Circular TwAIn

AI Platform for Integrated Sustainable and Circular Manufacturing

Deliverable

D5.1 Circular TwAIn Data4AI Platform and AI Toolkit - 1st version

Actual submission date: 03/04/2023

Co-funded by the European Union

Project Number:	101058585		
Project Acronym:	Circular TwAIn		
Project Title:	AI Platform for Integrated Sustainable and Circular Manufacturing		
Start date:	July 1st, 2022	**Duration:**	36 months
D5.1 Circular TwAIn Data4AI Platform and AI Toolkit - 1st version			
Work Package:	WP5 - AI-enabled Digital Twins for distributed manufacturing		
Lead partner:	NISSATECH		
Author(s):	Nenad Stojanovic (NISSATECH), Sotiris Koussouris (SUITE5)		
Reviewers:	Elisa Rossi (ENG)		
Due date:	31.03.2023		
Deliverable Type:	OTHER	**Dissemination Level:**	PUBLIC
Version number:	1.0		

Revision History

Version	Date	Author	Description
0.1	20/01/2022	NISSATECH	Deliverable ToC
0.2	01/02/2023	Suite5	Initial Content for AI Toolkit
0.3	02/03/2023	NISSATECH	Initial Content
0.9	19/03/2023	NISSATECH	Section 2 details
0.91	25/03/2023	ENG	Review
1.0	31/3/2023	ENG	Final coordinator review before submission

Table of Contents

List of Figures

List of Tables

Definitions and acronyms

AI	*Artificial Intelligence*
API	*Application Protocol Interface*
CNN	*Convolutional Neural Network*
DL	*Deep Learning*
DoA	*Description of actions*
DT	*Digital Twin*
GS	*Grand Scenario*
KPI	*Key Performance Indicators*
LSTM	*Long Short Term Memory*
MEWMA	*Multivariate Exponentially Weighted Moving Average*
ML	*Machine Learning*
PCA	*Principal Component Analysis*
RNN	*Recurrent Neural Network*
SME	*Short and Medium Enterprises*
SR	*Stability Ratio*
WP	*Work Package*
XAI	*Explainable AI*

Disclaimer

This document has been produced in the context of Circular TwAIn Project. The Circular TwAIn project is part of the European Community's Horizon Europe Program for research and development and is as such funded by the European Commission. All information in this document is provided 'as is' and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability with respect to this document, which is merely representing the authors' view.

Executive Summary

D5.1, Circular TwAIn Data4AI Platform and AI Toolkit - 1st version, is the first deliverable of WP5, one of the two technical Work Packages of the Project. The type of the deliverable is 'OTHER' as per the GA, in fact it contextualizes and presents in a readable way what has been and will be implemented under T5.1, Data4AI Platform, Data Quality and Pipelines, and T5.2, AI for Circularity and Resilience applications Toolkit.

While dealing with data coming from multiple sources that should feed and train AI algorithms, the major challenges regard the techniques to be adopted for data preparation. The study over these strategies is linked on one side with the pilots' expectations (the last information on that have been reported in parallel of this deliverable - M9, in D2.2 and D6.1), and on the other with the technical resolutions on Data Spaces defined in WP4, that should pave the way for the AI Platform infrastructure. The WP5 objective "creation of the data processing pipeline for ensuring data quality", is discussed in Section 2: it provides some foundational aspects related to the Data4AI Platform are presented, as user journey and basic requirements. It explains the development of the data.

Section 3 instead is related to the second objective of WP5, i.e., "the development of various tools for learning models required for AI applications". It explains the basis for the development of the AI toolkit that will be able to not only provide a set of baseline machine learning and deep learning algorithms to address the diverse intelligence problems at hand in an out-of-the-box manner leveraging state-of-the art AI and XAI libraries/frameworks, but will go further and cater for the easy integration of already trained models/pipelines to address specific circular and sustainable manufacturing needs.

This deliverable accompanies the software code developed in the scope of the tasks T5.1 and T5.2.

The final results of T5.1 and T5.2 will be reported at M27 in the second version of this deliverable, namely, D5.4, Circular TwAIn Data4AI Platform and AI Toolkit - 2nd version.

1 Introduction

1.1 Scope of the Deliverable

D5.1, Circular TwAIn Data4AI Platform and AI Toolkit - 1st version, presents the first results of WP5, AI-enabled Digital Twins for distributed manufacturing, reporting on the activities performed from M3 to M9 of the Project. WP5 is strongly linked with the work carried out in WP4, Data Space for Circular and Resilient Manufacturing, enabling the integration and combination of different data from various sources over the entire product life cycle, considering sustainability aspects.

The report is related particularly to tasks T5.1 and T5.2:

- Task 5.1 enables an efficient preparation of the data for being used in various innovative services for circular and sustainable manufacturing.
- Task 5.2 is responsible for the development of a toolkit for applying different ML/DL models, designing-executing-orchestrating appropriate AI/XAI pipelines.

The objectives of this tasks are the creation of the data processing pipeline for ensuring data quality, and the development of tools for learning models required for AI applications.

This deliverable accompanies the software code developed in the scope of the tasks T5.1 and T5.2.

1.2 Structure of the Document

Apart from the Introduction, the deliverable is structured as follows:

- Section 2 presents the results related with task T5.1, providing requirements for Data4AI Platform and describing methods for AI-based analysis,
- Section 3 focuses on AI toolkit for applying different ML/DL models in Circular TwAIn use cases,
- Section 4 contains concluding remarks and next actions.

2 Data4AI Platform, Data Quality and Pipelines

In this section some foundational aspects related to the Data4AI Platform are presented, as user journey and basic requirements.

2.1 Data4AI Platform - overview

Data4AI Platform is a new generation of platforms for ensuring data quality for AI applications. The main goal is to provide an efficient and easy for usage infrastructure for enabling manufacturing SMEs to prepare own data for the usage in AI applications. It includes the adapters for connecting relevant data sources and recipes (workflows) for defining data preparation pipelines (or using the available pre-configured ones). The main objective is to include the domain expertise in the data preparation process, but in a convenient way for non-technical expert. In addition, to ensure the reliability of collected data, methods for checking the completeness and validity of data will be applied on the edge (avoiding two most important problems for the AI: missing and corrupted data).

