CIVIL-426: Machine Learning for Predictive Maintenance Applications

Final Project – Data Challenge

Predictive Analytics for a Hydropower Unit

Version 1.0

October 2024

1 Introduction

Alpiq is a leading Swiss energy company. It operates across Europe and offers their customers comprehensive services in the field of power generation, electricity trading and in the marketing and optimization of energy. Their core business is the generation of power using flexible, zero-carbon electricity via a fleet of Swiss hydropower plants. Alpiq holds shares in 18 hydroelectric companies in Switzerland and manages 2 '910 MW of hydropower. They also generate electricity through wind, photovoltaic and nuclear power as well as several modern and flexible gas-fired combined cycle power plants. Alpiq has an installed capacity of 5 '223 MW, 55% of which is hydropower. Equipped with state-of-the-art technology, their power plants are milestones of Swiss engineering history. Alpiq continuously optimizes and controls them digitally to achieve a high degree of safety and efficiency.

Plant Information

In this data challenge, we will focus on the pumped storage power plant of FMHL (FMHL: Forces Motrices Hongrin-Léman). This plant acts as a water battery. Water is pumped from lake Léman into lake L'Hongrin during low electricity prices and turbined back when prices are higher. Pumped-storage plants do not act as electricity producers but rather as a stabilizing mechanism for the electricity grid.

The FMHL asset is constituted of:

- The L'Hongrin dam
- The waterways between the dam and the power plants
- 2 power plants: Veytaux 1 (4 units) and Veytaux 2 (2 units)

All the units in Veytaux 1 and 2 are ternary units. A ternary unit is composed of a generator, turbine, coupler and a pump sharing the same shaft. The coupler can be used to decouple the pump from the rest of the unit. In turbine mode, the pump is decoupled, water flows through the turbine to generate electricity at the generator output. During pumping mode, the pump is coupled, electricity is imputed to the generator which powers the whole unit and pumps water up to the reservoir. In Veytaux 2, a hydraulic-short circuit mode exists where, in pump mode, some of the pumper water is diverted back to the turbine, this enables control of the pumping power consumed by the machine. Veytaux 1 is composed of 4 horizontal units of 60MW (VG1 to VG4). Veytaux 2 is composed of 2 vertical units of 120MW (VG5 and VG6).

