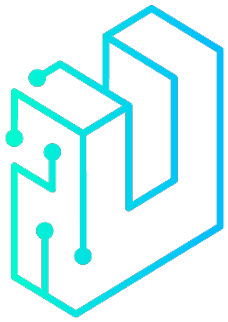

UNDERPIN

DATA SPACE FOR MANUFACTURING

Pan-European data space for holistic asset management in critical manufacturing industries

D4.2 Use case validation and lessons learned – mid-term report

30/11/2024

Co-funded by the European Union

This project has received funding from the Digital Europe Programme under grant agreement No 101123179

Document information			
Project name	Pan-European data space for holistic asset management in critical manufacturing industries		
Project acronym	UNDERPIN		
Grant Agreement No	101123179		
Start / Duration	1/12/2023		
Project Coordinator	MOTOR OIL (HELLAS) DIILISTIRIA KORINTHOU AE		
Deliverable	D4.2 Use case validation and lessons learned – mid-term report		
Work Package	WP4		
Authors	WM		
Dissemination level (PU = Public; PP = Restricted to other program participants; RE = Restricted to a group specified by the consortium; CO = Confidential, only for members of the consortium; SEN= Limited under the conditions of the Grant Agreement)	PU		
Type	Document, Report		
Due date (M)	M12	Actual delivery date	29.11.2024

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

Document history

Version	Date (DD/MM/YYYY)	Author(s)	Comments / Description
V1	22/08/2024	Georgios Chrysokentis	Working document, Table of Contents
V2	30/09/2024	Georgios Chrysokentis, Axel Weißenfeld	Details on implementation of UC2.1, description of each Use Case
V3	29/10/2024	Axel Weißenfeld	Section 3
V4	10/11/2024	Apostolos Garos	Section 2
V5	13/11/2024	Axel Weißenfeld, Apostolos Garos, Georgios Chrysokentis	Finalized all sections, document sent to reviewers
V6	19/11/2024	Robert David	Reviewed by SWC
V7	25/11/2024	Fanis Christakopoulos	Reviewed by MOH
Final	28/11/2024	Axel Weißenfeld, Apostolos Garos, Georgios Chrysokentis	Final version to be submitted

List of Authors and Contributors

Name	Organisation
Georgios Chrysokentis	WM
Christos Papaleonidas	WM
Axel Weißenfeld	AIT
Aristotelis Ntafalias	MOH
Fanis Christakopoulos	MOH
Sofia Visvardi	MORE
Odysseas Kokkinos	INNOV
Victoria Katsarou	SPH
Apostolos Garos	SPH
Robert David	SWC
Amela Kurtić	SWC

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

UNDERPIN has received funding from the Digital European Programme under grant agreement No 101123179.

Table of Contents

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

List of Figures

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

List of Tables

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

Acronyms and Abbreviations

UC	Use case
DS	DataSpace
ML	Machine learning
MAPE	Mean Absolute Percentage Error
WTG	Wind turbine generator
WF	Wind farm
SCADA	Supervisory Control and Data Acquisition
RPCA	Recursive principal component analysis
PDF	Probability density function
RPA	Robotic process automation

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

Executive summary

This document reports on the initial implementation of the use cases (UCs) that form the core of the UNDERPIN project, following their identification and planning, as reported in deliverable D4.1: "*Use case planning report*". The work presented here builds on the results of the aforementioned deliverable, reflecting on the efforts of tasks T4.2: "Use case trials execution" and T4.3: "Trials performance evaluation and lessons learned". The present document forms the initial version of D4.2: "Use case validation and lessons learned", hereby labelled as "mid-term report", serving as the basis for the final report, due in M24.

The primary goal of the document pertains to discussing the initial stage of UC implementation, presenting first results and deducing lessons learned that can inform future decisions and steps to be followed. The UCs that were selected in order to showcase the functionality and benefits of UNDERPIN are presented in the following table:

Identifier	Title	Pilot
UC1.1	Monitoring and predictive maintenance in the refinery	Refinery
UC2.1	Predictive maintenance in wind farms	Wind farms
UC2.2	Wind turbine blade repair prediction	Wind farms

For each UC, a status update is provided including the initial set of results, which offer the basis for a first evaluation of the UC implementation and processes. This is followed by a roadmap where future steps to be followed are delineated in order to streamline the procedure and ensure alignment between the involved parties. The deliverable also reflects on the lessons learned so far from the implementation of the UCs, taking different points of view into account.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

1 Introduction

1.1 Purpose and scope

UNDERPIN is a manufacturing Data Space (DS) that facilitates the efficient use of industrial data with a focus on dynamic asset management and predictive maintenance procedures. Through providing a trusted platform for data sharing and exchange, UNDERPIN will foster the collaboration between large industry players and SMEs in a bid to improve products, services, as well as business operations of the involved parties.

To showcase this potential, a set of use cases (UCs) has been identified, planned and is currently at the implementation stage. Following the work presented in D4.1: "*Use case planning report*", the present document looks to elaborate on the implementation of the three identified UCs, providing a status update on the ongoing work, while also attempting to draw useful conclusions, that can be utilized as lessons learned, not only for improving the implementation of the selected UCs, but also as guidance for other stakeholders joining UNDERPIN, as well as similar efforts in other DS.

This deliverable contains the work performed so far in the context of tasks T4.2: "Use case trials execution" and T4.3: "Trials performance evaluation and lessons learned". This mid-term report comprises a first version of deliverable D4.2: "Use case validation and lessons learned", setting the building blocks for a final version, due in M24 of the project (coinciding with its completion).

1.2 UNDERPIN use cases

After a multi-step process that involved a technical workshop, followed by specialized meetings with subject experts from MOH and MORE, the following three UCs were identified:

- **UC1.1: Monitoring and predictive maintenance in the refinery**: A predictive maintenance algorithm will be developed that will allow for predicting equipment failure as well as detecting abnormal behaviour trends in the refinery's compressors, with the aim of reducing downtime and increasing the lifetime of the machinery.
- **UC2.1: Predictive maintenance in wind farms**: A predictive maintenance algorithm will be developed that will allow for predicting equipment failure as well as detecting abnormal behaviour trends in wind turbines, acting as a benchmark for the maintenance operations carried out by contractors.
- **UC2.2: Wind turbine blade repair prediction**: Statistical analysis of wind turbine blade damages from lightning strikes and prediction of necessary blade repairs based on relevant historical data, in order to understand the impact of lightning strikes on blade damage and avoid catastrophic blade failure through timely repairs.

For a more detailed description of each UC, the reader can refer to D4.1: "Use case planning report", although additional information in a consolidated form is also provided in the respective sections of the present document.

1.3 Structure of the document

The structure of the deliverable is as follows:

- Section 1 presents a brief overview of the UCs and the scope of this deliverable

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

- Section 2 elaborates on the work performed regarding UC1.1: “Monitoring and predictive maintenance in the refinery”.
- Section 3 gives a detailed description on the implementation of UC2.1: “Predictive maintenance in wind farms” to date.
- Section 4 consists of an initial examination and literature review of UC2.2: “Wind turbine blade repair prediction”.
- Section 5 highlights the lessons learned from the inception and initial execution of the UCs with respect to the DS itself, the stakeholders and UC implementation.
- Section 6 summarises the outcomes of this deliverable.

For each UC specific section, an overview of the UC is initially provided, followed by a detailed description of the work that has been performed until now. The sections conclude with a roadmap of next steps that are planned for subsequent work on the implementation and execution of the UCs.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

2 UC1.1: Monitoring and predictive maintenance in the refinery

2.1 Use case description

The goal of this UC is to develop a predictive maintenance model for selected machinery from the refinery. The process pertains to collecting data from five different compressor machine groups along the main refinery process, which is subsequently analysed and processed through specialized machine learning (ML) algorithms with the aim of monitoring equipment performance and predicting impeding failures.

The datasets used for this UC consist of sensor data collected from multiple refinery components, covering operational periods from 2017 to 2020, and additional data from 2022. These datasets are stored in HDF5 files and, for initial exploration, in CSV samples. The data's temporal granularity varies between every five minutes for the years 2017-2020 and every one minute for 2022. This diverse data, collected through sensor networks, offers rich insights into refinery processes. However, the UC faces several challenges, including significant differences in the operational data between the years, which affect model training, as well as data preprocessing complexities such as missing values and timestamp corrections.

For each sensor in the dataset, thresholds were provided. Our approach is to create **Timeseries prediction models** that forecast values **1 day in advance** (as required). In case the forecasted value violates a threshold, we consider this an anomaly of the system, and an alert is raised.

Consolidated information regarding this use case is presented in Table 1. For a more detailed description, the reader can refer to D4.1: “Use case planning report”.