The development started in the MIDHI project, where the focus was on the (traditional) data cleaning process.

Platform has been evolved in AI REGIO project, by introducing StreamPipes-based realization of the data quality pipelines (https://streampipes.apache.org/) and more advanced processing elements, including intelligent process contextualization and Statistical learning.

In the task 5.1 the extensions in the AI and Circularity direction have been planned. In the first period the focus is on the inclusion of more AI-based processing in the data quality pipelines and in the collection of the requirements for the Circularity-based scenarios.

2.2 User Journey and Grand Scenarios

The main potential users/adopter of the Data4AI Platform identified up to now are the following which are described together with a first scenario explaining their main role:

Stakeholder Group Name	**Process Quality Monitoring and Continuous Process Improvement (Circular economy)**
Tasks	Ensuring the best possible performances of the process and its continual improvement
Personal interests	To be able to consider automatize the process improvement approach
Required expertise (description and level)	Statistical process control, basics Root-cause analysis, basics Circular economy
Comments	

Grand Scenario GS1.1	**Understanding process behaviour in the circularity context**
Stakeholder(s)	Process owner

Scenario	Process owner is interested in understanding behaviour of the process involved in circularity scenario (e.g. remanufacturing). Usually, KPIs are the most suitable way for determining how properly does the process behave. However, KPIs are summary views on the process performances and usually don't reflect the process behaviour, but only its outcome. Consequently, they are not providing insights into improvement from the circularity point of view. Thanks to the introduction of Data4AI platform, Process owner can understand how the process behaves, esp. its variations in time, interpreted in the circularity contexts. A very important feature is the root cause analysis, which for each issue finds the most relevant causes. This analysis is multivariate, i.e., issues are related to several parameters together contributing to the process variations.

GS1.2	**Understanding process quality in the circularity context**
Stakeholder(s)	Process owner
Scenario	Process owner is interested in ensuring the quality of the process by considering circularity aspects. However, KPIs are not providing insights for the circularity context. Thanks to the introduction of Data4AI platform, Process owner can understand how the process quality depends on circularity aspects. In that way it is possible to detect periods of the process execution which are very energy consuming or environmental unfriendly. A very important feature is the root cause analysis, which for each issue finds the most relevant causes. This analysis is multivariate.

Stakeholder Group Name	**Data quality assurance (circularity context)**
Tasks	Ensuring the best possible quality of the data used in AI / data analytics tasks in the circularity context
Personal interests	To be able to develop advanced data-driven solutions
Required expertise (description and level)	Statistics, basics Data quality
Comments	

GS2.1	**Data preprocessing**
Stakeholder(s)	Data Quality engineer
Scenario	Data Quality engineer wants to ensure that the data used in the possible AI applications has high quality, esp. to be used in the circularity scenarios (e.g. remanufacturing) She/he should define relevant criteria for defining the data quality.

	Thanks to the introduction of Data4AI platform, the Data Quality engineer can prepare the data in an optimal way. She/he can select the most important characteristics from the process (like stability, consistency, …) and the circularity (like completeness, accuracy, …) context to be checked and the system defines relevant data preprocessing pipelines automatically.

GS2.2	**Monitoring the quality of the data collection process (data observation)**
Stakeholder(s)	Data Quality engineer
Scenario	Data Quality engineer wants to ensure that the data is collected in the best possible way, i.e. that the process in which data is collected is well designed Thanks to the introduction of Data4AI platform, the Data Quality engineer can monitor the performance of the data collection process. She/he can select the most important characteristics of the data collection process (like completeness, …) and the system defines relevant data monitoring rules automatically.

Stakeholder Group Name	**Data valuation**
Tasks	Defining the value of the data to be used in processing / sharing and esp. circularity scenarios
Personal interests	To understand the value of chosen data – to be offered for sharing
Required expertise (description and level)	Statistics, basics
Comments	

GS3.1	**Calculate data value**
Stakeholder(s)	Data Quality engineer
Scenario	Data Quality engineer wants to understand what is the value of selected data. The value is related, in the first line, to the potential of data for being used for the process improvement tasks. In that context, she/he wants to check some of the process characteristics, like stability and capability. However, it requires a complete Statistical Process Control for a multivariate case, which is usually not available. Thanks to the introduction of Data4AI platform, the Data Quality engineer can check the value of data in an efficient way. She/he can select the most important parameters to be analyzed and the system defines relevant values (of data) automatically.

GS3.2	Calculate value of data for circularity scenarios
Stakeholder(s)	Data Quality engineer
Scenario	Data Quality engineer wants to understand what is the value of selected data for circularity scenarios. The value is related, in the first line, to the potential of data for being used for improving energy and environmental aspects. Thanks to the introduction of Data4AI platform, the Data Quality engineer can check the value of data in the given circularity context. She/he can select the most important parameters to be analyzed and the system defines relevant values (of data) automatically.

2.2.1 Requirements

In the following table summarizes the requirements for Data4AI Platform.

Table 2-1: Requirements for Data4AI Platform

ID	Description	Priority	Stakeholders
R1.1	Enabling to connect to a data source (industry settings)	H	Process owner
R1.2	Enabling to define a set of criteria for process quality	H	Process owner
R1.3	Enabling to upload / process past data (labeled)	H	Process owner
R1.4	Enabling to define the process/product quality (beyond traditional KPIs)		Process owner
R1.5	Providing visual presentation of results	H	Process owner
R1.6	Providing explanations	H	Process owner
R2.1	Enabling to select the data source	H	Data Quality engineer
R2.2	Enabling data observation (monitor data collection)	M	Data Quality engineer
R2.3	To define a set of criteria for data quality / circularity context	M	Data Quality engineer
R2.4	Enabling define various monitoring rules	H	Data Quality engineer
R2.5	Enabling to check / validate created pipeline	H	Data Quality engineer
R2.6	Provide automatic reporting about the data quality / circularity context	H	Data Quality engineer
R2.7	Provide report about the data preprocessing results	H	Data Quality engineer
R2.8	Enable reconfiguration of the pipeline	M	Data Quality engineer
R2.9	Enable to check / validate created pipeline in the circularity context	M	Data Quality engineer

ID	Description	Priority	Stakeholders
R2.10	Enable automatic validation of the pipeline (based on defined KPIs)	M	Data Quality engineer

2.3 Data4AI Conceptual model

This section presents the model, which the Data4AI Platform is based on.

