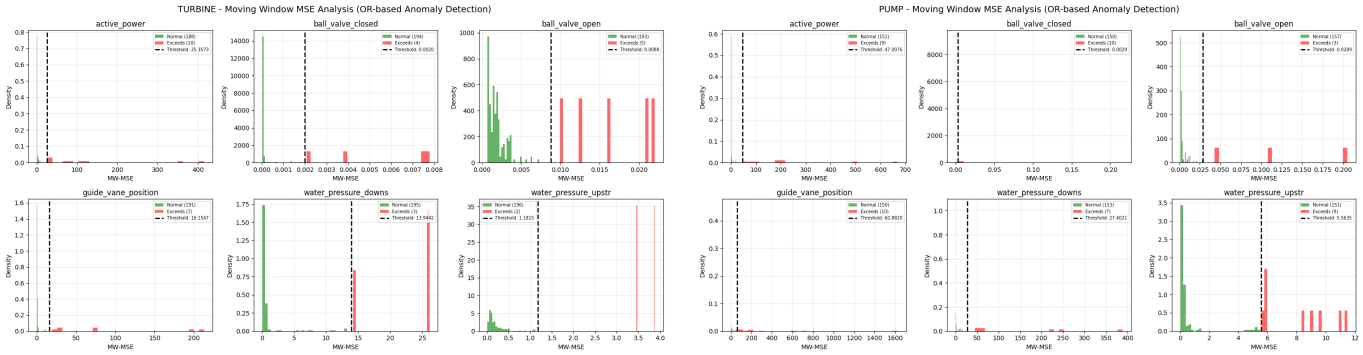
In the turbining regime, the 12 detected anomalies are primarily attributed to spikes, delayed closures, and pressure-related events, with confidence values mostly between 40% and 60%. In the pumping regime, a stronger bias toward spike-like anomalies is observed, often with higher confidence, although scores remain moderate overall. Across both regimes, no anomaly reaches high-confidence attribution, confirming that while the latent space enables coarse categorization, it does not support precise physical fault identification.

4.6 Results and Evaluation

Detection performance. The detection framework was evaluated on 198 turbine transitions and 160 pump transitions. Per-signal thresholds were defined as the 97.5th percentile of reconstruction errors computed on healthy training data. Using this criterion, anomalies were detected in 6.1% of turbine events and 12.5% of pump events. The higher detection rate observed in pumping is consistent with the larger signal variance typically associated with pumping operations. This difference is also reflected in training performance, as the pump model converged to a higher validation loss (0.003029) than the turbine model (0.002578), partly due to the smaller number of available training samples.

Threshold sensitivity was driven primarily by active power, guide vane position, and downstream pressure, which are the most informative signals during valve closure. The percentile was tuned to ensure the detection of all injected synthetic anomalies (F1 = 1.00 for turbine and 0.996 for pump). This ensured that synthetic anomalies could be used as reference points in the latent space without ambiguity.

Per-signal detection statistics. Tables 6 and 7 summarize thresholds, mean MW-MSE, and exceedance rates for each signal. The flagged events constitute the subset of transitions passed to latent-space clustering and type inference. While detection is robust, the limiting factor is the subsequent attribution of multi-signal deviations to specific failure modes with high confidence.



((A)) Turbine regime testing results.

((B)) Pump regime testing results.

FIGURE 15
Detection evaluation summaries for turbining and pumping regimes.

