CIVIL-426: Machine Learning for Predictive Maintenance Applications

Final Project – Data Challenge

Ball Valve Anomaly Detection

Version 0.2

October 2025

Introduction

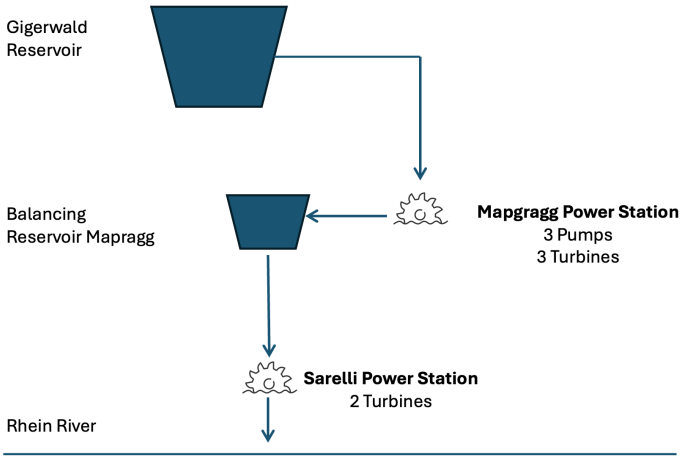
Axpo at a glance

Axpo is Switzerland's largest power producer and the country's largest producer of renewable energy. With 7,000+ employees, 100+ years of experience and activities in 30+ countries, Axpo leads in hydropower within Switzerland with an annual production of 9TWh per year and 4'300MW installed capacity spread across 60 hydropower plants. In the following tasks, you will receive data from the hydropower plant KSL.

Intro into KSL

Kraftwerke Sarganserland AG (KSL) is a high-pressure hydropower scheme in the Swiss canton of St. Gallen that has been generating electricity since the 1970s. The system collects water from the Calfeisen and Weisstannen valleys into the Gigerwald reservoir and uses a two-stage configuration to produce power at Mapragg (MAP) and Sarelli (SAR) near Vadura and Bad Ragaz.

KSL's upper stage is the Mapragg power station, which operates as a pump-storage facility: it turbines water from the Gigerwald reservoir to the Mapragg balancing reservoir and can also pump water back up from Mapragg to Gigerwald, providing flexible, grid-balancing capacity in addition to energy generation. The lower stage is the Sarelli power station, a conventional storage hydropower plant that turbines the outflows from Mapragg along with additional inflows from the Taminatal catchment, completing the cascade to maximize energy recovery across the scheme.



Hydropower basics and the role of spherical (ball) valves

In hydropower plants, water is conveyed under pressure from the intake to the turbine through penstocks. For high-pressure applications, spherical (ball) valves are commonly used as turbine inlet safety valves to rapidly shut off flow while respecting the maximum admissible pressure in the penstock during quick or emergency

closures. Typically, such spherical valves can close automatically against unidirectional pressure and often feature dual sealing systems to enable maintenance without fully depressurizing the pipe. Rapid closures must be managed carefully due to transient effects like the water hammer, which is essentially a pressure wave travelling up and down the penstock after a pressure change. You can check out how a ball valve works in this video, the principle remains the same even in large hydropower plants.

Which phenomena we want to detect

We target malfunctions related to ball valves. Seal degradation, e.g., seal wear, debris accumulation, or material aging can cause leakage and altered closing dynamics. These issues are known contributors to ball valve failures. In time-series data, this can manifest as a deviation in the pressure decay during valve closing events, such as elevated terminal pressure and a longer tail toward steady state. It can also manifest in changes of the opening and closing time of a spherical valve.

Picture of damage to a sealing ring of a ball valve due to a permanent water stream resulting from a leak eroding the sealing ring.

Challenge objective

Design an **unsupervised** anomaly detection model that:

- Learns normal closing behavior across multiple valves and sensors.
- Scores each closing event for deviation from normal.
- Robustly detects the target deviation with low false alarms and minimal detection delay.
- Has a well generalizing approach (i.e. domain adaptation is theoretically feasible)

Tasks

This final project consists of several tasks and emulates a real-world situation. An estimated importance of each task in the final grading is given in brackets. Each task must be documented in your report.

Exploratory Analysis [~10% of report]

Start with a global exploratory analysis to get familiar with the provided datasets. In particular, pay attention to: the different types of sensors, different sampling, missing data, correlation between variables, preprocessing steps etc. It is your task to read, resample, preprocess the data before analysis. Summarize your findings and highlight which findings led to your final solution.

Task 1 [~20% of report]

Your warmup task is to detect outliers among spherical valve closing and opening time. Observing the closing/opening time of spherical valves is crucial to determine the condition of a spherical valve.

Your task is to write a model that first computes the opening and closing times of the spherical valve and then detects anomalies. In this task, you **should not** use the real valve signals (ball_valve_closed & ball_valve_open).

Time Calculation:
For each cycle, calculate:
The time taken for the valve to open (from closed = 1 to open = 1).
The time taken for the valve to close (from open = 1 to closed = 1).