Table 1: UC1.1 consolidated information

Title	Monitoring and predictive maintenance in the refinery
Description	A predictive maintenance algorithm will be developed that will allow for predicting equipment failure as well as detecting abnormal behaviour trends
Use case owner	MOH (maintenance)
Involved partners	MOH, AIT, SPH, INNOV
Assets	5 compressor machine groups from the main process of the refinery
Expected outcomes	Asset owner will be able to appropriately schedule maintenance works for impeding failure, as well as apply corrective actions without interrupting operations based on detected abnormal behaviour
Datasets involved	Sensor data (temperature, pressure, vibration and axial displacement)
Data structures	Format: .xls/.csv Granularity: 5 minutes
Existing infrastructure	- Operations monitored through SAP and SCADA systems - Preexisting predictive maintenance model
Challenges	Insufficient failure data may lead to low accuracy of predictive algorithm

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

2.2 Status update

2.2.1 Initial results

Data provided for this UC consist of two pillars – the operational data of 2022 (sampled at 1 minute) as well as historical data ranging from 2017 to 2020, sampled at 1 minute and 5 minutes, respectively.

Initial exploration of the data revealed substantial differences between operational data from 2022 and earlier data from 2017-2020. To better understand these differences, statistical tests, specifically the Kolmogorov–Smirnov test [1] and Student's t-Test [2], were applied to compare the distributions of sensor readings across different years.

The *Kolmogorov–Smirnov* test was used to determine whether the sensor readings from different years follow the same distribution, focusing on identifying differences in their cumulative distributions.

Student's t-Test, on the other hand, was applied to compare the means of the sensor readings from 2022 with those of previous years, helping to determine whether any significant difference existed between their average values.

Results from these tests confirmed that the operational conditions in 2022 were **significantly different** from those observed in previous years. Therefore, to ensure the models being developed are trained on the most relevant and representative data, the decision was made to use only the 2022 dataset for the predictive maintenance model. This dataset captures the most current state of refinery operations and is more likely to reflect recent changes in equipment or process conditions, ensuring that predictions are aligned with real-world scenarios.

Data Preprocessing:

The preprocessing of this data was an essential and complex undertaking, involving several key steps:

1. **Data Loading:** First, the sensor data was loaded and merged from multiple HDF5 and Excel files, and a consistent datetime format was ensured across the dataset.
2. **Timestamp correction:** In the originally stored operational data of 2022, some individual timestamps were incorrect, as only the date was recorded instead of both the date and time. The preprocessing checks if the timestamps are incomplete and attempts to fill in the missing time by using the timestamp from the next row, provided it's available and valid. Each modified timestamp is reformatted as a string and saved back.
3. **Data Cleaning:** The data was cleaned to handle missing values and duplicates, with any duplicated timestamps being removed to ensure the chronological integrity of the time series. The aggregated yearly values contained duplicate entries, which were also removed to ensure data accuracy and consistency.
4. **Resampling:** The data was resampled to a uniform sampling rate, specifically every 60 seconds, using custom aggregation techniques for different sensors.
5. **Inactive periods removal:** Based on some control sensors, we were able to identify periods during 2022 that each machine was inactive (for unidentified reasons). We removed those periods from our datasets (by replacing the sensor values to NaN). We

added 1 day padding before and after each period, to avoid utilizing skewed values in our model training and evaluation.

6. **Feature engineering:** This step involved the addition of an 'under_maintenance' flag to indicate whether a given data point occurred during a maintenance period. Additionally, the column 'missing_values,' was also added to indicate rows with missing sensor data rather than imputing these missing values, which could introduce bias.

Statistical Analysis:

A tool was developed, which allows time-series datasets to be analysed efficiently. It provides an overview of key patterns and trends, helping data consumers who are potentially interested in the dataset to gain a clearer understanding of its content and significance. By offering insights into the data's structure, the tool aids in identifying important features and areas of interest, facilitating a more informed exploration of the dataset. In Figure 1, the histogram of a specific sensor, which measures the temperature of a component, is exemplarily presented.

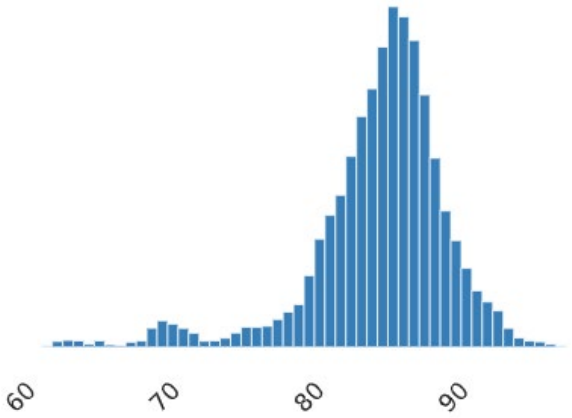

Figure 1: Histogram of the sensor "22TI123".

In addition, in order to assist model development, statistical analysis was performed to better understand the structure and properties of the dataset.

Descriptive statistics were calculated for each sensor, including metrics such as mean, standard deviation, minimum, maximum, and quartiles, with summaries saved for documentation. Sensors common to both datasets were identified, and percentage differences in statistical properties, like mean and standard deviation, were calculated to compare data from different years. Visualizations such as histograms and time series plots were also created to observe trends and highlight key differences, with operational thresholds overlaid to facilitate anomaly identification.

In Figure 2, we can see the linechart of a sensor, along the thresholds that were provided.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

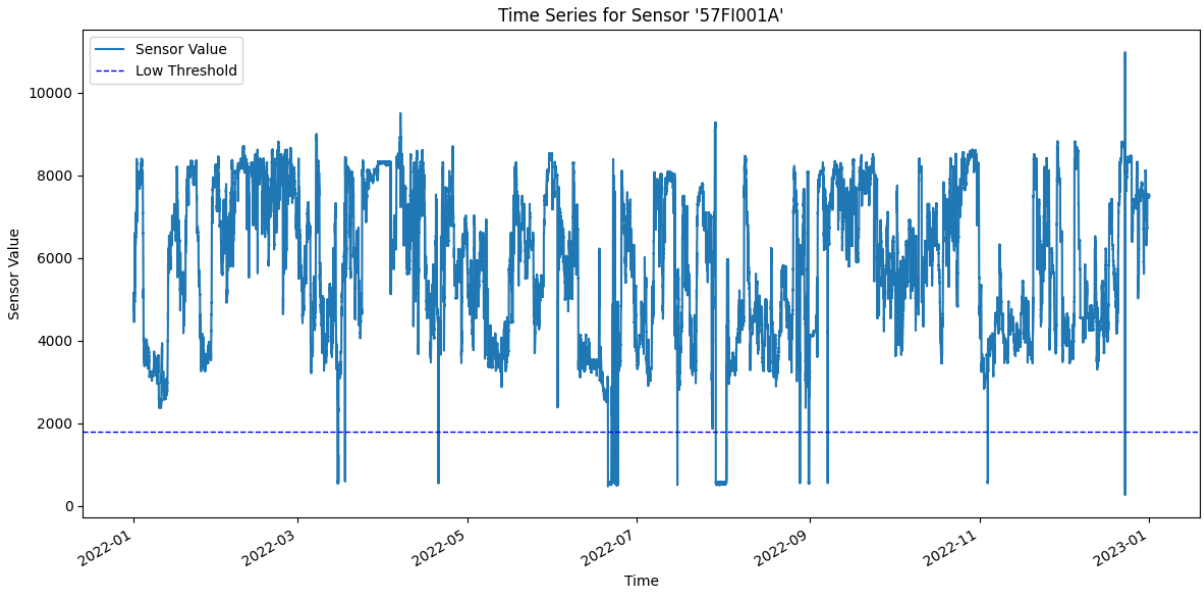

Figure 2: Linechart of the sensor "5FI001A".

The primary objective of model development is to predict sensor readings and detect anomalies that could indicate potential equipment failures. We are planning to test both **univariate and multivariate** modeling approaches, including **statistical, machine learning (ML), and deep learning** methods, to identify the best fit for predictive maintenance in refinery settings.

So far, only univariate statistical and ML models have been tested. The models tested so far include both traditional statistical methods and more advanced ML techniques. For statistical methods, **Exponential Smoothing** [3] was tested with a 10-day history window, and **ARIMA** [4] was applied using a 20-day history window. These models provided a baseline understanding of the data's temporal dynamics.

In addition, ML methods such as **Random Forest Regressor** [5] and **Linear Regression** [6] were implemented, both using a 10-day history window in a univariate setting. This means that we trained a model for **each sensor** and created a wrapper function that calls all of them to make predictions for the entire sensor pool.

To keep the results consistent and comparable, an 80-20 split was done, using approximately 10 months for training and the last 2 months (November and December) for testing. The data was also resampled to an hourly frequency.

No hyperparameter tuning was done for any of the models; instead, we aimed to observe their performance out of the box in order to decide which ones to fine-tune later. The statistical models were evaluated using **rolling predictions**, while the ML models were tested using **one-shot predictions**—predicting one value 24 hours ahead, without using the previous 24-hour predictions for subsequent time points.

One difficulty we faced was the evaluation (and comparison) of the models. In every experiment, we are building separate models for each sensor, but we need to somehow evaluate aggregational metrics over all of the sensors. Since the sensors are different in nature, their

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

scales significantly differ, so we cannot use absolute value metrics. We decided to move forward with Mean Absolute Percentage Error (MAPE), which we calculate over each sensor and then aggregate.