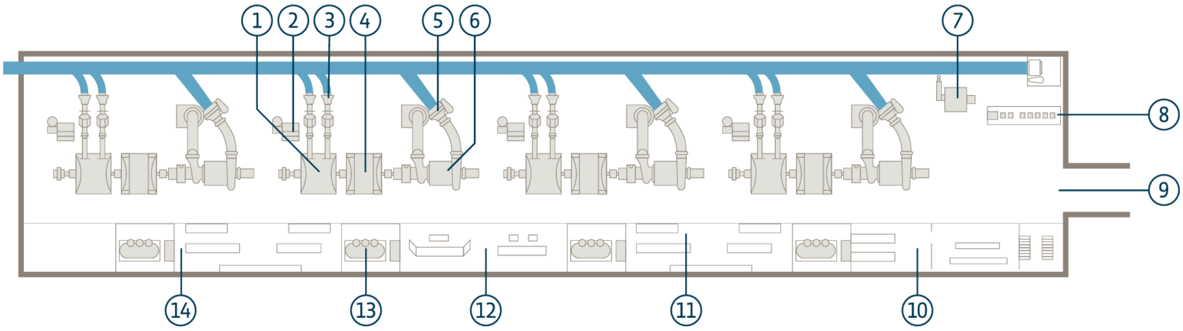


Fig 1: layout of Veytaux 1

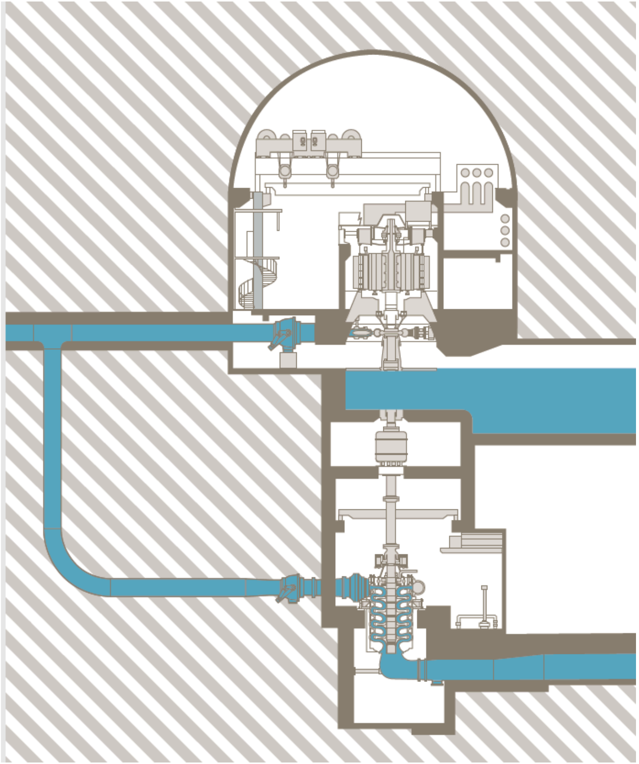


Fig 2: Veytaux 2 unit

Unit Information

This challenge will focus on the thermal state of the generator subsystem of units 4, 5 and 6.

The generator is composed of the following elements:

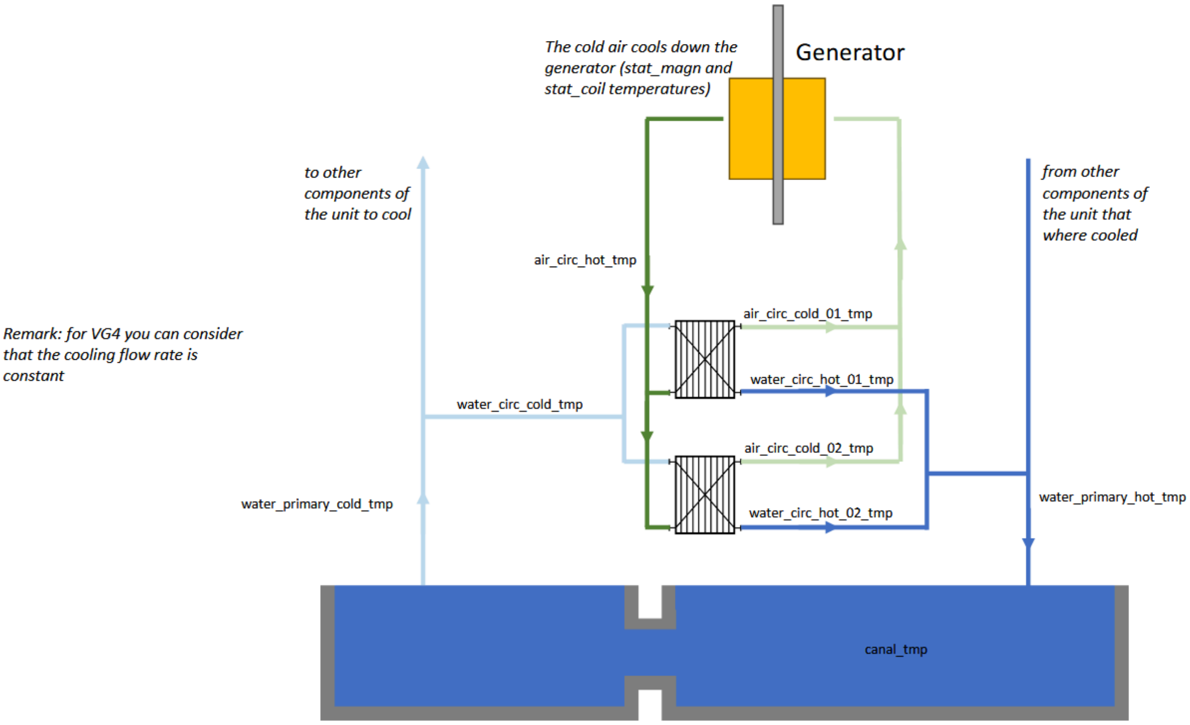
- **Rotor:** The rotor is the rotating component of the generator. It is composed of several electro-magnets that generate the main field flux. The rotor is generally not instrumented due to technical constraints as it is rotating at a couple hundred rpm.
- **Stator:** The stator is the stationary part of the alternator that carries the armature winding and the magnetic circuit. The stator generates the output high voltage field. The stator is heavily monitored with temperature, current and voltage sensors