2.3.1 AI Processing Pipeline

The following figure presents a high-level view on the AI applications (general case).

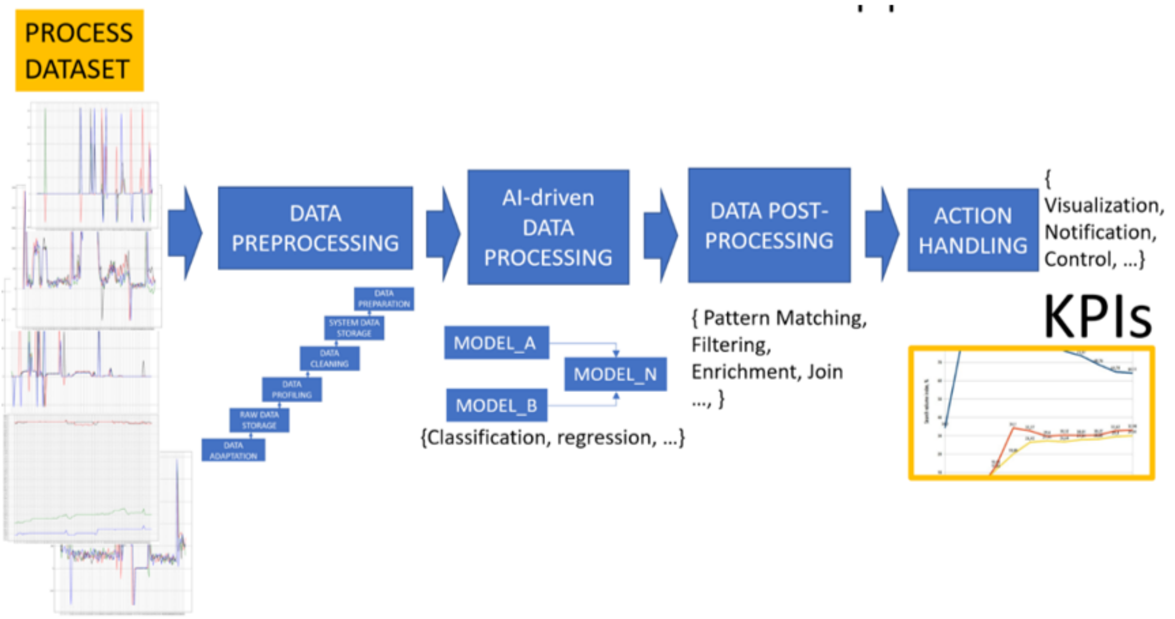

Figure 2-1: Data processing pipeline for AI applications

The steps in the pipeline are briefly described:

- Data Preprocessing is a processing pipeline which transforms raw data in the well-formed data (valid structure) that can be processed by various data analysis methods
- AI-driven Data Processing is data analysis which can be done within or outside Data4AI Platform
- Data Postprocessing enables preparation of the data for output (e.g., filtering)
- Action Handling is related to the delivery of the output to other (control, notification, visualization) systems.

2.3.2 Data4AI pipeline

Figure 2-2 presents the Data4AI approach for introducing AI-based processing in the data processing pipelines. This is one of the main innovations to be introduced: using AI for generation of the additional context (e.g., process) to be used in the data processing pipelines.

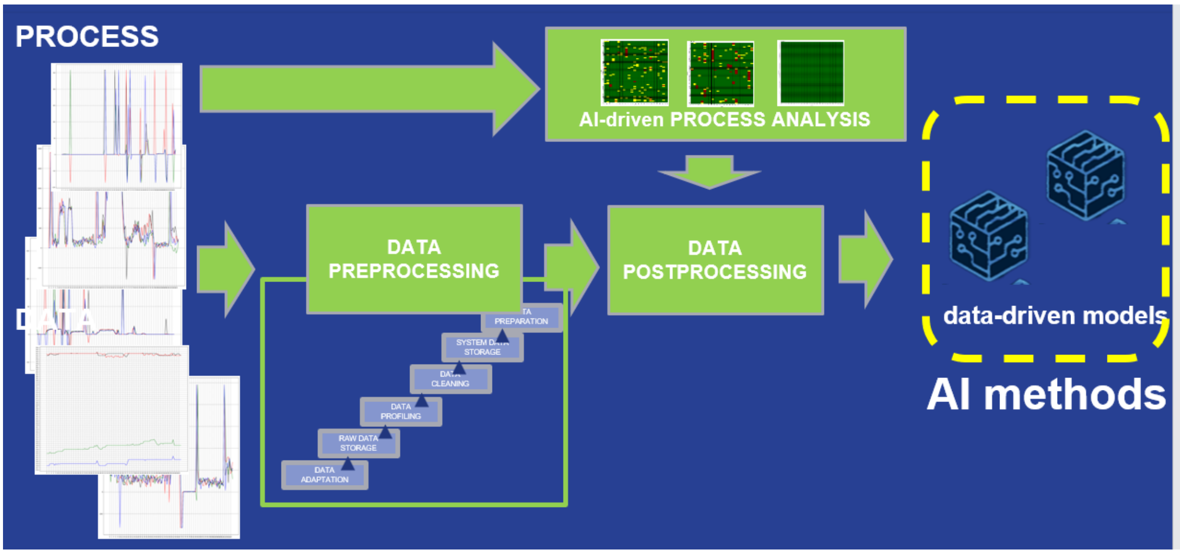

Figure 2-2**:** Data4AI AI-driven Data Processing approach

Through Circular TwAIn the circularity context will be introduced in the Data4AI Platform.

2.3.3 AI-driven analysis

In this section the AI-driven analysis is presented, which corresponds to the AI-driven data processing depicted in Figure 2-1.

The following figure (Figure 2-3) illustrates the three methods for AI-driven analysis, which are described in the following subsections.