Our goal was to utilize the mean MAPE over all sensors to evaluate the performance of a model (across all sensors). However, we are getting very high values of mean MAPE. We researched this issue by checking the median value of MAPE and some indicative percentile values. We determined that even though the models are performing at an acceptable level in the vast majority of the sensors, there are a few problematic cases with very large values of MAPE (even 50000% in some cases).

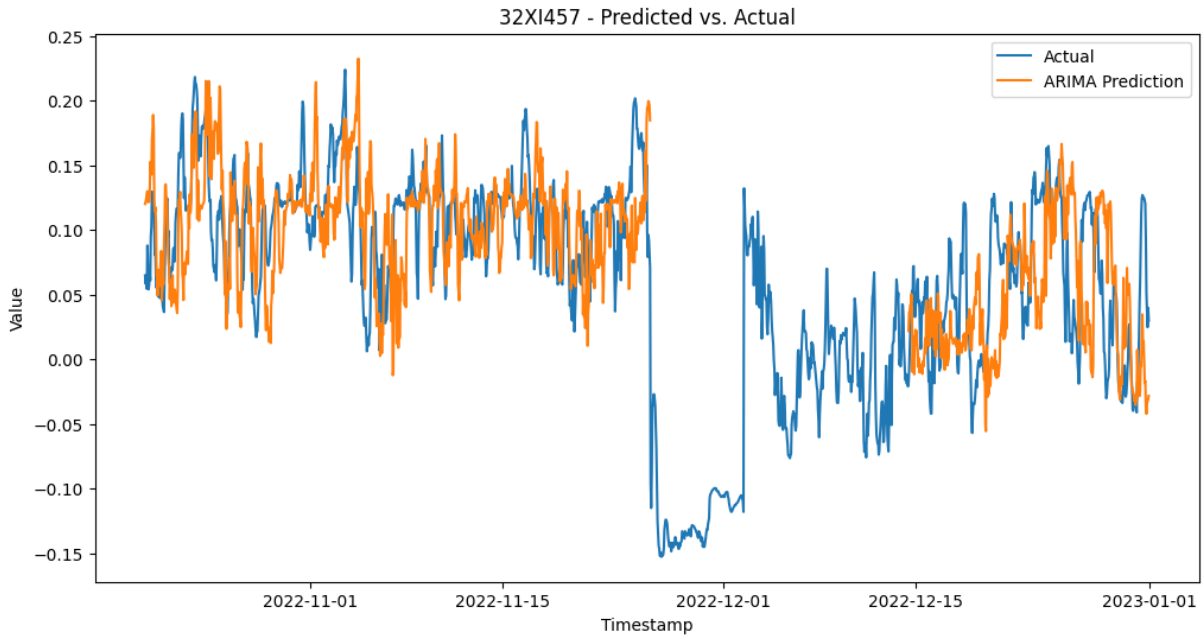

Figure 3: Predicted vs Actual values of the sensor "32XI457"

Upon visual inspection of the problematic sensors, we confirmed that the models generally perform well, so they are not responsible for the large values of MAPE. However, the issue of extreme values arose due to a "division by zero" error, a common challenge when using the MAPE evaluation metric. As illustrated in Figure 3, numerous values in the dataset are either very close to or exactly zero.

Initially, we considered filtering out values near zero to mitigate this issue, but this approach was ultimately deemed scientifically unsound, as it risked discarding non-problematic data points.

Instead, we focused on exploring alternative evaluation metrics that could avoid the "division by zero" problem. After thorough research, we identified two suitable MAPE variations:

- **Median Absolute Percentage Error (MdAPE)**: Unlike MAPE, which uses the mean, MdAPE uses the median of the absolute percentage errors. This makes it less sensitive to outliers and the division by zero problem, providing a more robust central tendency.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

- **Weighted Mean Absolute Percentage Error (WMAPE)**: WMAPE adjusts the MAPE formula by weighting errors based on the magnitude of the actual values, reducing the impact of small or zero values and allowing a more balanced assessment across different scales of data. A clear advantage of this metric is that the mathematical formula that calculates it eliminates the "division by 0" problem.

We tested both MdAPE and WMAPE and compared them to the previously calculated MAPE values for sensors with both reasonable and extreme MAPE scores. The comparison revealed that, while both metrics effectively addressed the division by zero issue, WMAPE proved to be more reliable. For sensors with typical MAPE values, WMAPE produced results close to MAPE, maintaining consistency. For sensors with anomalous MAPE scores, WMAPE provided reasonable, stable results.

MdAPE, although effective for problematic sensors, also altered values we deemed reasonable, primarily due to the inherent differences between mean (used in MAPE) and median (used in MdAPE). Consequently, we chose **WMAPE as our preferred metric for evaluations.**

The results for the models we have tested are in Table 2.

Table 2: Mean Absolute Percentage Errors (MAPE) for each model

Model	**Mean WMAPE (%)**
Exponential Smoothing	9.2
Arima	6.6
Random Forest Regression	19
Linear Regression	**5.7**

As we can see, **Linear Regression** is the model with a slight **advantage over the others**. In order to get a deeper understanding on the model's performance across different metrics, we have visualised the predictions along with the actual values on the test set (Figures 3-4).

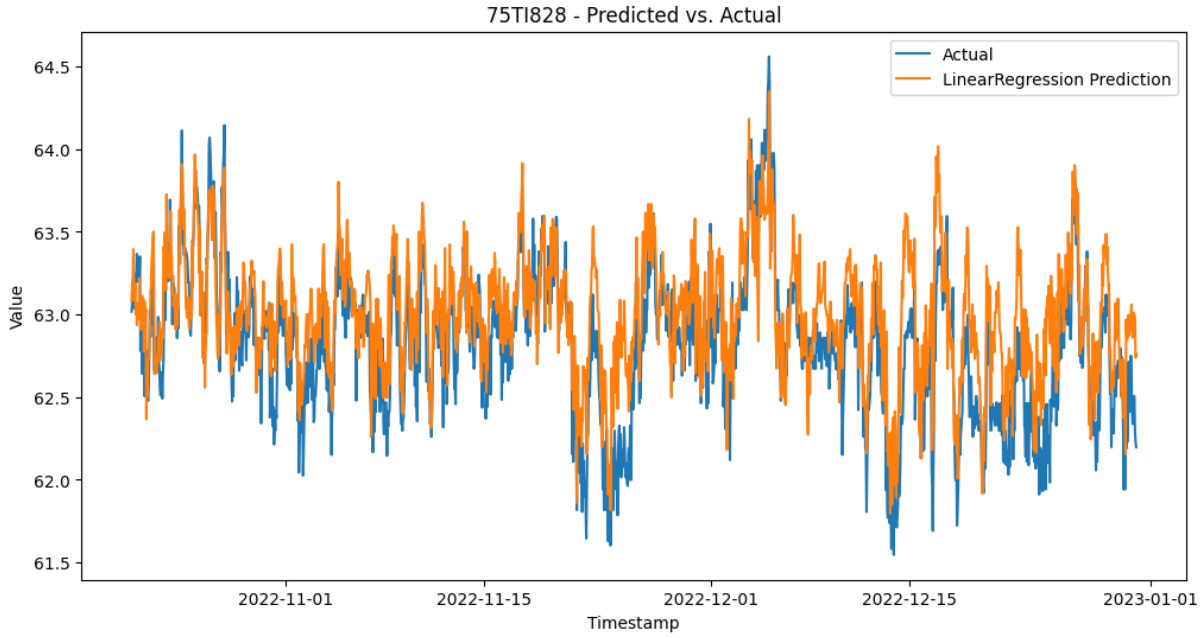

Figure 4: Linear Regression predictions and Actual values for the sensor "75TI828".

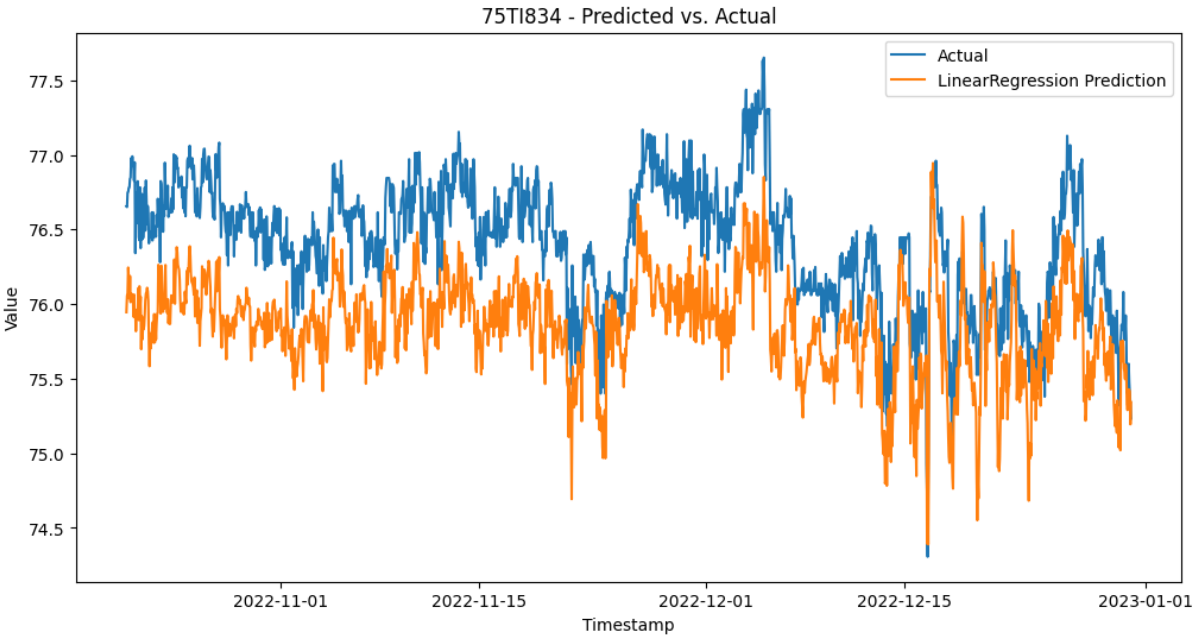

Figure 5: Linear Regression predictions and Actual values for the sensor "75TI834".