Outlier Detection:
Analyze the list of opening and closing times. Use a statistical method (such as the interquartile range or Z-score) to identify outliers—cycles where the times are unusually long or short.

Attention: Do **not** use ball valve signals in this task

Task 2 [~70% of report]

Anomaly Detection

Your main task is to detect outliers among transients of ball valve closings. You will be provided with a training dataset of measurement data for multiple valves in different hydropower plants. The training set represents an anomaly-free, healthy data, whereas the test set contains anomalies.

Imputation of different outliers that capture faults of the kind found in the test dataset is up to you.

Types of Anomalies

In the test data you can expect different degrees of the ball valve not closing entirely, not closing fast enough and further anomalies. One of the anomalies is a real world anomaly observed in one of our hydropower plants due to a defective seal on a ball valve.

Generate Synthetic Anomalies

In order to train your models, you'll have to generate synthetic anomalies with the help of the provided python script (described below). You can also think of your own anomalies to improve the anomaly detection.

Provided Data

signal_descriptions.csv

This is the CSV file containing the metadata for the measurement data of six signals for each of the five machine groups in KSL, three in Mapragg and two in Sarelli. Each signal can be uniquely identified by its signal_id both in the measurements parquet file for each machine group and the signal_descriptions.csv file.

signal_name	description_english
active_power*	Active power measurement
ball_valve_closed	Ball valve position rotating body closed (true if valve is fully closed)
ball_valve_open	Ball valve position rotating body open (true if valve is fully open)
guide_vane_position	Guide vane position measurement
water_pressure_downstream	water pressure downstream of the turbine
water_pressure_upstream	water pressure upstream of the turbine

*For Sarelli active_power is always negative. This is simply a convention and Sarelli only supports turbine (i.e. power generation) operation. In Mapragg, negative active power means pumping operation, while positive active power means turbine operation.

Pumping Operation Start

Pumping operation in Mapragg is jumpstarted by the turbine since both the pump and turbine are on the same axis. This can be seen by both water_pressure_downstream and water_pressure_upstream increasing at the beginning of the pump operation and water_pressure_downstream then decreasing again as active_power becomes negative. This phenomenon can be seen at the beginning of every pumping operation.

Synthetic anomalies

generate_anomalies_template.py

We provide the python script to add synthetic anomalies to the real data that should resemble anomalies that could happen in reality. The script also integrates the ground truths for all these anomalies in order to verify your results. The anomalies implemented for now are specified as follows:

Type of anomaly	Description
Drift of sensor values	Certain sensor values are increased linearly for a period of N days, then back to normal
Offset on sensor	Add an offset on the sensor for a duration of N days
Your own anomalies	You are encouraged to implement your own type of anomalies and add it to sensor signals

The synthetic sets contain synthetic anomalies that should be more easily detectable and can be used to calibrate your models. However, do **not** tune any decision thresholds to detect anomalies on the synthetic sets, since you cannot be sure that the real anomaly follows the same pattern. In cases this is done nonetheless, we will penalize the final grading.

Hint: You can also define your own anomaly functions

Attention: Do **not** tune any decision thresholds to detect anomalies on the synthetic sets

Real plant data

Mapgragg_MGX_training_real_measurements.parquet,
Mapgragg_MGX_testing_real_measurement.parquet,
Sarelli_MGX_training_real_measurements.parquet,
Sarelli_MGX_testing_real_measurements.parquet

Each of the files contains four columns:

- **plant**: the power plant (in your case that's always KSL)
- **signal_id**: a unique identifier for each signal. You will find six unique signal IDs in each file. By joining the signal descriptions (described below) onto the measurements, you can get the human-readable name of the signal and a description of what it measures.
- **value**: measured value
- **ts**: timestamp at the time of measuring

The signals have not yet been resampled. Whenever a signal changes, a new value is saved with the timestamp of the change and the new value. Hence, the values in the parquet are not uniformly sampled. This means it is up to you to transform the raw data such that you can use it to train and score your model.

Hint: When resampling to a uniform sampling rate, keep data leakage in mind when choosing the resampling method.

General Tips

Tip for Task 2: Modelling

The main idea in anomaly detection is to model the normal behavior and define anomalies as observations that are unlikely under this model. There are several ways to model it mathematically. Moreover, there is a distinction between control variables (linked to the operation of the plant or the environment, denoted X) and the generator variables (denoted Y). There is a clear causality relationship X → Y (no independence).

Strategy 1 : Modeling the conditional distribution p(Y|X)

One strategy is to model the relationship between control and measurements, i.e. the distribution p(Y|X). It can be seen as sensor modeling.

Point mapping: $X_t \rightarrow Y_t$

Sensor forecasting (autoregressive) :

$$Y_{t-1} \rightarrow Y_t X_t,\ X_{t-1},\ Y_{t-1} \rightarrow Y_t X_{t-k},\ \cdots,\ X_{t-1},\ Y_{t-k},\ \cdots,\ Y_{t-1} \rightarrow Y_t$$

Strategy 2 : Modeling the joint distribution p(X, Y) – Unsupervised learning approach

In this strategy, we don't assume such a relationship and model the normal joint distribution of (X, Y) or the marginal distributions of X and Y, using an unsupervised approach. For example, autoencoders ; clustering ; traditional algorithms for anomaly detection.