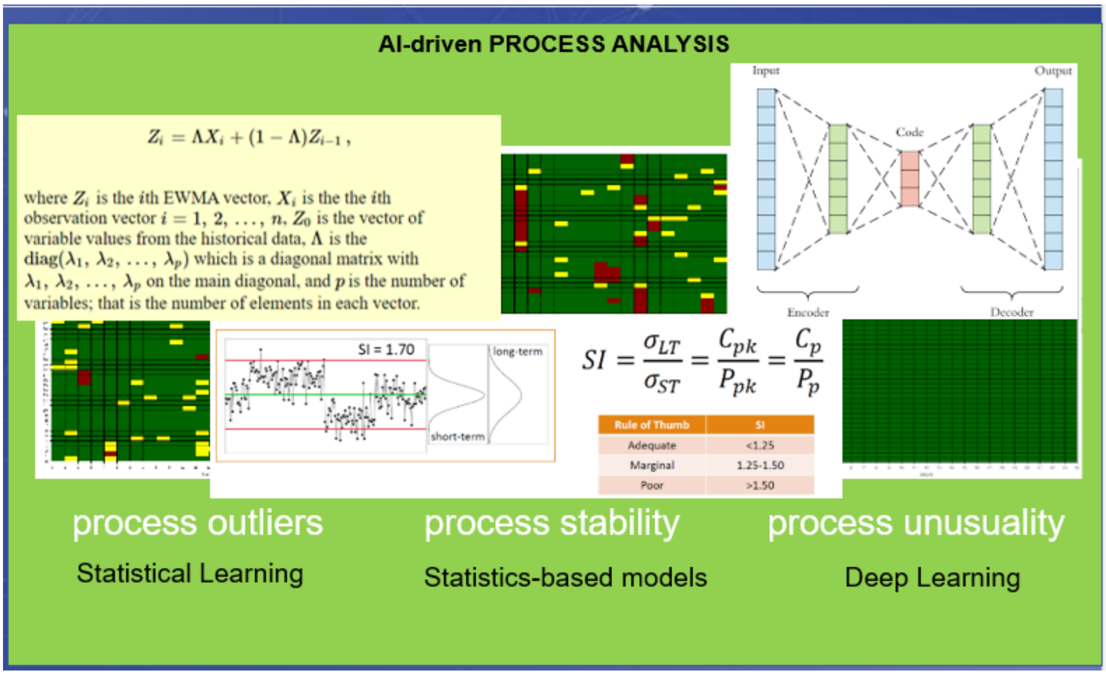

Figure 2-3**:** AI-driven Data Processing methods

2.3.3.1 Outlier detection

The outlier detection uses three different analyses: Multivariate Exponentially Weighted Moving Average (MEWMA), Principal Component Analysis (PCA), and Hotelling.

MEWMA Analysis

The Multivariate Exponentially Weighted Moving Average[1] (MEWMA) analysis is a statistical method used for detecting small shifts in the mean vector of a multivariate time series. It is used to track the deviation of a new data point from the expected value of the series based on the previous data points. The analysis is performed by calculating the weighted average of the previous data points, where the weights decrease exponentially as moving further away from the most recent data point.

PCA Analysis

The Principal Component Analysis[2] (PCA) analysis is a technique used to reduce the dimensionality of a dataset by identifying the most significant variables in a multivariate dataset. It transforms the dataset into a set of orthogonal components, where each component explains the maximum amount of variance in the data. This analysis is used to identify the direction of maximum variation in the data, and it can be used to detect outliers.

Hotelling Analysis[3]

The Hotelling analysis is a statistical method used to detect outliers in multivariate datasets. It is based on the T^2 statistic, which measures the distance between a data point and the center of the dataset in the direction of maximum variation. The analysis is performed by calculating the T^2 statistic for each data point, and then identifying the data points that are significantly different from the others.

Outlier aggregations

Aggregations are hierarchically organized and use one of the aggregation methods:

- **Simple vote** – If any of three selected methods returns that the datapoint on that timestamp is an outlier, then the timestamp is outlier.
- **Majority vote** - If at least two of the analyses identify the data point as an outlier, it is classified as such.

2.3.3.2 Root Cause detection

If a data point is identified as an outlier, it is then matched with a root cause that was previously calculated. This matching helps to understand the reason for the outlier and to take corrective actions to prevent future occurrences.

However, it appears that finding a “cheap” and fast algorithm to find the root cause of an out-of-control signal was a very hard task and in last few years many researchers were trying solve this problem. For example, to solve this problem some authors purposed to make all possible combinations of parameters and then check if correlation is “broken”. However, this method can be used only with a small number of parameters. For example, if the parameters are 10, check all combinations of 2 parameters and then all combination of 3 parameters and so on, will end up with checking 10! Combination, which is equal to 3 628 800. This method is naive and same as brute force. It has exponential time complexity and in practice it is not applicable, since even the fastest computers will need hours or maybe days to check

[1] https://www.itl.nist.gov/div898/handbook/pmc/section3/pmc343.htm
[2] https://en.wikipedia.org/wiki/Principal_component_analysis
[3] https://online.stat.psu.edu/stat505/lesson/7/7.1/7.1.3

all of these combinations, while Project need is to run this algorithm whenever an out-of-control signal on multivariate control charts is found.

T2 decomposition is the fastest and the best algorithm found so far. The algorithm works as follows:

1. Run root cause detection algorithm only when multivariate control chart finds an out-of-control signal.
2. Compute T2 based on all parameters (named as T2_total for p parameters).
3. Compute T2 without one parameter and repeat process for all parameters (T2 based on p-1 parameters).
4. For each parameter make a difference between T2_total and T2 with that one parameter and store results into temporarily list.
5. The biggest difference indicates that specific parameter is responsible for an out-of-control signal.
6. There can be more than one parameter which is responsible for out-of-control signal, thus steps 2-5 need to be repeated, but staring with p-1 parameters. Parameter which caused the biggest difference in 1st iteration is removed from further calculating and added to list of responsible parameters for out-of-control signal.
7. Return the list of responsible parameters for out-of-control signal.