The requirement set out by the refinery for this metric is **MAPE < 3%**. The best model right now is achieving a mean WMAPE of 5.7%, which is not sufficient. However, keeping in mind that we will be experimenting with several other models, as well as hyperparameter tuning the best ones, we are very optimistic that the goal will be reached in the later stages of our implementation.

2.2.2 Infrastructure

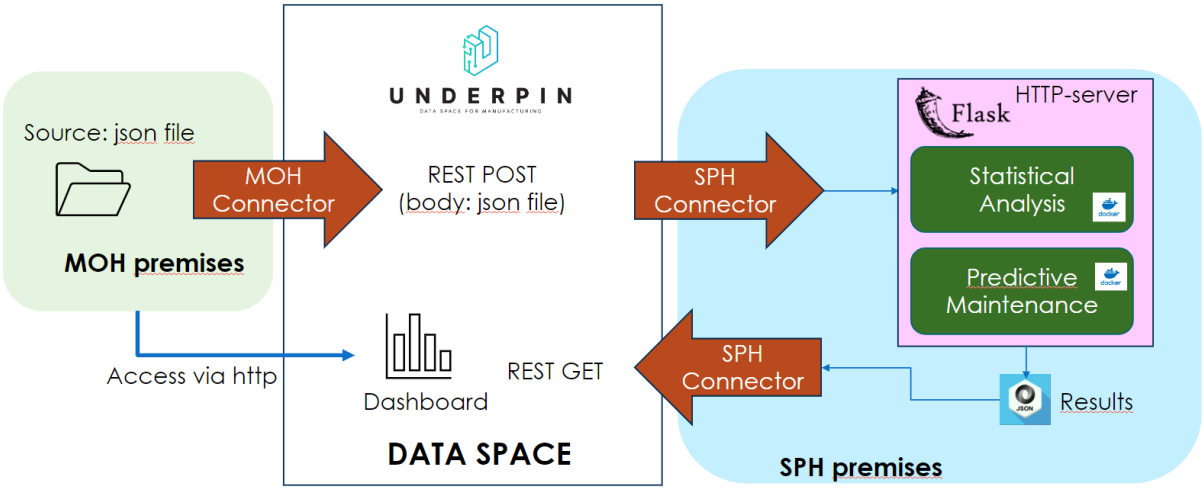

Figure 6: Dataflow of refinery demonstration

As we progress in developing the Predictive Maintenance models, it is essential to remain cognizant of the operational context and the target deployment environment. Consequently, we have integrated our current models into the data processing pipeline illustrated in Figure 6.

In this scenario, **MOH** serves as the **data provider**, supplying refinery sensor readings and anticipating statistical analyses as well as the outcomes of the Predictive Maintenance model. These results will be visualized on the UNDERPIN dashboard, which operates within the DS.

SPH functions as the data consumer, receiving the data transmitted by MOH, conducting the required analysis and model inference, and generating the results.

The whole process described above will follow these steps:

Data analysis & PrM workflow:

1. **MOH** negotiates a **contract** with **SPH** in order to send them refinery sensor readings (a contract is established between the **MOH** and **SPH connectors**)
2. **MOH** sends a **REST POST** request to **SPH** (forwarded through the DS), containing the data they want processed
3. **SPH** receives the request and analyses the data
4. The results are stored in **SPH's** premises
5. A "Success" message is returned to MOH (not results)

Result visualisation:

1. **MOH** connects to the **Dashboard** (operating on the DS infrastructure) using their credentials (not a connector)
2. The **Dashboard** makes a **REST GET** request to SPH through the connector to receive some results
3. **SPH** receives the request and responds with the results that were asked
4. The **Dashboard** visualizes the results

2.3 Roadmap for next steps

We are currently at a stage in model development where the training and evaluation pipeline is established, and data preparation is at a good point. While these components are not yet finalized, they are stable, and we do not anticipate making significant changes to them in the near future. Our primary focus now will be to explore a broader range of models and approaches, identify the most promising candidates, and fine-tune them to achieve optimal performance.

The roadmap for the next steps to commence within Q1 2025 is as follows:

1. Continue experimentation with Univariate Prediction models
2. Conduct experiments with Multivariate Prediction models
3. Evaluate and compare model performance
4. Select the best models for further tuning
5. Ensure the selected model is containerized (e.g., via Docker) for integration into the deployment pipeline

By following these steps, we aim to develop a robust Predictive Maintenance model that is not only optimized for performance but also prepared for deployment and scalability within a real-world system.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

3 UC2.1: Predictive maintenance in wind farms

3.1 Use case description

The goal in this UC is the timely prediction of failures in major components of wind turbine generators (WTG). Similarly to UC1.1, this procedure is performed through a predictive maintenance model, making use of specialized ML algorithms. In this case, the operator is not directly responsible for performing maintenance on the wind turbines. Instead, maintenance is performed by a third party through a relevant long-time service agreement. Therefore, the expected outcome of this UC is for the wind farm (WF) operator to be able to monitor potential abnormalities in the operation of the WF, while also benchmarking the maintenance works performed by the contractor. A summary of important information regarding UC2.1 is provided in Table 3, and the reader is referred to D4.1: "Use case planning report" for a more detailed description of the UC.

Table 3: UC2.1 consolidated information

Title	Predictive maintenance in wind farms
Description	A predictive maintenance algorithm will be developed that will allow for predicting equipment failure as well as detecting abnormal behaviour trends
Use case owner	MORE (operations)
Involved partners	MORE, AIT, SPH, INNOV
Assets	Wind turbine generators from MORE's wind farm portfolio
Expected outcomes	Asset owner will be able to optimize operations based on detected abnormalities, as well as benchmark the maintenance works carried out by contractor
Datasets involved	Sensor data from multiple components within the wind turbine Fault alarms and warnings
Data structures	Format: .xls/.csv Granularity: Every 10 mins
Existing infrastructure	Operations monitored through proprietary SCADA systems offered by wind turbine manufacturers
Challenges	Loss of communication with wind turbines means alarms and errors may not always be detected

An overview of the failures of major components shows that the generator (with over 72%) is the component that most frequently fails and needs to be replaced. Therefore, we have decided to focus on this component first.

3.2 Status update

3.2.1 Initial results

The onshore WF under consideration is located on a Greek island and has been operational since 2009. It consists of ten WTGs. The WF's data used in this project is communicated by the Object Linking and Embedding (OLE) for Process Control (OPC) and stored in the Open Database Connectivity (ODBC) server. The initial data provided was collected from the Supervisory Control and Data Acquisition (SCADA) systems of the WTGs over a 12-month period (01/01/2019-

31/12/2019) using a variety of sensors configured to measure key operational variables, such as wind speed, pitch angle, temperatures etc. at 10-minute raw intervals.

Data Preprocessing:

In the originally stored production data, some individual timestamps were incorrect, as only the date was recorded instead of both the date and time. This resulted in incomplete or inaccurate timestamp entries for certain data points. To correct these false timestamps, we leveraged the information from neighbouring timestamps and the known time interval of 10 minutes between consecutive entries. By using the correct neighbouring timestamps as reference points, we were able to reconstruct the missing time values and ensure that the corrected timestamps adhered to the expected 10-minute interval. There are different methods to fill missing values in cells, such as forward fill, backward fill, interpolation, or using a constant value. We decided, however, not to fill any missing values but instead added the column "missing data" which indicates if a timestamp is available but any sensor data missing.

Statistical Analysis:

A tool was developed, which allows time-series datasets to be analysed efficiently. It provides an overview of key patterns and trends, helping data consumers who are potentially interested in the dataset to gain a clearer understanding of its content and significance. By offering insights into the data's structure, the tool aids in identifying important features and areas of interest, facilitating a more informed exploration of the dataset. In Figure 7, the histogram of a specific sensor is exemplarily presented.