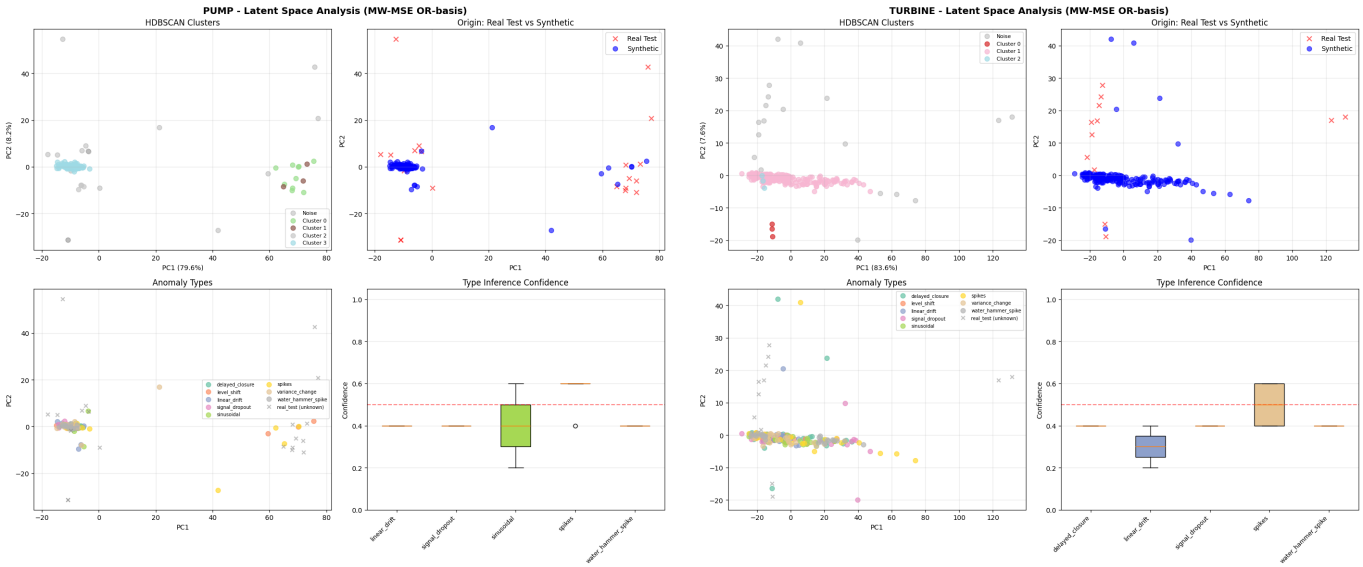
TABLE 6
Turbine Detection Results (198 Test Samples)

Signal Name	Threshold	Mean MW-MSE	Exceedance Rate
active_power	25.357	8.136	5.1% (10/198)
ball_valve_closed	0.002	0.00019	2.0% (4/198)
ball_valve_open	0.009	0.0022	2.5% (5/198)
guide_vane_position	16.155	4.043	3.5% (7/198)
water_p_downstream	13.944	1.085	1.5% (3/198)
water_p_upstream	1.181	0.242	1.0% (2/198)
Overall Detection			**6.1% (12/198)**

TABLE 7
Pump Detection Results (160 Test Samples)

Signal Name	Threshold	Mean MW-MSE	Exceedance Rate
active_power	47.008	15.074	5.6% (9/160)
ball_valve_closed	0.003	0.0025	6.2% (10/160)
ball_valve_open	0.029	0.0056	1.9% (3/160)
guide_vane_position	60.882	28.039	6.2% (10/160)
water_p_downstream	27.402	8.344	4.4% (7/160)
water_p_upstream	5.563	0.884	5.6% (9/160)
Overall Detection			**12.5% (20/160)**

Latent-space clustering results. Clustering was performed on the latent representations of all windows flagged as anomalous by the OR-basis MW-MSE detector. In turbining mode, 203 windows (12 real, 191 synthetic) were clustered into three main groups, with 9.4% of samples labeled as noise. In pumping mode, 140 windows (20 real, 120 synthetic) formed four clusters, with 11.4% noise. While synthetic anomalies formed dense and repeatable structures, real test anomalies were sparsely distributed across clusters and frequently assigned to noise, indicating limited alignment between real operational deviations and the predefined synthetic failure archetypes. PCA projections of the latent space for both regimes are shown in Figure 16, highlighting overlap between synthetic classes and weak separability of real anomalies.



((A)) Pumping regime: PCA projection of latent space (main clustering).

((B)) Turbining regime: PCA projection of latent space (main clustering).