- **Cooling system:** The generator heats up significantly when in operation and requires constant cooling. The cooling system is generally air/water based:
    - Cold air is pumped at the base of the generator, between the rotor and the stator.
    - The cold air is heated by the generator, collected at the top of the stator, and directed into a heat exchanger.
    - The heat exchanger is water based. It takes as input the hot air from the generator and cold water from the primary cooling system (common for all subsystems of a unit) and outputs cooled air and heated water.
    - The cooled air is pumped back to the generator.

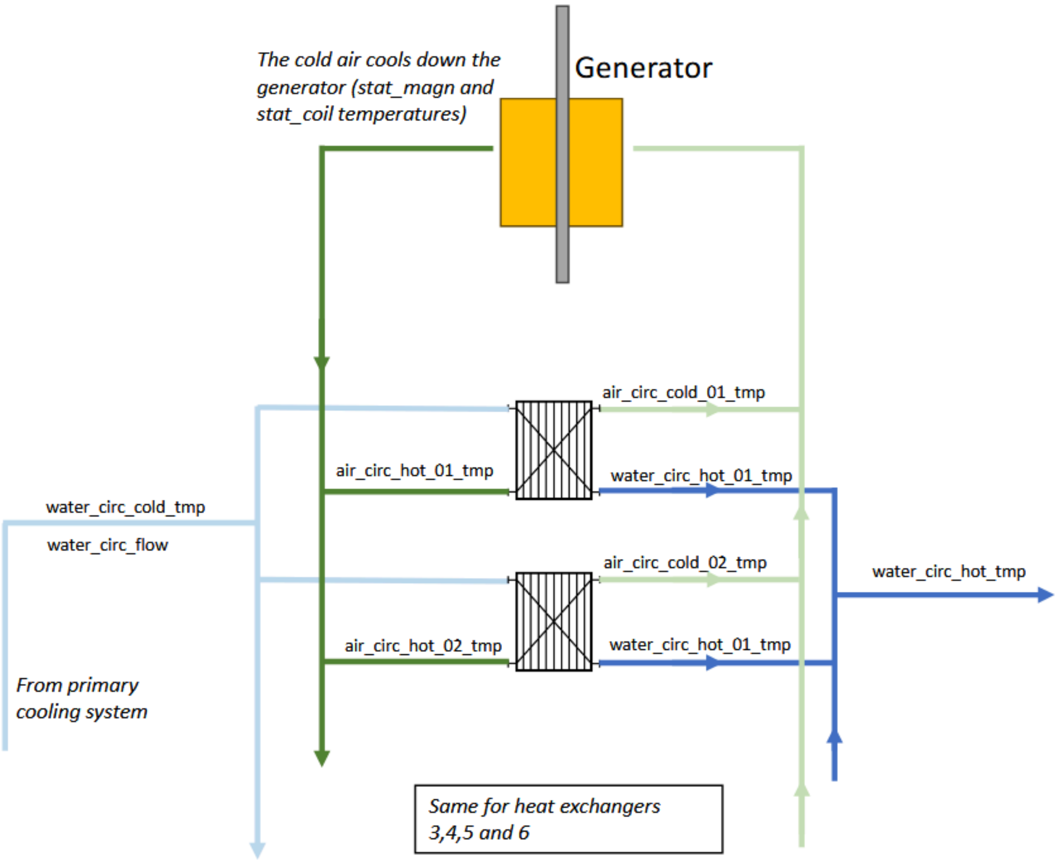
VG4 has a 2 heat exchanger while units VG5 and VG6 have six heat exchangers radially distributed around the generator.