Each iteration has quadratic time complexity due to matrix multiplying and the maximum iterations can be p, which means that total time complexity of this method is O(p4), where p is the number of parameters. This time complexity is much better than a brute force approach and it is acceptable because usually not too much parameters are involved in the cases of interest. In fact, most of Circular TwAIn use cases have less than 20 parameters. However, knowing that it can be problematic for use cases when p > 50 the research of a better solution is still an ongoing activity.

2.3.3.3 Variation detection

Variations are defined as changes in the normal flow of a process.

The method is illustrated in Figure 2-1. Briefly, Variation detection is performed using Nelson rules[4], which are a set of statistical process control methods used to detect shifts in process data. The following Nelson rules are used for detecting variation: 'rule1', 'rule3', 'rule4', 'rule5', 'rule6', and 'rule8'.

In addition to the Nelson rules, three more steps are used for further analysis:

[4] Nelson rules - Wikipedia

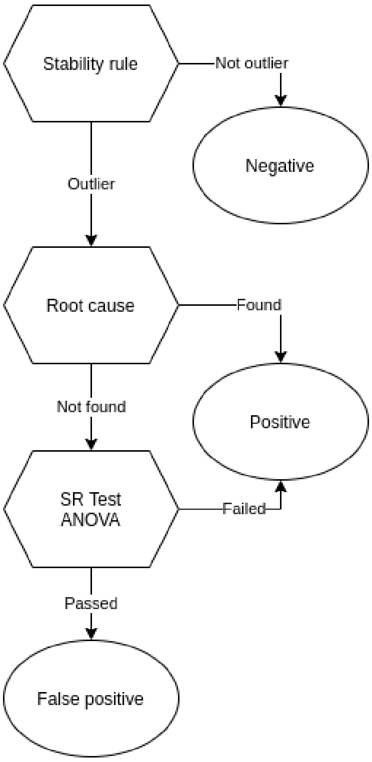

Figure 2-4: Variation detection process

1. **Length of the period –** if the length is 1 datapoint, this period (or one datapoint) is considered as unstable and should be marked as true positive.
2. **Root cause –** for each datapoint the root causes are calculated, if there is more than set percentage (currently 50%) of datapoints which have at least one parameter as root cause, the period is considered as true positive.
3. **Stability Ratio (SR) test**[5] –current period is compared with the batch of data for each parameter. If there is at least one parameter which has SR more than 1.5, the period is considered as true positive.

Results of Nelson rules are periods, and if any Nelson rule has over 15% of all timestamps (when periods for each rule are combined), then that rule is considered as a bad rule for current dataset and it is removed.

All periods which are deduced by the algorithm as true positives are further aggregated in the same way as Outlier detection.

2.3.3.4 Unusuality detection

Based on performed testing, the usage of LSTM autoencoder network for the purpose of Unusuality detection for time series is very promising. The main idea is to build an autoencoder to learn the data distribution used for encoding which represents the normal behaviour of a time series. The encoding is validated and refined by attempting to regenerate the input from the encoding. If output data differs significantly from the input data, it is considered an anomaly. In ideal cases, only data with normal instances are used to train the autoencoder, in others, the frequency of anomalies is small compared to the observation set

[5] [(PDF) Stability assessment with the stability index (researchgate.net)](#)

so that its contribution to the learned representation could be ignored, meaning that the number of anomalies in the data set must be less than 5%. When facing anomalies, the model would worsen its reconstruction performance. After training, the autoencoder will accurately reconstruct "normal" data, while failing to do so with anomalous data.

The proposal is to test the stability of the process by calculating stability index or employing some other method to make sure that given data set satisfies the condition mentioned above.

LSTM is employed to process time series data. LSTM is capable of automatically extracting effects of past events, both long- and short-term effects.

Method description

The Circular TwAIn applications involve a multivariate time-series data. The LSTM autoencoder should be built on this multivariate time-series to perform detection of rare events (anomalies/unusual behavior). This is achieved by using an anomaly detection approach:

1. Build an autoencoder on mainly normal data (data without anomalies or data with at most 5% anomalous points)
2. Use data to reconstruct a new sample
3. if the reconstruction error is high, label it as anomaly

LSTM is a type of Recurrent Neural Network (RNN). RNNs, in general, and LSTM, specifically, are used on sequential or time series data. LSTM requires few special data-preprocessing steps. The input to LSTMs are 3-dimensional arrays created from the time-series data. The shape of array is *samples x lookback x features:*

1. samples: this is simply the number of observations
2. lookback: this is how many points the LSTM will process up to the observed point
3. features: this is the number of features present in the input data

It is usually better to use a standardized data for autoencoders, and the standardization has to happen with respect to the original 2D data.

Model architecture

The objective of fitting the network is to make this output close to the input. For better understanding the LSTM autoencoder flow diagram as only 2 features, and sequence length is 3. The diagram illustrates the flow of data through the layers of an LSTM Autoencoder network for one sample of data. A sample of data is one instance from a dataset.