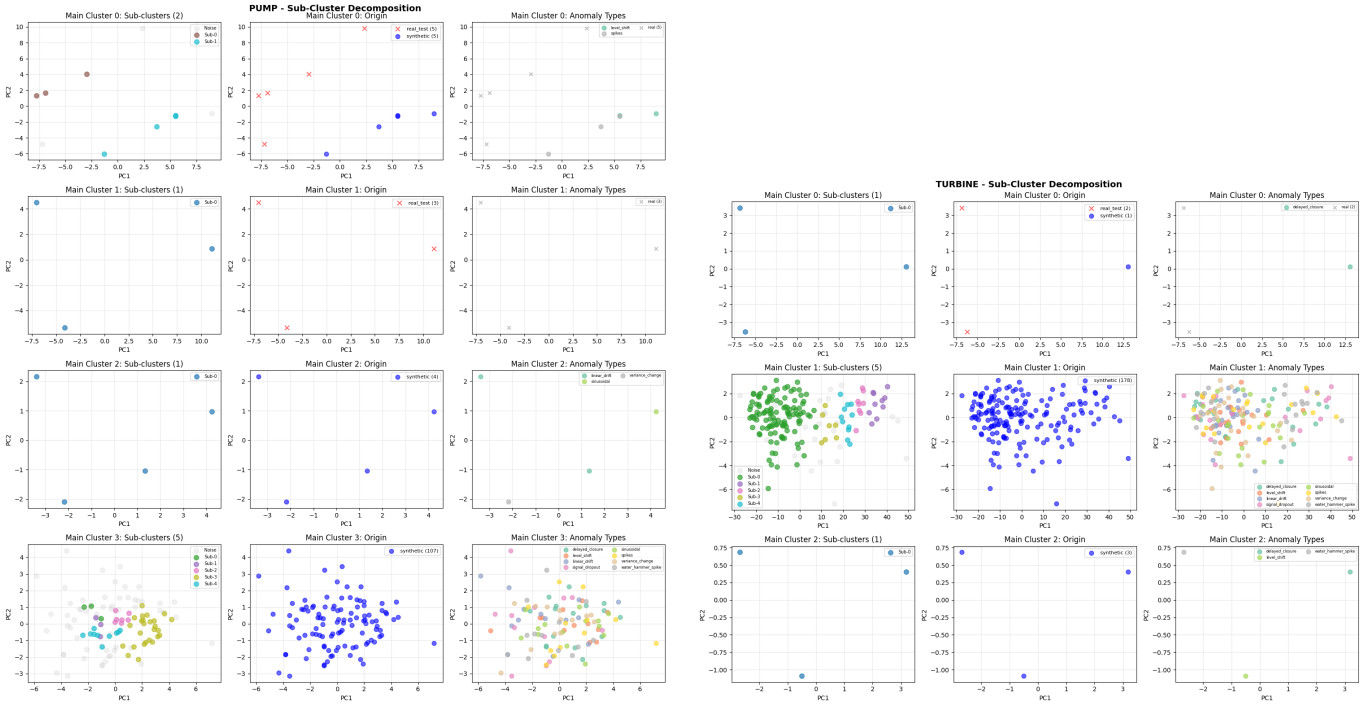
FIGURE 16
Latent-space PCA visualizations after HDBSCAN clustering.

TABLE 8
Main clustering statistics

Regime	Flagged windows	Real	Clusters	Noise (%)
Turbine	203	12	3	9.4
Pump	140	20	4	11.4

Sub-clustering evaluation. To assess whether finer structure could be recovered, sub-clustering was applied within sufficiently large main clusters. In the turbining regime, the dominant cluster (178 samples) decomposed into five sub-clusters and a noise group, all composed exclusively of synthetic anomalies. The largest sub-cluster grouped level shifts, linear drifts, and variance changes, while smaller sub-clusters mixed water hammer, signal dropout, delayed

closure, and sinusoidal patterns. In pumping mode, sub-clusters were smaller and less compact, with a substantial fraction of samples assigned to noise. Real test anomalies rarely formed coherent sub-clusters and were typically isolated or labeled as noise. PCA projections after sub-clustering (Figure 17) confirm strong overlap between anomaly types and limited gains in separability.



((A)) Pumping regime: PCA projection after sub-clustering.

((B)) Turbining regime: PCA projection after sub-clustering.

FIGURE 17
Latent-space PCA visualizations after sub-cluster decomposition.

TABLE 9
Sub-clustering outcome summary

Regime	Main clusters	Sub-clusters	Real in sub-clusters	Synthetic mixing
Turbine	3	5	0	High
Pump	4	6	≤ 2	High

Final anomaly type probability attribution. Following clustering and sub-clustering, anomaly types were assigned using k-NN inference in latent space. For each real test anomaly, the final attribution corresponds to the majority label among the $k = 5$ nearest synthetic neighbors, with the confidence score reflecting neighborhood homogeneity. Tables 10 and 11 summarize the resulting probability attribution for turbining and pumping regimes, respectively. The choice of $k = 5$ reflects the minimum neighborhood size observed to yield stable majority votes within dense synthetic clusters, while remaining small enough to avoid excessive smoothing across adjacent failure types.