The thermal state of the generator is mostly dependent on the operating conditions of units. The more power produced (or consumed in pump mode) the more heating is generated and the more the cooling system is active. The thermal state of a generator must be heavily monitored as any deviations can lead to severe failures that have important implications in terms of safety, downtimes, and costs.

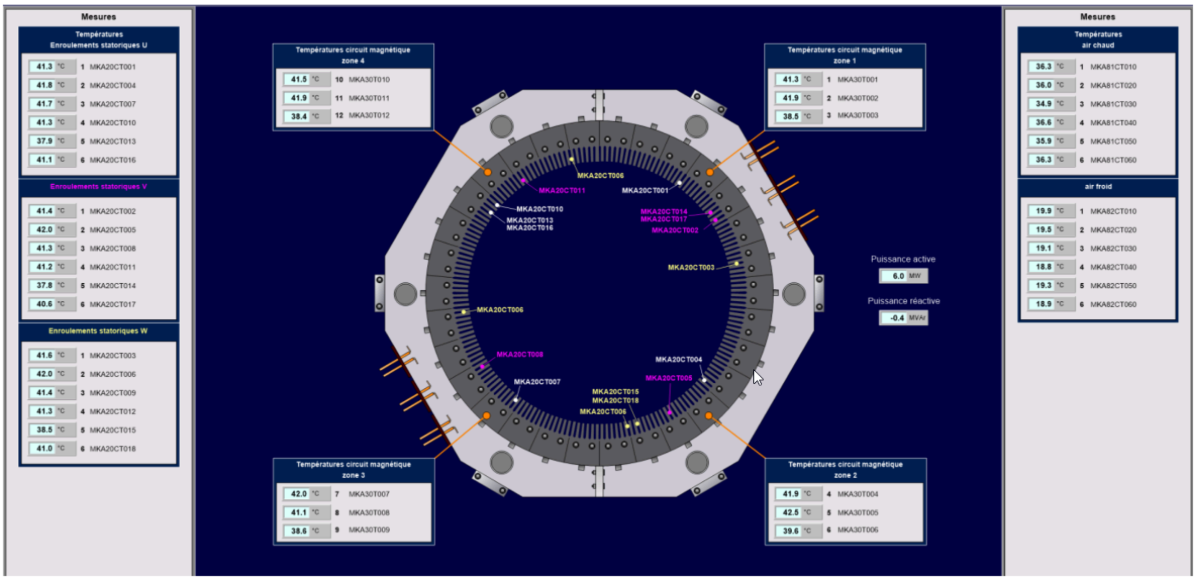
VG4 Cooling System Schematic Overview


The cold air cools down the generator (stat_magn and stat_coil temperatures)
Generator
to other components of the unit to cool
from other components of the unit that where cooled
air_circ_hot_tmp
air_circ_cold_01_tmp
Remark: for VG4 you can consider that the cooling flow rate is constant
water_circ_hot_01_tmp
water_circ_cold_tmp
air_circ_cold_02_tmp
water_primary_cold_tmp
water_circ_hot_02_tmp
water_primary_hot_tmp
canal_tmp

VG5 & 6 Cooling System Schematic Overview



VG5 & 6 Stator Temperatures



Mesures

Températures
Enroulements statoriques U

41.3	°C	1	MKA20CT001
41.8	°C	2	MKA20CT004
41.7	°C	3	MKA20CT007
41.3	°C	4	MKA20CT010
37.9	°C	5	MKA20CT013
41.1	°C	6	MKA20CT016