Tip for Task 2: Procedure

Further, we also recommend including the following two steps in your process:

(i) Analyze the detected anomalies to provide as many insights as possible; you should:

- Define a meaningful anomaly score.
- Derive the anomaly score for each timestamp.

- Derive a robust decision rule for anomalies (number of time steps over a threshold value).

(ii) Root cause identification

- In case of anomalies, highlight which variables have the highest influence on the anomaly score and give, if possible, a physical explanation for each anomaly.
- The causes of the anomalies may be multiple or might come from deviations of other variables. Develop a model that automatically outputs the most probable cause of a given anomaly.

Timeline: Submission, presentation of results and grading

We expect that you apply the tools, methods and guidelines that you have seen during the class such as the typical steps to follow in a machine learning project.

Introduction

On October 16th at 10:15, Axpo engineers will join us in the exercise session on Zoom to give an overview of the project and explain the data you will be using, directly sourced from the power station. This session will give you key insights into the challenges and context of your work. During this session, form groups of 4 students and fill in your group in this document. After signing the data access form, retrieve the data and materials from this switchdrive (password will be sent after signing the form). After reading carefully this project description, start brainstorming with your team on how you would proceed to detect the synthetic anomalies.

Milestone Submission (Pitch 13.11)

On the **13th November**, we will assess your intermediate understanding of the project. Your objective is to pitch to a TA (1) your approach to tackle the tasks of the final submission based on your data analysis (you present your proposed way to arrive at the goals of the project). Additionally, you have to present your findings on the synthetic anomalies dataset (2), focusing on their detection, and their relevance to real-world anomalies. Based on your results, propose (and justify) one or several anomaly scoring strategies.

- Describe the planned data preprocessing, splitting (train/validation/test), and evaluation scheme.
- Come up with a workflow diagram of your proposed methodology, showing the training and testing phases.

Each group presents and discusses briefly (for 5 minutes) their ideas on the whiteboard to the TAs. The milestone will be graded on a pass/fail basis, so ensure your pitch is clear, concise, and well-structured. Take this opportunity to demonstrate your understanding in the context of our project. You'll also gain valuable feedback on your methodology and approach, which will help you to refine your final

submission. Be aware that, independent of the pitch, we expect you to briefly summarize the steps that led you to your final solution later in the introduction section of your report (whatever you pitch to us also needs to be documented in the final report if it leads to your solution).

Final Submission

Provide a technical report of your findings (PDF), the corresponding (clean and documented) source code with a list of python packages that are required to run your code. Make sure that your final methodology describes only **one** solution approach (f.e. comparisons between different models can go to appendix). Please make your submission on the course Moodle **before 15. December**. The presentation of results is planned for **18. December 2025** at EPFL in front of Axpo engineers and TAs. The presentation itself is based on a conference A0 poster that you create from your methodology and findings and should be 2–3 minutes long. All team members must contribute to the final project. In the final report, provide a detailed explanation of each member's specific contributions.

Please interpret your findings and provide recommendations for further developments. If you had more time to work on it, how would you develop the methodology further? Which other questions would you be interested to analyze?

The grading of the final project will be based on:

1. The milestone pass/fail performance (-0/-0.25 of the final grade)
2. The poster presentation (20%)
3. The submitted report, including code (80%)

Prizes

A jury will evaluate the final presentations of all teams. There will be an award for the best technical performance, creativity, and data storytelling.

Data License

Note that access to the data sets has been provided exclusively for educational purposes, specifically for the data challenge and related tasks outlined in the description. Please note that redistribution, publication and commercial use of the data set and insights specifically linked to the data set are not permitted. After completion of the challenge, at latest by 01 February 2026, all copies of the data set need to be erased. Any use of the data set outside the intended purpose requires prior written consent by Axpo Holding, CH-5401 Baden

Important Key dates

- **16.10 : Project introduction, group forming, data sharing**
- **06.11: Support slot**
- **13.11 : Milestone 1 support slot & project pitches**
- **27.11 : Support slot**

- **04.12 Support slot**
- **11.12 Support slot**
- **15.12 : Submission Milestone 2 (Presentation & Report)**
- **18.12 : Final poster presentations**
- **End of December : grading**
- **01.02.2026 : Axpo data must be erased**

Contacts

- Nicolas Pelzmann nicolas.pelzmann@axpo.com
- Bastian Röhrig bastian.roehrig@axpo.com
- Olga Fink olga.fink@epfl.ch
- Leandro Von Krannichfeldt leandro.vonkrannichfeldt@epfl.ch
- Raffael Theiler raffael.theiler@epfl.ch
- Arthur Bizzi arthur.bizzi@epfl.ch
- Sergei Garmaev sergei.garmaev@epfl.ch
- Vinay Sharma vinay.sharma@epfl.ch

Appendix

I confirm that I have read and understood the data license and intended purpose of the data usage.

Last Name	First name	Place, date	Signature