TABLE 10
Turbine regime anomaly type probability attribution

Test idx	Inferred type	Confidence	Agg. MSE
11	Signal dropout	0.40	107.02
35	Water hammer spike	0.40	27.00
36	Water hammer spike	0.40	26.16
42	Delayed closure	0.40	9.29
57	Signal dropout	0.40	10.74
67	Spikes	0.40	6.71
84	Linear drift	0.40	6.92
112	Delayed closure	0.40	10.22
124	Spikes	0.40	95.26
165	Linear drift	0.20	9.65
196	Spikes	0.60	28.30
197	Spikes	0.60	26.63

TABLE 11
Pump regime anomaly type probability attribution

Test idx	Inferred type	Confidence	Agg. MSE
2	Spikes	0.60	3.82
3	Spikes	0.60	24.23
4	Spikes	0.60	5.52
5	Spikes	0.60	6.01
7	Spikes	0.60	152.67
8	Spikes	0.60	16.63
12	Spikes	0.60	25.67
13	Spikes	0.60	31.84
14	Spikes	0.60	256.92
15	Spikes	0.60	4.13
27	Linear drift	0.40	40.59
43	Signal dropout	0.40	397.42
55	Spikes	0.40	21.72
63	Linear drift	0.40	39.36
77	Sinusoidal	0.60	15.52

5 Conclusion

This project developed a data-driven anomaly detection framework for spherical ball valve opening and closing operations using multivariate industrial time-series data. A preprocessing pipeline was implemented to address heterogeneous and irregular sampling rates, including temporal standardization, segment-wise resampling, forward interpolation, and exponential moving average smoothing. Exploratory analysis of turbine operating modes revealed physically consistent relationships between active power, guide vane position, and penstock pressures, validating the quality of the processed data and motivating the choice of modelling.

Valve opening and closing events were detected without relying on Boolean valve status signals by combining rolling-average–based heuristics with a Temporal Convolutional Network that inferred probabilistic valve states. While this approach successfully identified most transient events and enabled estimation of opening and closing durations, it exhibited notable shortcomings. In particular, the initial event flagging strategy produced a significant number of false positives, necessitating additional filtering based on downstream pressure. Even after refinement, accurately determining the precise start and end of valve transients remained challenging due to subtle signal changes and oscillatory behavior during hydraulic transitions. These difficulties are reflected in the event duration estimation errors,

which, although centered near zero on average, occasionally reached deviations of up to 10 seconds. Such outliers may negatively impact the reliability of downstream anomaly detection.

For transient anomaly detection and characterization, a multivariate autoencoder combined with density-based clustering in latent space was employed to model normal valve-closing behavior across six correlated signals. While this approach effectively captured cross-signal dependencies and enabled robust detection of anomalous transitions, its performance in anomaly characterization was limited. Real test anomalies were frequently isolated or assigned to noise during HDBSCAN clustering, and sub-clustering primarily decomposed synthetic anomalies rather than revealing coherent structures among real events. This indicates that, although the latent representation is well suited for separating normal and abnormal behavior, it does not provide sufficient separability to reliably distinguish between fine-grained physical failure mechanisms in real operational data.

Overall, while the proposed framework demonstrates strong potential for predictive maintenance of valve systems, its performance is currently constrained by data quality, transient estimation accuracy, and reliance on synthetic anomalies. Addressing these limitations would improve robustness, reduce false positives, and move the method closer to deployment in real-world monitoring scenarios.

A Required Python packages/ Libraries

- numpy
- pandas
- glob
- os
- pickle
- ipywidgets
- IPython.display
- matplotlib.pyplot
- torch
- sklearn
- random
- tqdm
- hdbscan
- seaborn
- sklearn.neighbors
- scipy
- pyarrow

B Contribution of team members

- Matas Jones: Data analysis & Task 1
- Shreyas Nara: Task 2: Exploration of models, generation of synthetic anomalies, anomaly detection methods
- Georg Schwabedal: Task 2: Data preprocessing, training and evaluation of implemented models, generation of synthetic anomalies, anomaly classification
- Benjamin Bahurel: Task 2: Implementation of models and training pipeline