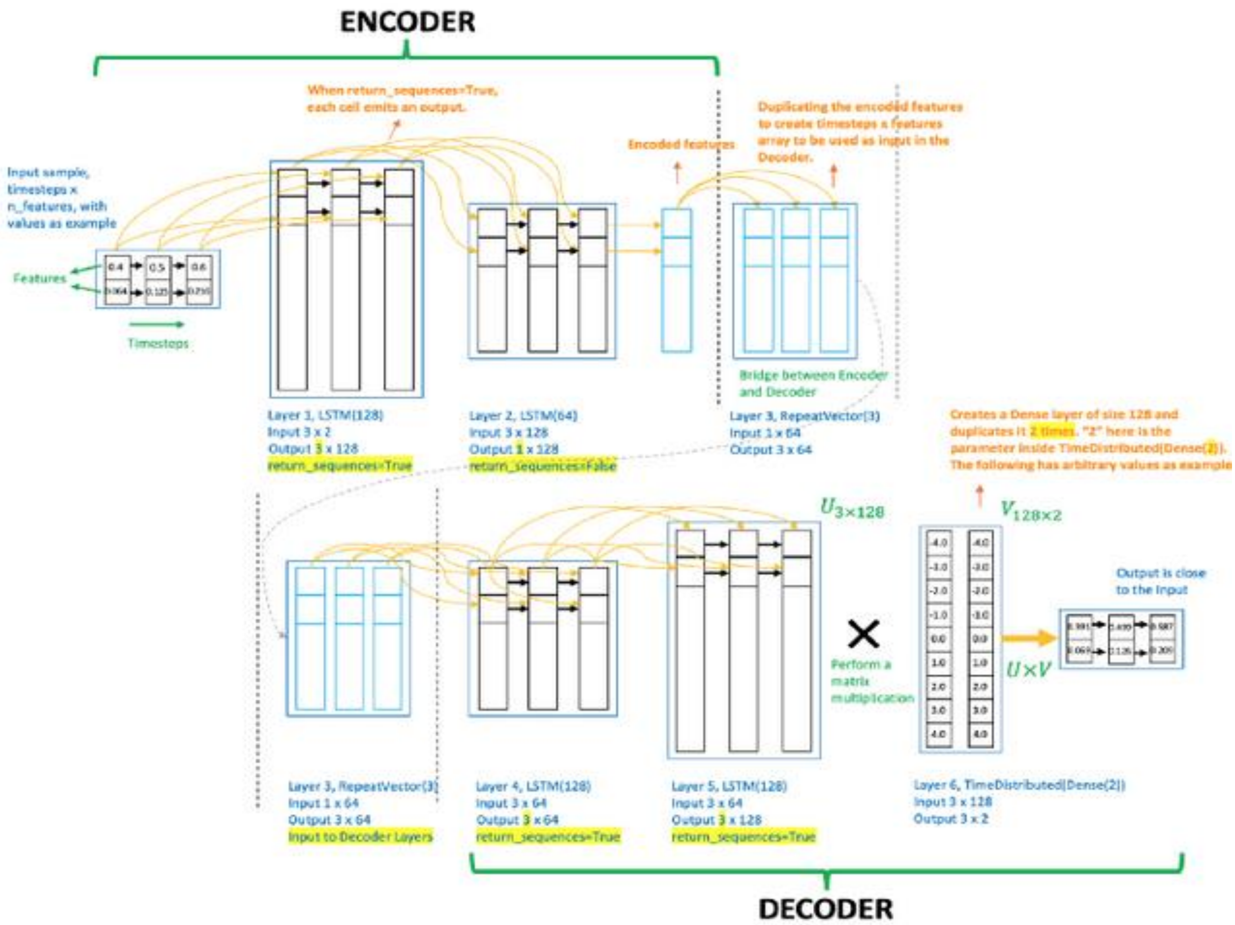

Figure 2-5: Unusuality detection model

A key attribute of recurrent neural networks is their ability to persist information, or cell state, for use later in the network. This makes them particularly well suited for analysis of temporal data that evolves over time. LSTM networks are used in tasks such as speech recognition, text translation and here, in the analysis of sequential sensor readings for anomaly detection.

The lesson learnt from this diagram is:

1. The LSTM network takes a 2D array as input.
2. One layer of LSTM has as many cells as the timesteps.
3. Setting the return_sequences=True, makes each cell per timestep emit a signal.

This becomes clearer comparing the difference between return_sequences set to true and false:

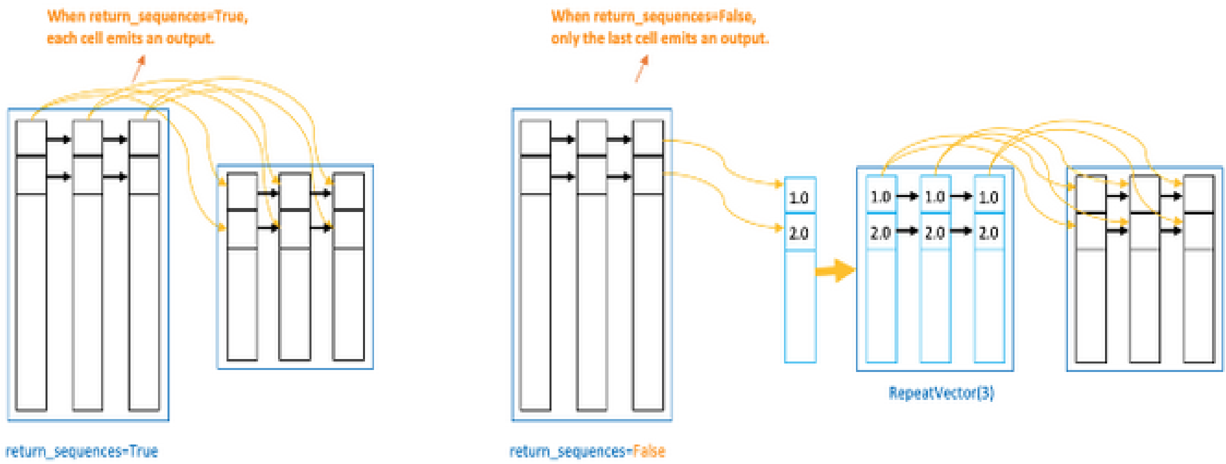

Figure 2-6: Unusuality detection: calculation methods

When return_sequence is set to 1, signal from a timestep cell in one layer is received by the cell of the same timestep in the subsequent layer. When return_sequence is set to 0, only the last timestep cell emits signals. The output is, therefore, a vector.

In the encoder and decoder modules in an LSTM autoencoder, it is important to have direct connections between respective timestep cells in consecutive LSTM layers, so return_sequence is set to 1.

The output of encoder part is the **encoded feature vector** of the input data. This encoded feature vector can be extracted and used for data compression.

Once the model is built, the expectation is the autoencoder to reconstruct anomalous points with high reconstruction error, such that anomalous from non-anomalous points can be separated. The threshold for this should be determined.

Model evaluation

The trained model may be evaluated through two different approaches. First approach has to be adopted while implementing the method, and while training the model on data with labeled anomalies. In such environment, since this problem is binary classification (labeling points as anomalous or normal), most commonly used metrics for evaluating anomaly detection solution are Accuracy, Precision, Recall and F1 score.