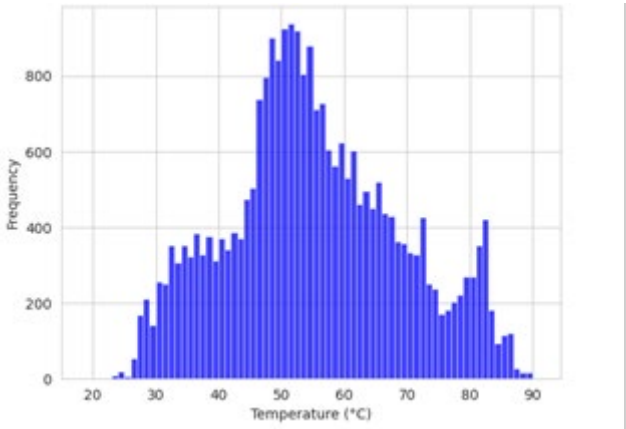

Figure 7: Histogram of the Generator Bearing Temperature of WTG01

Model design:

A literature review indicates that the generator bearing temperature is commonly used to predict the health of the generator [7]. Consequently, the target variable we aim to model is the generator bearing temperature – our specific WTG has two bearings. Since we are conducting post hoc monitoring, we can leverage measurements of any unidirectional causal signals at time T to model the normal operating range at the same time T. The input features we will use include generator rotational speed, nacelle temperature, and generated active power.

The following features are selected – all values are measured over a ten-minute interval:

- Generator Bearing Temp. Avg. [°C]: Average temperature of the generator bearing.
- Generator Bearing2 Temp. Avg. [°C]: Average temperature of the generator bearing 2.

This project has received funding from the Digital Europe Programme under grant agreement No 101123179

- Generator RPM Max. [RPM]: Maximum rotations per minute.
- Nacelle Temp. Avg. [°C]: Average temperature of the nacelle.
- Production Latest Average Active Power Gen 0/1 Avg. [W]: Average Power Production by generator 0/1.
- Generator Cooling Water Temp. Avg. [°C]: Average temperature of the water circulating through the generator's cooling system.
- Rotation on/off: The generated feature tracks the number of time steps since the last instance when Generator RPM Max crossed a specific threshold (e.g. we use 1000 RPM). A positive value is recorded if it crossed from below the threshold to above, and a negative value if it crossed from above to below. It serves as a proxy for heat build-up and stagnation over time, capturing the frequency and duration of operational changes in the generator's RPM.
- Missing data: Generated feature - whether data is missing in the dataset.

All the selected features demonstrate positive correlations with the generator bearing temperature, meaning that as these features increase, so does the temperature of the generator bearing (Figure 8). Among these, nacelle temperature—which varies due to factors like daily temperature cycles and seasonal changes—and power generation show the strongest positive correlations, indicating that they have the most significant influence on the bearing temperature.

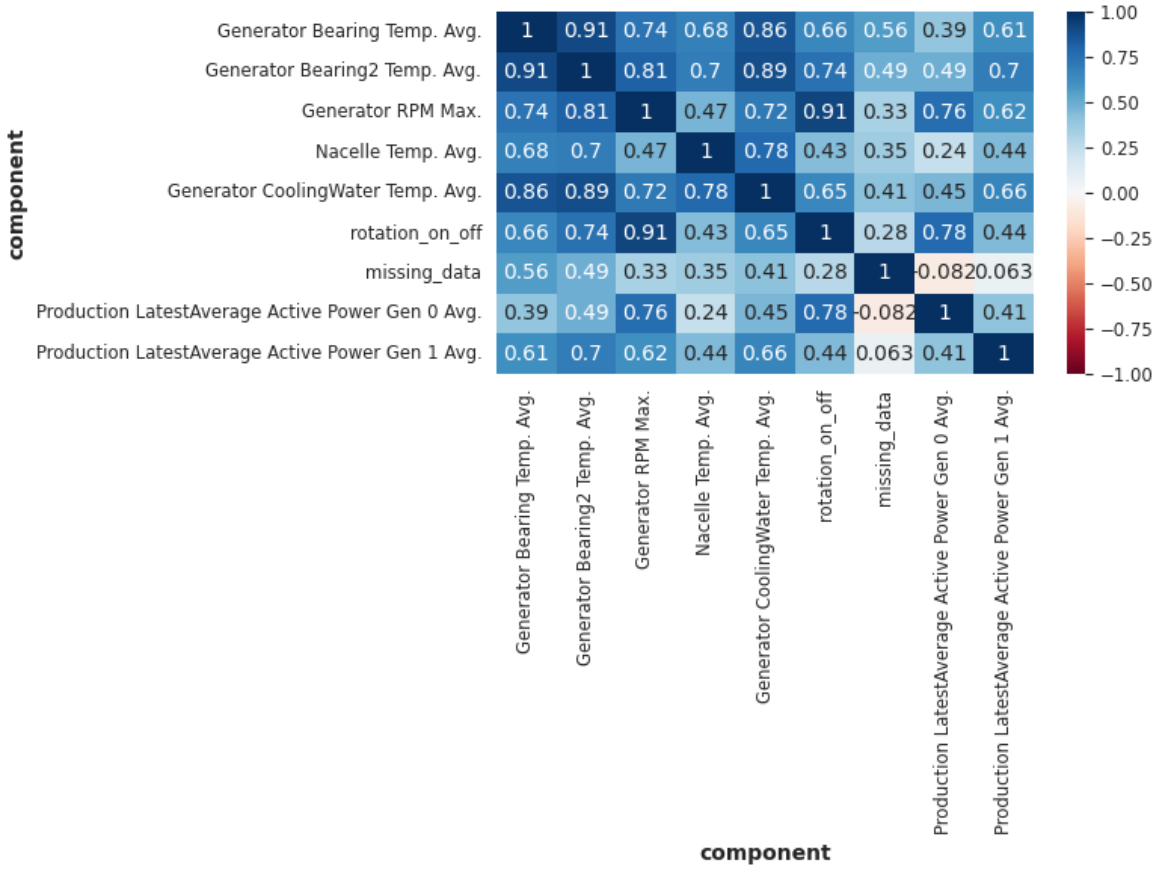

Figure 8: Visualizing the strength of relationships between features and generator bearing temperature, highlighting key positive correlations with features like nacelle temperature and power generation.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

In cases where some features exhibit high variability or noise due to frequent fluctuations (e.g., rapid changes in environmental conditions or operational parameters), this noise can obscure underlying patterns in the data. To address this, we can apply a moving average filter to smooth out short-term variations and highlight longer-term trends. By reducing the impact of noise, this filtering technique can potentially increase the observed correlation between the features and generator bearing temperature, leading to clearer insights.

The boxplot in Figure 9 reveals the difference in the generator bearing temperatures between both bearings. This variation suggests that the temperature distribution and behaviour differ substantially across the bearings. As a result, it indicates that a separate model needs to be trained for each bearing to accurately capture the unique characteristics and operating conditions of their respective generator bearing temperatures. Training individual models ensures that the predictions are tailored to the specific behaviour of bearing, leading to more reliable and precise monitoring.

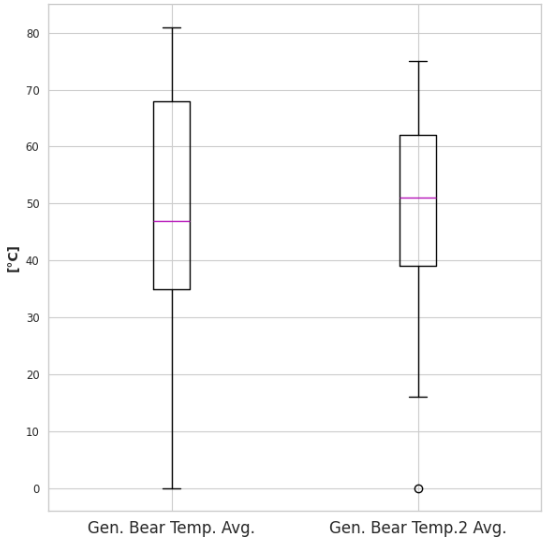

Figure 9: Boxplot comparing the generator bearing temperatures of two wind turbines, highlighting the significant differences in temperature distribution between both bearings.

Since our predictive maintenance model is applied post hoc, we can model the generator temperature at any given time using actual sensor data from that moment. To accomplish this, we configure the model to predict one time step at a time. We use a probabilistic model to forecast generator bearing temperature quantiles, serving as a proxy for the normal temperature range, with the likelihood set to "quantile" and quantiles defined as [0.05, 0.50, 0.95]. To ensure causal flow, past temperature data is not used as input. Instead, we rely solely on the previous introduced features, as data from other sensors is available for temperature prediction at a specific time step. Additionally, we include calendar features, such as the month and hour of the day, to capture seasonal and diurnal patterns (e.g., temperature variations influenced by the time of year or day/night cycles).