Enroulements statoriques V

41.4	°C	1	MKA20CT002
42.0	°C	2	MKA20CT005
41.3	°C	3	MKA20CT008
41.2	°C	4	MKA20CT011
37.8	°C	5	MKA20CT014
40.6	°C	6	MKA20CT017

Enroulements statoriques W

41.6	°C	1	MKA20CT003
42.0	°C	2	MKA20CT006
41.4	°C	3	MKA20CT009
41.3	°C	4	MKA20CT012
38.5	°C	5	MKA20CT015
41.0	°C	6	MKA20CT018

Températures circuit magnétique
zone 4

41.5	°C	10	MKA30T010
41.9	°C	11	MKA30T011
38.4	°C	12	MKA30T012

Températures circuit magnétique
zone 1

41.3	°C	1	MKA30T001
41.9	°C	2	MKA30T002
38.5	°C	3	MKA30T003

Puissance active
6.0 MW

Puissance réactive
-0.4 MVAr

Températures circuit magnétique
zone 3

42.0	°C	7	MKA30T007
41.1	°C	8	MKA30T008
38.6	°C	9	MKA30T009

Températures circuit magnétique
zone 2

41.9	°C	4	MKA30T004
42.5	°C	5	MKA30T005
39.6	°C	6	MKA30T006

Mesures

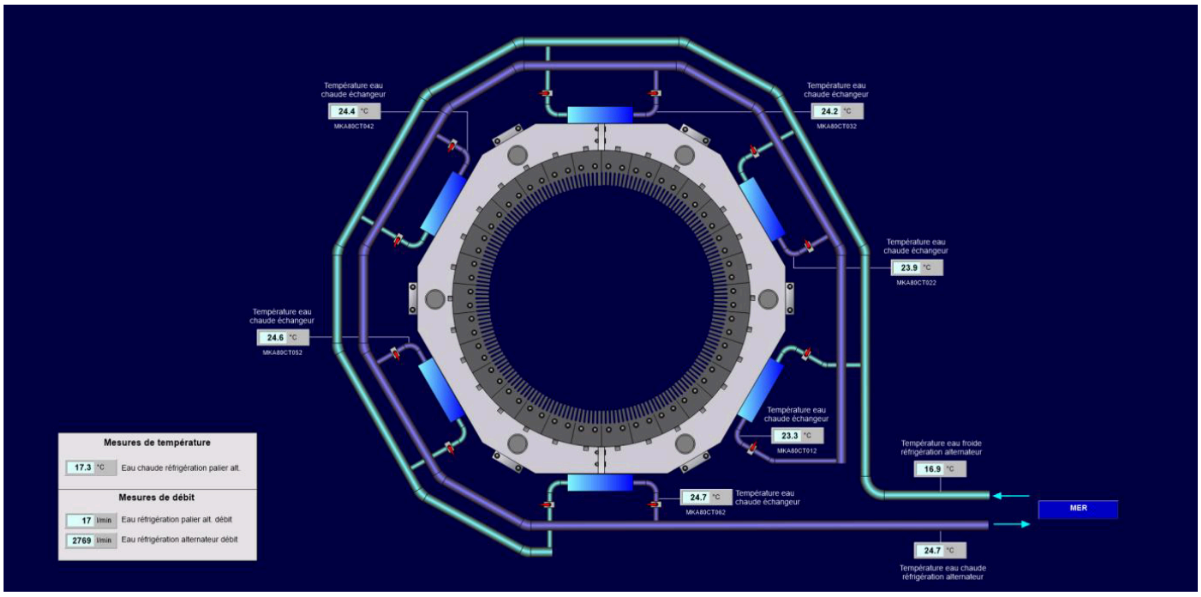
Températures
air chaud

36.3	°C	1	MKA81CT010
36.0	°C	2	MKA81CT020
34.9	°C	3	MKA81CT030
36.6	°C	4	MKA81CT040
35.9	°C	5	MKA81CT050
36.3	°C	6	MKA81CT060

air froid

19.9	°C	1	MKA82CT010
19.5	°C	2	MKA82CT020
19.1	°C	3	MKA82CT030
18.8	°C	4	MKA82CT040
19.3	°C	5	MKA82CT050
18.9	°C	6	MKA82CT060