Accuracy= TP+TN/TP+FP+FN+TN

Precision = TP/TP+FP

Recall = TP/TP+FN

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Second approach is applied to a real usage of model, when training data without information about anomalous points. The only way to determine the correctness of model in this situation is to compare results with other methods, giving the ratio of matching between models. For this purpose, in Circular TwAIn different methods from statistical modeling and different machine learning techniques will be employed. So far, the investigation is directed to control charts and prediction with CNN.

3 AI Toolkit for Circularity and Resilience Applications Toolkit

3.1 Overview

As the Project's DoA reflects and the Reference Architectures points out, the different Digital Twins to be operating within the Project will be powered by an infrastructure that will be in the position to offer to AI and XAI methods that are essential for the operations of the DTs.

For this purpose, an AI toolkit is being developed in the frame of the Project that will be able to not only provide a set of baseline machine learning and deep learning algorithms to address the diverse intelligence problems at hand in an out-of-the-box manner leveraging state-of-the art AI and XAI libraries/frameworks, but will go further and cater for the easy integration of already trained models/pipelines to address specific circular and sustainable manufacturing needs. In order to do so, the AI toolkit takes inspiration and reuses results of reference AI implementations that are being recorded by Task 3.5 of the Project, and aims to deliver models and explainability methods that are both tailored to the needs of the DTs to be developed in the Project, but also offers generic methods and algorithms that can satisfy more broader needs, allowing in this manner future DTs to onboard and develop their own models and algorithms.

3.2 AI Toolkit Architecture View

To satisfy the needs set in the previous chapter, the architectural design of the AI Toolkit has been constructed in order to guarantee the following principles:

- Be in a position to offer DT agnostic AI pipelines and XAI services, and at the same time satisfy the needs of the DTs of the Project,
- Be able to ingest FAIR datasets, depending on and interoperating with services that guarantee data security, trustworthiness, interoperability and quality,
- Cater for performance and resilience, supporting the local and cloud-based execution of the different models, depending on the constrains set by the operational environment of the DTs and the data and demonstration scenarios to be chosen.

In this context, the overall AI toolkit can be seen as a complex infrastructure that offers the following:

- The ability to design AI Pipelines both directly with using code (experimentation – targeting data scientists) as well as via a user friendly and more production-grade interface (cloud-based AI designer) that can schedule and coordinate the execution of the models.
- The ability to deploy and execute these AI pipelines and provide XAI methods either locally in the DTs or over the designated cloud-based infrastructure.
- The ability to expose the AI results and make use of XAI methods, making them available directly to the directly operations of the DTs.

The following figure presents the high-level architecture of the AI Toolkit, which has been used to deliver the toolkit version at M9 of the Project.

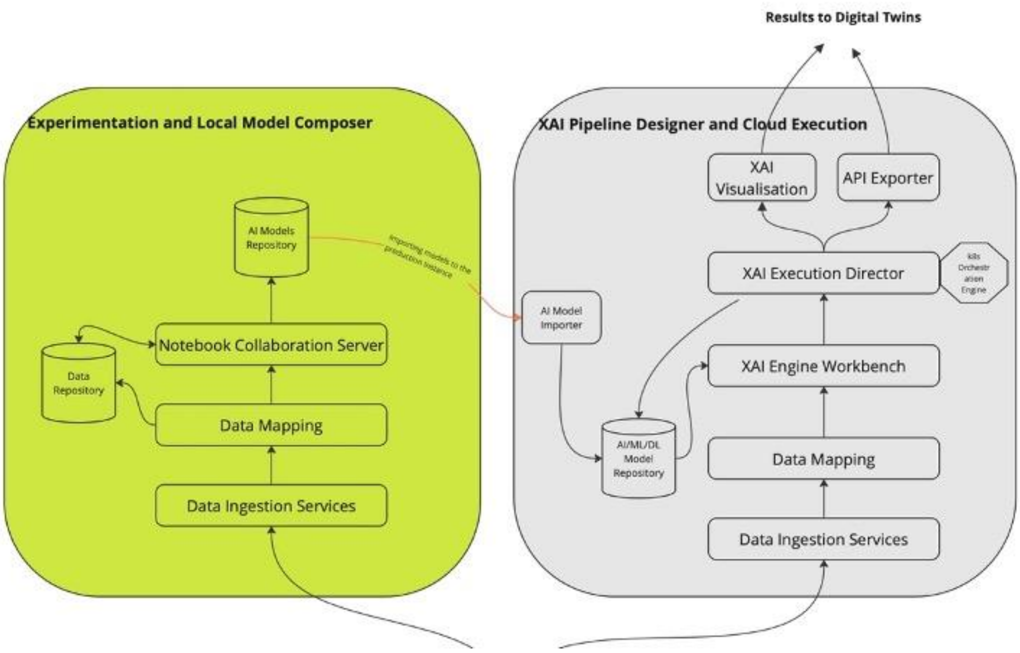

Figure 3-1: AI Toolkit High Level Architecture

As shown in the diagram above, the complete toolkit consists of two major components, which are the “Experimentation and Local Models Composer Interface” found on the left side, and the “XAI Pipeline Designer and Cloud Execution service”.

Both of those components are fed with data (southbound) coming from the Data Spaces and/or directly from the Data Pipelines which are responsible for treating the data and make them usable foe the AI to be executed.

The results of both components are offered to the DT via APIs, either internal one (in case the execution is placed inside the DT, or external ones, that communicate with the DT as the execution is taking place in an external to the DT’s virtual environment.

3.3 Experimentation and Local Models Composer Interface

The Experimentation and Local Model Composer is designed to enable data scientists to effortlessly set up AI experiments. Within this environment, incoming data (i.e., from data pipelines) can be leveraged to build proper datasets that can be stored and used later in the model building phase. The playground is a notebook-based server equipped with dashboarding tools, and its primary objective is to assist data scientists in managing their own projects, manipulating data interactively, and creating AI model artifacts that can be further stored in a local repository. This approach facilitates tracking and reproducing experiments, resulting in a continuous evolution of AI models.