We are training multiple models, including the CatBoost model [8], the XGBoost model [9] and theLightGBM model [10], which support probabilistic forecasts. CatBoost (Categorical Boosting) is a gradient boosting algorithm designed to handle categorical features efficiently without requiring extensive preprocessing, making it highly suited for datasets with mixed feature types.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

Its ability to automatically encode categorical variables and mitigate overfitting through advanced regularization techniques makes it particularly robust in complex tasks. On the other hand, XGBoost is another powerful gradient boosting algorithm known for its efficiency and scalability. XGBoost excels in handling large datasets and provides flexible customization options, such as tree-based models and regularization parameters, allowing us to fine-tune the model for higher accuracy and better performance. XGBoost is particularly known for its speed and precision in classification and regression tasks. LightGBM (Light Gradient Boosting Machine) is a fast, high-performance gradient boosting framework, which excels in both regression and classification tasks, especially with structured data, and includes methods for handling missing values and imbalanced datasets. All models are based on the implementation of Darts [11].

For training and evaluation, the data was divided into two segments: the training period (01-2019 to 09-2019) and the evaluation period (10-2019 to 12-2019). During this time, only a single major generator failure occurred (November 25th, 2019, to December 18th, 2019), limiting the robustness of any thorough evaluation. Consequently, this report focuses primarily on demonstrating the principal workflow and verifying the feasibility of the approach. Fine-tuning and enhancements will be conducted in a follow-up phase, utilizing additional data from 2020 onward to improve model reliability and evaluation depth. The XGBoost models show good performance and are used in a first anomaly detection.

Both trained models are used to generate a probabilistic forecast of the generator bearing temperature, relying solely on the previously defined covariates. The forecasted bearing temperature values have been replaced with *NAN* wherever the 'missing data' column is true, as the forecast cannot be considered reliable when portions of the covariate data are missing.

Figure 10 and Figure 11 show the ground truth bearing temperature of both generators respectively as a black line, accompanied by the forecasted median temperature (quantile 50) represented by a dark blue line. The forecast uncertainty range, spanning quantiles 0.05 to 0.95, is illustrated by a light blue shaded area. Residuals, which may indicate anomalies, are marked in violet. In both figures, residuals are displayed. As a next step, we plan to implement a sliding window over the residuals - if enough values within the window exceed a predefined threshold, the corresponding time point will be flagged as an anomaly. This feature is yet to be implemented, but we aim to incorporate additional data, particularly more instances of generator bearing failures, to refine the sliding window design.

The actual component failure of the generator occurred on November 25th, 2019. During the displayed timeframe, two significant residuals occur on November 12th and November 24th, potentially indicating anomalies. This suggests that the model based on the bearing temperature can predict failures. However, the residual on November 12th may represent a false alarm. In contrast, the model using bearing temperature 2 appears unable to predict the component failure. Instead, it generates several false alarms on November 11th, 15th, and 20th. These findings require further evaluation, incorporating additional data to draw more definitive conclusions.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

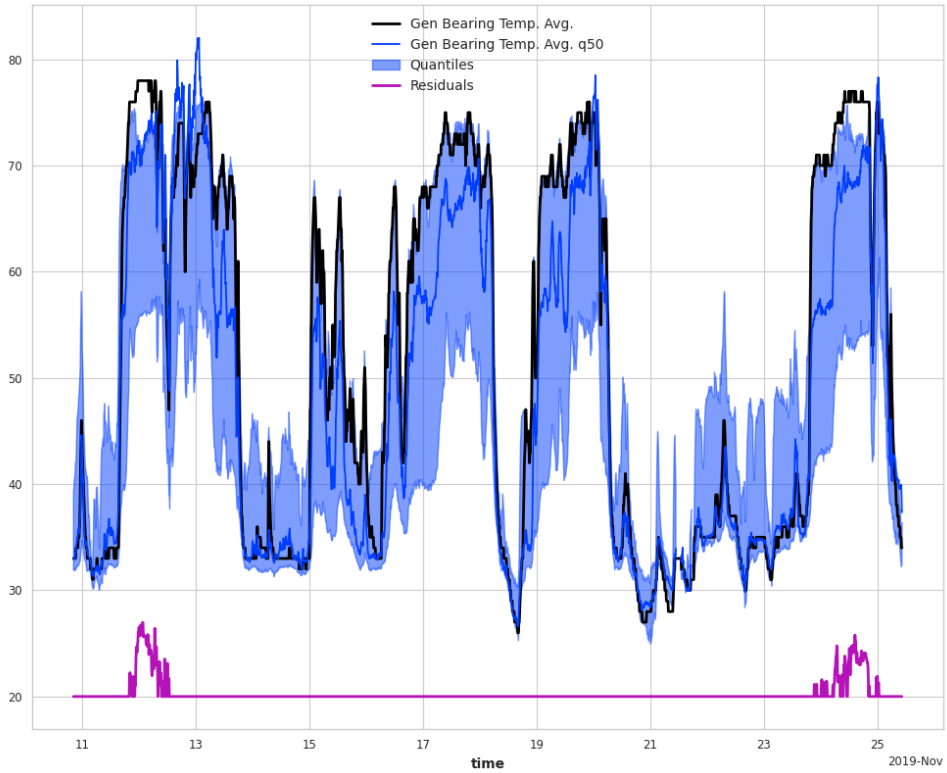

Figure 10: The graph depicts the ground truth bearing temperature (black line) alongside the forecasted median temperature (quantile 50, dark blue line) and the forecast uncertainty range (quantiles 0.05 to 0.95, light blue shaded area). Detected residuals, potentially indicating anomalies, are highlighted in violet.

An additional step involves analysing the changes in generator bearing temperature of a WTG over time. There may be underlying trends, indicating that the models require periodic updates to maintain accuracy. Furthermore, variations in generator bearings across different WTGs could pose challenges, potentially necessitating the training of separate models for each WTG. This would complicate generalization, or in some cases, make it unfeasible.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

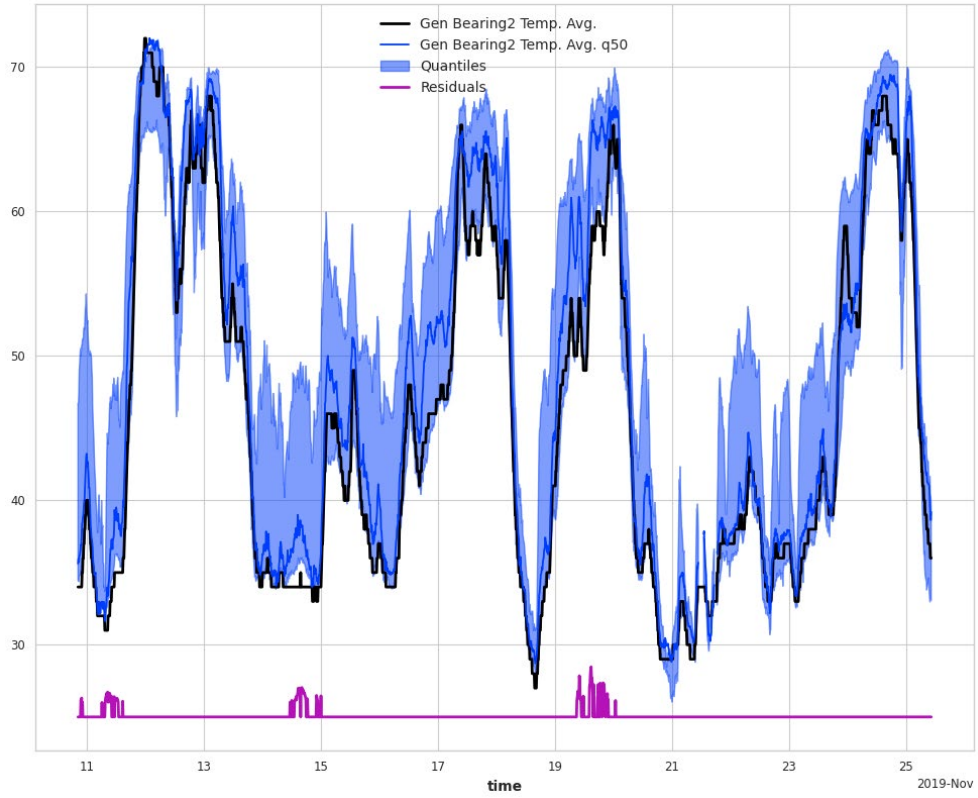

Figure 11: The graph depicts the ground truth generator bearing 2 temperature (black line) alongside the forecasted median temperature (quantile 50, dark blue line) and the forecast uncertainty range (quantiles 0.05 to 0.95, light blue shaded area). Detected residuals, potentially indicating anomalies, are highlighted in violet.