VG5 & 6 Cooling System


Température eau chaude échangeur
24.4 °C
MKA80CT042
Température eau chaude échangeur
24.2 °C
MKA80CT032
Température eau chaude échangeur
23.9 °C
MKA80CT022
Température eau chaude échangeur
24.6 °C
MKA80CT052
Température eau chaude échangeur
23.3 °C
MKA80CT012
Température eau froide réfrigération alternateur
16.9 °C
24.7 °C
Température eau chaude échangeur
MKA80CT062
MER
24.7 °C
Température eau chaude réfrigération alternateur

Mesures de température

17.3 °C Eau chaude réfrigération palier alt.

Mesures de débit

17 l/min Eau réfrigération palier alt. débit

2769 l/min Eau réfrigération alternateur débit

2 Datasets

You will receive the following datasets:

- VG5
    - Training: 02-01-2020 to 31-12-2020
    - Testing_real: 01-01-2021 to 31-03-2021
    - Testing_synthetic_01: 01-06-21 to 31-07-2021
    - Testing_synthetic_02: 01-11-21 to 31-12-2021
- VG6
    - Training: 02-01-2020 to 31-12-2020
    - Testing_real: 01-01-2021 to 31-03-2021
    - Testing_synthetic_01: 01-06-21 to 31-07-2021
    - Testing_synthetic_02: 01-11-21 to 31-12-2021
- VG4
    - Training: 02-01-2020 to 31-12-2020
    - Testing_real: 01-06-2021 to 31-08-2021

For each unit there is a 1 year long training set and 1 to 3 test sets. The testing_real sets contain real anomalies that happened during the operation of the plant. The testing_synthetic sets contain synthetic anomalies that should be more easily detectable. These sets can be used to calibrate and optimize your models. The data will be provided as Parquet (.parquet) files. The temporal resolution is 30 seconds.

The datasets contain all the sensors located on or related to the state of the generator which includes:

- **Voltage and currents**
    - exc_freq, exc_current, exc_voltage: excitation current, voltage and frequency
    - ph_current and ph_voltage: current and voltage of the 3 different phases
    - elec_freq: frequency of the electrical output
- **Stator temperature measurements**

    - stat_coil_tmp: temperatures of the stator coils
    - stat_magn_tmp: temperatures stator magnetic circuit
- **Cooling system state**
    - air_circ_hot_tmp and air_circ_cold_tmp: hot and cold temperatures of the cooling system
- **Heat exchanger**
    - water_circ_flow: cooling water flow
    - water_circ_hot_tmp and water_circ_cold_tmp: hot and cold temperatures of the cooling water

In addition to the generator measurements the control signals and the operating conditions defined below are also given. The control signals can give you a representation of the state of the unit. The different operating modes can be computed based on the configuration of needle opening, rotational speed, and active power.

- **Control time series:**
    - tot_activepower: power measurement at the generator
    - ext_tmp and plant_tmp: the exterior and plant temperatures
    - injector openings (more information below)
    - pump_rotspeed: rotational speed of the pump
    - turbine_rotspeed: rotational speed of the turbine
    - turbine_pressure (VG5 et VG6): pressure at the input of the turbine
    - injector_01_pressure and injector_02_pressure (VG4): pressure at the input of each turbine of Veytaux1

- **Operating modes:**
    - turbine_mode (and equilibrium_turbine_mode):
        - P>0MW
        - equilibrium where needle position stabilized
    - pump_mode (and equilibrium_pump_mode):
        - P~-60MW for VG4
        - P~-120MW for VG5 and VG6
        - equilibrium when rotational speed stabilized
    - short_circuit_mode (and equilibrium_short_circuit_mode)
        - -120 MW < P < 0 MW, only for VG5 and VG6
        - equilibrium where needle position stabilized
    - machine_on:
        - turbine_rospeed>0
    - machine_off:
        - P=0 MW, turbine_rospeed=0
    - dyn_only_on:
        - all the dynamic transients between equilibrium operating modes