Figure 3-1 illustrates the Experimentation and Local Model Composer, which is equipped with the following set of tools:

- **Notebook Collaboration Server**, being the core of the experimentation playground, provides data scientists with a collaborative environment that enables them to manipulate data and create AI models. This server is based on Jupyter Notebook, which facilitates real-time collaboration, allowing multiple users to edit the same file simultaneously while others can view the changes in real-time, ultimately leading to increased team productivity. Furthermore, this component permits users to utilize various programming languages and ML/DL frameworks and libraries, thereby fostering cooperation among data scientists with different backgrounds. The notebook server receives data from two main sources: the first source, implemented with the data ingestion and mapping services, acts as an input for real-time data, while the second one, the data repository, provides historical data already prepared in previous experiments. Additionally, the experimentation server is equipped with features that enable data scientists to seamlessly repeat experiments, visualize intermediate results, and browse file history to track the steps that led to a particular result. The artifacts produced by the experimentation playground are data models and AI models.

- **Data Repository** component is responsible for storing all data flowing through the system. It receives data from the underlying data pipeline component, which are correctly mapped into structured datasets, or from the notebook server after appropriate processing steps decided by data scientists. In this way, the repository stores both raw data from input sources and different versions of datasets that are ready to use in experiments.
- **AI Model Repository** is a component that stores different versions of the AI models generated in the experiment playground. It serves as a repository of models that can be further improved in future experiments, as well as a catalog of validated artifacts that can be accessed from the XAI Pipeline Designer and Cloud Execution Service macro-component.

3.4 XAI Pipeline Designer and Cloud Execution services

The XAI Pipelines designer and Cloud execution services provide to users a configurable framework that is fed with data coming from the Data Twins and/or directly from the Data Pipelines and is responsible for treating the data and make them usable for the AI to be executed. It consists of an intuitive dashboard that enables users to design, orchestrate and execute advanced XAI pipelines based on existing analytic models and XAI algorithms, configured appropriately to comply with their needs, towards extracting meaningful insights from the data they own and/or data coming from the DTs. Overall, the XAI Pipelines designer provides various data blocks to be used during the pipeline's design, ranging from statistics computations and visualisations to pre-trained machine learning (ML) algorithms (e.g., regression, classification, clustering, forecasting, etc.) deep learning (DL) models, well-developed analytics workflows, and configurable model training workflows.

As shown in Figure 3-1, the XAI Pipelines designer and Cloud execution services provide a set of data analytics components which are namely:

- the **XAI Workbench,** responsible for delivering a configurable framework, supporting Spark[6] Scikit[7], TensorFlow Python[8] execution libraries, enabling users to design analytics pipelines, utilizing pre-processed data assets (directly from the users or through the AI/ML/DL Model Repository which in turn is responsible for storing and retrieving data model), so as to be fed into the various analytics functions and XAI

[6] https://spark.apache.org
[7] https://scikit-learn.org
[8] https://www.tensorflow.org

algorithms appropriately configured by the users to comply with their needs. The XAI Workbench offers to its users various blocks to be utilised in the analytic pipelines, including data input blocks, data preparation blocks (drop, create columns, sort, datetime and datatype-transformations, aggregations, etc.), data control blocks (e.g., For Loop), ML blocks including Application blocks (such as classification, clustering, Keras model, etc.) Evaluation blocks (such as regression) and Train block (such as DBSCAN Clustering, Decision Tree regressor, etc.) as well as output blocks for saving and visualising of results.

- the **XAI Execution Director**, responsible for running (executing) the user-configured analytics pipelines and generating the corresponding results. Users are also enabled to schedule the execution of the analytics pipelines also defining their preferred execution frequency (e.g., hourly, daily, weekly, or monthly). The user can easily shift between 3 different execution runs. Through the “Graph view” users can undertake a dry run of the pipeline, through the “Table view” users can have a test run using a sample of the selected dataset and finally through the “Results view” users can see the corresponding results of the pipeline, where a full run is undertaken utilising the whole dataset.
- the **XAI Visualisation** component that generates the visualisations of the respective analytics pipelines' results, enabling users to edit how these will be visualised by defining the format of the legends, the type of the graph, etc.
- the **API Exporter** which enables users to export the analytics results, providing the suitable APIs, either internal ones, in case the execution is placed inside the DT; or external which communicate with the DT as the execution is taking place in an external to the DT’s virtual environment.
- The **XAI Pipelines designer** includes also data management/manipulation components (the Data Collection Services and the Data Mapping), which undertake the pre-processing of the data assets in an appropriate format that serve the needs of the analytics services.

4 Conclusion and Future Outlook

This deliverable is related to the WP5, particularly to tasks T5.1 and T5.2, aiming to achieve two objectives:

- creation of the data processing pipeline for ensuring data quality,
- development of various tools for learning models required for AI applications orchestrating appropriate AI/XAI pipelines.

Regarding the first objective, some foundational aspects related to the Data4AI Platform are presented, as user journey and basic requirements. In particular, six grand scenarios and sixteen requirements have been described. In addition, the initial design of the Platform is depicted. Furthermore, the set of AI-based methods which will enable dealing with unknown situations, i.e., ensuring the data quality in the case of dynamic environments is investigated.

Regarding the second objectives, the basis for the development of the AI toolkit are explained. It will be able to not only provide a set of baseline machine learning and deep learning algorithms to address the diverse intelligence problems at hand in an out-of-the-box manner leveraging state-of-the art AI and XAI libraries/frameworks, but will go further and cater for the easy integration of already trained models/pipelines to address specific circular and sustainable manufacturing needs.

Regarding Data4AI Platform, next activities will be related to the testing and refinement of the data pipelines on the pilot datasets. In the first phase we will test the "traditional" aspects of the data quality pipelines, whereas in the second phase we plan to test circularity aspects.

As imminent next steps for the evolution of the AI toolkit focuses on the following points:

- Design of tailored Explainability interfaces for the different scenarios of the DTs,
- Seamless integration with all different data connection points offered by the data pipeline methods,
- Ability to transfer models from the Experimentation side to the Production site,
- Local deployment of Experimentation interface notebooks,
- Local execution of compressed AI models designed with the AI Pipeline Designer.

Co-funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058585