3.2.2 Infrastructure

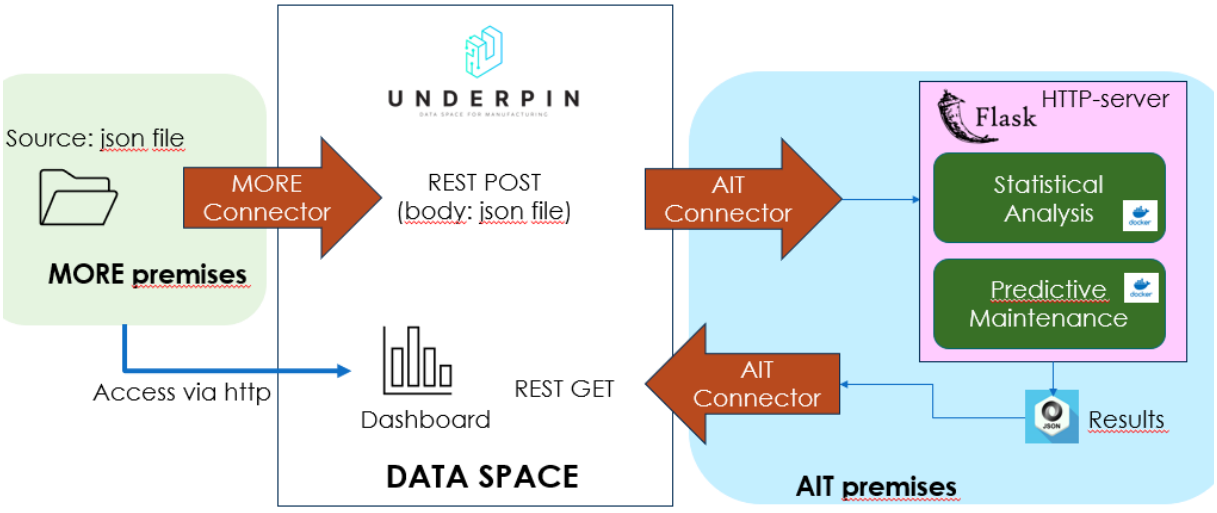

Figure 12: This diagram outlines the project's first iteration, highlighting how the DS enables advanced analytics like predictive maintenance as a service.

A first demonstrator of a data processing pipeline using the designed DS for predictive maintenance is presented in Figure 12. In this specific scenario, the participant MORE would like to analyse its data and displays the results in the dashboard. The participant AIT offers a service for doing a statistical analysis of time-series data as well as carrying out predictive maintenance

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

based on SCADA data. The actual data exchange between both participants is carried out by their connectors. Before any data exchange takes place, the involved participants need to agree on a contract which is stored in the connectors.

In this scenario, MORE serves as the data provider, supplying SCADA readings and anticipating statistical analyses as well as the outcomes of the predictive maintenance model. These results will be visualized on the UNDERPIN dashboard, which operates within the DS. AIT functions as the data consumer, receiving the data transmitted by MORE, conducting the required analysis and model inference, and generating the results.

The whole process described above will follow these steps:

Data analysis & PrM workflow:

1. MORE negotiates a contract with AIT to send their SCADA data. A contract is established between the connectors of MORE and AIT.
2. MORE sends a REST POST request to AIT using its connector, where the request's JSON-formatted body contains the data needed for processing.
3. AIT receives the request and analyses the data – data analysis or predictive maintenance.
4. The results are stored in *json* format on AIT's premises.
5. MORE is notified that the analysis was carried out.

Result visualisation:

6. MORE connects to the Dashboard via an HTTP request using their credentials (not a connector).
7. The Dashboard makes a REST GET request to AIT through the connector to receive the results.
8. AIT receives the request and responds with the results that were requested.

The Dashboard visualizes the results.

3.3 Roadmap for next steps

Up to this point, the basic model, based on bearing temperature, shows significant promise and will remain unchanged. In 2019, the generator of a single wind turbine (WTG5) experienced a failure and required replacement, providing only one instance of failure data for model training and validation. This limited dataset restricts the model's ability to generalize to multiple failure events, which may impact the robustness and accuracy of any anomaly detection models derived from it.

However, we now have access to an extended dataset spanning from 2020 to 2024. During this period, the generators of the wind farm failed on seven additional occasions. These additional instances of generator failure provide a richer dataset that captures a broader range of operational conditions and potential failure signatures. This allows the optimization of hyperparameters for training the gradient-boosted decision tree (GBDT) models, and the development of the sliding window mechanism. A key challenge in designing the sliding window lies in defining pre-set thresholds and parameters to ensure it operates both accurately and robustly. Since the data is available, the optimization of the GBDT models and design of the sliding window is expected to be finalized by Q1 2025. Moreover, the additional data enables a meaningful evaluation, which is not feasible with only a single anomaly.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

The first iteration of our system (based on the initial demonstrator as depicted in Figure 12) will involve deploying predictive maintenance and statistical analysis components to the DS. This will allow all participants in the DS to take advantage of these tools. In addition, we plan on deploying the dashboard to provide users with a streamlined way to visualize their results.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

4 UC2.2: Wind turbine blade repair prediction

4.1 Use case description

The final UC pertains to the development of a statistical and predictive model that analyses wind turbine blade damage from lightning strikes, traces the type and intensity of the lightning strikes and offers predictions for necessary blade repairs. In that sense, this UC acts as a subset of UC2.1, albeit one with a very specialized subject. As a result, UC2.2 follows a delayed timeline in order to take full advantage of the processes and knowledge gained from UC2.1, thus avoiding duplicate work. To that end, no results have been produced for this UC as of yet, and instead the outcomes of a literature review on the subject are presented. The expected outcome of the UC is that the wind farm operator is able to detect potential blade damage in a timely manner, aiming to prevent impeding blade failures by carrying out the required blade repairs. While lightning strikes will be the focus, as they are the main contributor to blade damage, the impact of strong winds will also be considered.

Consolidated information regarding this use case is presented in Table 4. For a more detailed description, the reader can refer to D4.1: "Use case planning report".

Table 4: UC2.2 consolidated information

Title	Wind turbine blade repair prediction
Description	Statistical analysis of wind turbine blade damages from lightning strikes and prediction of necessary blade repairs based on relevant historical data
Use case owner	MORE (operations)
Involved partners	MORE, AIT, SPH
Assets	Wind farms from MORE's portfolio with focus on two wind farms equipped with lightning strike monitoring equipment
Expected outcomes	Asset owner will: - gain deeper understanding on the impact of lightning strikes on wind turbine blades - be able to prevent impending blade failures by carrying out blade repairs in a timely manner
Datasets involved	Sensor data from multiple components within the wind turbine Fault alarms and warnings Historical data for peak current, time of lightning strike, strike intensity Blade repair historical data
Data structures	Format: .xls/.csv /.txt Granularity: Every 10mins
Existing infrastructure	Data collected through proprietary SCADA systems offered by wind turbine manufacturers Additional specialized data collected through Lightning Key Data (LKD) systems currently installed in two wind farms
Challenges	Low availability of lightning data since only two wind farms have LKD systems installed

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

4.2 Status update

We conducted a comprehensive literature review to identify state-of-the-art (SOTA) fault detection methods for wind turbines, limiting our focus to those that rely on SCADA data since it is the source of operational data, which MORE provides.

Yang and Zhang [12] proposed a novel conditional convolutional autoencoder (CCAE) for monitoring wind turbine blade breakages, which are trained on SCADA data. The results showed that the CCAE-based monitoring method achieved good performance in terms of the monitoring effectiveness and robustness with an accuracy of over 90%. Rezamand et al.in [13] developed a real-time hybrid fault detection strategy for wind turbine blades based on SCADA data. This approach combines Generalized Regression Neural Network Ensemble for Single Imputation (GRNN-ESI), Recursive Principal Component Analysis (RPCA), and Wavelet-based Probability Density Function (PDF) to accurately detect incipient blade failures. Experiments using SCADA data from a wind farm in southwestern Ontario showed that the wavelet-based PDF with RPCA method could enhance the reliability of fault detection by improving accuracy and reducing false alarm rates. It also demonstrated better early detection of blade faults compared to alternative approaches like wavelet-based PDF with Principal Component Analysis (PCA), wavelet-based PDF with Dynamic Principal Component Analysis (DPCA), and Support Vector Machine (SVM) techniques. Chandrasekhar et al. [14] considered the uncertain nature of operational wind turbine blades' environments, where they proposed a new diagnostic methodology based on novel structural health monitoring (SHM). Gaussian Processes (GPs) were used to predict one blade's edge frequencies using another blade's edge frequencies and ambient temperature as inputs. The system successfully identified damage onset months before it was remedied. Significant losses in wind power occur when turbine blades are damaged beyond repair, making early detection crucial. However, diagnosing early damage is challenging due to operational constraints and complex diagnostic models. The study of Tang et al. [15] used noise signals from turbine operation and the k-nearest neighbour (k-NN) algorithm with generalized fractal dimensions (GFDs) as diagnostic features, achieving 98.9% accuracy. The k-NN algorithm is simple to implement, and an optimal combination of three parameters—GFD scale index, neighbour count, and range formula—was identified, providing a quick, efficient, and accurate diagnostic method.

4.3 Roadmap for next steps

This use-case builds upon our experience from UC2.1 and leverages the curated data and existing literature to quickly develop appropriate models for blade failure detection. The plan includes deploying these models in the DS, offering predictive maintenance as a service using established interfaces. This approach allows us to focus on model training and take advantage of our prior work in UC1.1 and UC2.1. By utilizing our knowledge base from the previous use-case, we can accelerate the development process and ensure seamless integration with current systems. This strategy enables rapid advancement towards our goal of delivering actionable insights for enhanced blade maintenance.