Additional info:

- **Injector opening:** The Pelton turbines in these units are driven by 5 water jets exiting from 5 different nozzles. The opening of these nozzles can be monitored by 5 independent time series. Depending on the rated power that the turbine must produce, either all or only a subset of injectors are opened, and each injector opening can range from 0 to 100%. In order to help you in your task, we recommend you compute a feature representing the scaled sum of all injector openings. This feature can be seen as an input or control feature in your models.

4 Preliminary work [Monday 20.11-27.11 sessions]

Thursday 10.10

On October 10th at 11:00, Alpiq and Hydro 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 3 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 (see section 5.B).

5 Tasks [17.10 - 12.12]

This final project consists of several tasks. An estimated importance of each task in the final grading is given in brackets. Each task must be documented in your report.

The training datasets (`VGX_generator_data_training_measurements.parquet` and `VGX_generator_data_training_info.csv`) contain sensor data under normal operating conditions. You can start with an analysis of these datasets and then use these datasets to build anomaly detection models based on the different signals.

On the other hand, the test sets may contain anomalies. In particular,

- The “synthetic” test sets (*`synthetic_anomalies/`*) contain synthetic anomalies (described below) that have been added artificially into the data. These anomalies should be more easily detectable and will help to guide you in designing your anomaly detection strategy.
- The “real” test sets contain real operational data of the plant `VGX_generator_data_testing_real_measurements.parquet` and `VGX_generator_data_testing_real_info.csv`). For each unit, it may or may not contain anomalous behaviors.

A Exploratory data analysis [~10% of the report]

Start with a global exploratory analysis to get familiar with the provided datasets. In particular, pay attention to: the different types and groups of sensors ; missing data ; the different operating modes ; correlation between variables ; preprocessing steps etc. Summarize your findings and highlight which findings led to your final solution.

You are provided you a notebook with data exploration elements: `01_data_exploration.ipynb`

B. Anomaly detection – steady-state [~60% of the report]

The main task of this final project is anomaly detection on the hydropower units. Each plant unit should be treated independently from each other.

As a first step, consider only the steady-state regime and not the transient regime, i.e. when the plant is operating at equilibrium. This information is already provided as flags in the data : *dyn_only_on*, *equilibrium_turbine_mode*, *equilibrium_pump_mode*, *equilibrium_short_circuit_mode*.

You may also consider anomalies in the transient phases. This is considered optional.

Potential strategies for anomaly detection

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.

Synthetic anomalies

Considered files: `VG5-6_generator_data_training_info.csv ; VG5-6_generator_data_training_measurements.parquet ; all data in synthetic_anomalies/`

We provide a total of 12 testing sets for units VG5 and VG where synthetic anomalies have been added that should resemble anomalies that could happen in reality. In total, each set contains between 1 and 3 anomalies. You are given the ground truths for all these anomalies in order to verify your results. The anomalies are specified as follows:

Type of anomaly	Description
Drift of sensor values	Certain sensor values are increased linearly for a period of ~10 days, then back to normal (for example, +0.5°C/day on temperatures).
Offset on temperature sensors	+5°C on different temperature sensors after each startup of the machine, for a duration of 5-10 days

You don’t have to test every case. The goal is to guide you and allow you to evaluate and show the limitations of your method by evaluating it against the different synthetic faults.

Real plant data

Considered files: `VG4-5-6_generator_data_training_measurements.parquet ;` `VG4-5-6_generator_data_training_info.csv ;` `VG4-5-6_generator_data_testing_real_measurements.parquet;` `VG4-5-6_generator_data_testing_real_info.csv`

(i) Perform anomaly detection on the test sets for steady-state conditions.