Leveraging the outcomes of UC2.1, as presented in Section 3.2, it is planned that the implementation of UC2.2 will commence in Q1 2025, with the first results expected in early Q2 2025.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

5 Lessons learned

5.1 Lessons learned for the Data Space

Semantic Layer

Data sharing with clear semantic descriptions is crucial for an effective implementation of predictive maintenance ML models. Each manufacturer may use different sensors, measurements and data formats, which makes it difficult to harmonize and integrate the data without formal definitions of the meaning.

By employing standardized ontologies, stakeholders can efficiently and in a unified way access relevant data, streamlining the integration process and facilitating predictive maintenance training data on a large scale, thereby potentially increasing the quality of the predictions.

The Vocabulary Hub is defined in the IDS-RAM as a basic building block to achieve semantic interoperability in DSs by provisioning common vocabularies. Vocabularies are expressed in Resource Description Frameworks (RDF) and use RDF Schema (RDFS) and Web Ontology Language (OWL) for ontologies, and Simple Knowledge Organization System (SKOS) for thesauri. The IDS-RAM allows for extending the function of the Vocabulary Hub towards providing ontology mappings that enable DS connectors to automatically convert data between user-specific and standard data formats. Additionally, these standardizations create a shared language across different industry players, reducing misunderstandings and ensuring each participant interprets data in the same way. Such automatic conversions can be performed by Data Transformation Apps by the Data Consumer, the Data Provider, or both. This interoperability is key for achieving cross-functional insights, enabling more effective diagnostics and maintenance recommendations.

We see the need for a Semantic Layer for DSs to provide such vocabulary-based services for advanced and automated semantic description of metadata and data.

Architectural Integration

The focus of the UNDERPIN UCs is to provide predictive maintenance services in a DS based on the consolidated data of multiple DS participants. The idea is to achieve higher quality of predictive services with higher amounts of training data for the ML algorithms. There is the need to eventually consolidate all the training data available in the DS into one storage to run the training process or to use federated learning approaches which can be especially useful when dealing with sensitive or private information. For UNDERPIN, and because of the harmonization needs via a Semantic Layer, we decided to store the integrated data into a time series database, which makes it easy to assemble specific training data for the predictive maintenance ML. This database selection enables fast querying and retrieval of historical data, which is crucial for developing accurate and responsive predictive models. However, the IDS-RAM is based on principles of decentralization for data sharing, which does not synergize with a central data storage. We can see that we need to have a clear view on the whole system architecture and define what is part of the DS and what is outside the DS, but still part of the UNDERPIN architecture. The DS components will be dedicated to sovereign data sharing, enabling participants to access predictive maintenance services under established contractual terms. Such transparency builds trust among DS participants, encouraging broader data contributions

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

and improving the predictive power of the ML models. Meanwhile, training processes can proceed independently outside the DS, ensuring transparency and efficiency without compromising decentralized data-sharing principles. This dual approach ensures that while data remains decentralized, its value is maximized through organized, cohesive processing and utilization.

Data Governance and Security

Ensuring data sovereignty and granular access control was a primary consideration in planning the DS, guiding us to define clear policies on who could access, process, and share data, and under what conditions. Compliance with data privacy regulations, was also central to our planning, prompting us to consider data anonymization techniques and a secure framework for data sharing to mitigate risks of unauthorized access. Additionally, we emphasized the importance of real-time auditing and tracking mechanisms in our design to support transparency and accountability in data usage and access. Altogether, these considerations influenced our planning, laying the groundwork for a robust, secure, and compliant DS infrastructure for future data sharing.

5.2 Lessons learned for stakeholder participation

Data Collection

One important issue that arose during the implementation of UC2.1 (and will similarly impact UC2.2) relates to the automation of data collection on the user's side, in this case MORE. In order to gain access to the data from the various sensors on the wind turbines, as well as the related faults and errors, the operator needs to connect to the wind turbine's SCADA system, provided by the manufacturer, through a dedicated virtual private network (VPN). This means that the user needs to go through several layers of credentials before accessing the data, which creates impediments in automating the data collection process and instead the data needs to be manually downloaded every time in batches. While this does not substantially affect the outcome of the use cases, it is nevertheless creating hurdles in our efforts to further streamline the process. A potential solution has been examined, by utilizing the robotic process automation (RPA) procedures already in place within the Motor Oil group of companies, however said RPAs are not allowed to access remote networks for security reasons. We believe this to be a serious issue that is not unique to MORE (nor the group as a whole) and we anticipate encountering similar issues with other stakeholders further down the line. Moreover, we consider this form of "gating" to be antithetical to the European strategy in regard to data sharing. Nevertheless, efforts to counteract this particular issue are still ongoing, in order to identify a potential solution.

5.3 Lessons learned for use case implementation

Data Quality

Data from industrial production systems and other industry sources is often flawed, not only due to intrinsic issues like sensor noise, equipment malfunctions, and human error, but also due to errors introduced during data export. Exporting large volumes of complex data can lead to misalignments, truncation, format inconsistencies, and even data loss, further affecting data quality. In both presented UCs, we identified such problems and pre-processed the data before the data can be used for further analysis. To assist in overcoming these challenges, the designed UNDERPIN DS offers a service that assesses key data statistics, providing a preliminary analysis

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

of data quality and common issues. This service allows participants to understand the quality of the provided data and better understand potential problems before a data exchange is triggered. By identifying and addressing issues early on, participants can avoid costly and time-consuming errors that might otherwise compromise the reliability of their analysis. Moreover, based on this analysis data scientists and engineers can be supported by their work of data cleaning and pre-processing. Such proactive measures in data quality management not only streamline the ML training process but also enhance the accuracy and consistency of the predictive maintenance models.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

6 Conclusion

This deliverable elaborates on the implementation and validation of the UCs that have been identified as suitable to highlight the benefits of the UNDERPIN DS, building upon the planning established in D4.1: "*Use case planning report*". Deliverable D4.2 provides a comprehensive status update in regard to the implementation of the UCs, potential hurdles and solutions, and lessons learned, and will be complemented with an updated version by the end of the project, which will describe the execution of the UCs and the final outcomes in full detail.

More specifically, a status update was presented for each UC, with initial results for UC1.1: "Monitoring and predictive maintenance in the refinery" and UC2.1: "Predictive maintenance in wind farms", as well as a literature review for UC2.2: "Wind turbine blade repair prediction", as it follows a delayed timeline in order to take advantage of the processes and experience gained from UC2.1. Furthermore, a data pipeline is established, showcasing how the UCs make use of the DS components. Finally, a roadmap with the next steps for each UC is outlined. The document closes with the lessons learned from the first UC implementation stage.

The outcomes of this first version of deliverable D4.2 create the basis for the successful execution of the UNDERPIN UCs, while also identifying potential avenues for improvement.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

7 Bibliography

[1] M. F. J., "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association,* vol. 253, no. 46, pp. 68-78, 1951.

[2] Student, "The Probable Error of a Mean," *Biometrika,* vol. 6, no. 1, pp. 1-25, 1908.

[3] R. J. Hyndman, A. B. Koehler, J. K. Ord and R. D. Snyder, Forecasting with Exponential Smoothing: The State Space Approach, Springer, 2008.

[4] A. C. Harvey, "ARIMA Models," in *Time Series and STatistics*, London, Palgrave Macmillan, 1990, pp. 22-24.

[5] L. Breiman, "Random Forests," in *Machine Learning*, Springer, 2001, pp. 5-32.

[6] G. A. F. Seber and A. J. Lee, Linear Regression Analysis (2nd ed.), Wiley, 2012.

[7] W. Udo and Y. Muhammad, "Data-driven predictive maintenance of wind turbine based on SCADA data.," *IEEE Access, 9, 162370-16238,* 2021.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features.," *Advances in neural information processing systems, 31.,* 2018.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System.," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.,* 2016.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree.," *Advances in neural information processing systems, 30,* 2017.

[11] J. Herzen, F. Lässig, S. G. Piazzetta, T. Neuer, L. Tafti, G. Raille, ... and G. Grosch, "Darts: User-friendly modern machine learning for time series.," *Journal of Machine Learning Research, 23(124), 1-6.,* 2022.

[12] L. Yang and Z. Zhang, "A conditional convolutional autoencoder-based method for monitoring wind turbine blade breakages.," *IEEE transactions on industrial informatics,* vol. 17(9), pp. 6390-6398, 2020.

[13] M. Rezamand, M. Kordestani, R. Carriveau, D. S. K. Ting and M. Saif, "A new hybrid fault detection method for wind turbine blades using recursive PCA and wavelet-based PDF.," *IEEE Sensors journal,* vol. 20(4), pp. 2023-2033., 2019.

[14] K. Chandrasekhar, N. Stevanovic, E. J. Cross, N. Dervilis and K. Worden, "Damage detection in operational wind turbine blades using a new approach based on machine learning.," *Renewable Energy,* vol. 168, pp. 1249-1264, 2021.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179

[15] Y. Tang, Y. Chang and K. Li, “Applications of K-nearest neighbor algorithm in intelligent diagnosis of wind turbine blades damage.,” *Renewable Energy,* vol. 212, pp. 855-864, 2023.

Co-funded by the European Union
This project has received funding from the Digital Europe Programme under grant agreement No 101123179