Hints:

- Remember to consider the operating conditions when analyzing the dataset. You might want to define anomaly detection models individually for each operating mode, or for multiple operating modes combined.
- As detailed above, some of the signals correspond to operating conditions of the unit and some are measurements that will vary based on these operating conditions. Take some time to look at the signals and think about the correlations between them.
- All the signals are related to the generator sub-system. However the signals can further be divided into groups that follow similar behaviors. Think about how these signals should be grouped. Can you model all of them in conjunction or should you separate them into different sub-models?

(ii) 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).

(iii) 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.

The current monitoring of the units at Alpiq relies on fixed threshold alarms that notify the technician when the value of a signal goes above a predefined value for at least a minimum duration. There is usually two thresholds:

- High limit where the technician is notified.
- Too high limit that acts as a mechanical protection to the machine and automatically shuts the generator off, if the signal reaches this threshold.

You will find the threshold values for the different signals in the table below:

Signal	**High**	**Too high**	**Duration**
Magnetic circuit temperature	100 °C	105 °C	10s

Stator coil temperature	103 °C	108 °C	10s
Hot air temperature	72 °C	74 °C	10s
Cold air temperature	35 °C	37 °C	10s

(iv) Can anomalies be detected by these alarms? Is your anomaly detection approach able to detect anomalies earlier than these predefined alarms ? In other words, if it allows one to have foresight, which would be very helpful for the plant operators.

Hints:

- You may want to track the evolution of the anomaly score from your anomaly detection model.

C Auxiliary tasks [~30% of the report]

In addition to the main anomaly detection task, each group receives one or two auxiliary tasks. Groups of 3 students work on the first task only, while groups of 4 students work on both tasks.

C.1 Transfer between different units

Until now, you have probably developed separate models for each plant unit. Would it be possible to develop an anomaly detection model on one unit, and to apply it on another one (for example, VG5->VG6, or VG5->VG4) ?

Consider that you have only access to the training data of one "source" unit to develop your AD approach. The goal is to detect anomalies in the test data of the other "target" unit.

In particular:

- Evaluate the results for the synthetic anomalies.
- Analyze whether the anomalies can be detected effectively in the real test data, and compare them with your previous model developed specifically for the target unit.
- Propose potential strategies to improve the transferability of the model. What if you had access to a limited amount of training data for the target unit?

C.2 Impact of seasonality and amount of training data

In this auxiliary task, the goal is to study the impact of the seasonality and amount of available normal training data on your developed models.

Reduce the time span of the training data and analyze the impact on the models you developed for the anomaly detection task. What happens if the training data was collected in different seasons/months compared with the testing data?

5 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.

Milestone Submission (Pitch 7.11)

On the **7th 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 from 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, as well as slides for a 10 min presentation of your findings. The presentation of results is planned for **12. and 16. December 2024** at EPFL in front of Alpiq engineers and the TAs. Please make your submission on the course Moodle **before 12 December**. We ask for all presentations and reports on that date independent of your assigned presentation slot. 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 final presentation (30%)
3. The submitted report (70%)

6 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.

7 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 2024, 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 Alpiq AG, CH-1001 Lausanne

8 Important Key dates

- **10.10 : Project introduction, group forming, data sharing**
- **14.10: Excursion**
- **7.11 : Milestone support slot & Project Pitches**
- **14.11 : Introduction to milestone 2 & Support slot**
- **28.11 Support slot**
- **5.12 Support slot**
- **12.12 : Submission milestone 2 (Presentation & Report) & final presentations (Slot 1)**
- **16.12 : Final presentations (Slot 2)**
- **End of december : grading**
- **01.02.2024 : Alpiq data must be erased**

9 Contacts

- Martin Boden martin.boden@alpiq.com
- Thierry Zufferey thierry.zufferey@hydro.ch
- Olga Fink olga.fink@epfl.ch
- Florent Forest florent.forest@epfl.ch
- Raffael Theiler raffael.theiler@epfl.ch
- Leandro Von Krannichfeldt leandro.vonkrannichfeldt@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

Last Name	First name	Place, date	Signature

Last Name	First name	Place, date	Signature