AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Deliverable 3.2

D3.2: Predictive AI for process quality assurance

WP3: Platform AI analytics and decision-making support

T3.2: Predictive AI for process quality assurance

Version: 1.0

Dissemination Level: PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

Table of Contents

List of Figures

List of Tables

Disclaimer

This document contains a description of the AI-PROFICIENT project work and findings.

The authors of this document have taken any available measure for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AI-PROFICIENT consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

AI-PROFICIENT has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957391.

Title: D3.2: Predictive analytics for quality assurance

Lead Beneficiary:	TEK
Due Date:	31/01/2023
Submission Date	24/02/2023
Status	Final ~~Preliminary Draft~~
Description	Contribution of AI technologies to ensure process and product quality within a manufacturing process
Authors	Susana Ferreiro (TEK)
Type	Report
Review Status	~~Draft WP Leader accepted.~~ PC + TL accepted
Action Requested	~~Contribution from partners leading in each UC the feedback collection and management.~~ ~~To be revised by partners~~ ~~For approval by the WP leader~~ For approval by the Project Coordinator & Technical Leaders

VERSION	ACTION	OWNER	DATE
0.1	First Draft for partners review	TEK	18/12/2022
0.2	Second Draft for WP Leader review	TEK	30/01/2022
0.3	Final version for approval	TEK	17/02/2022
1.0	Final version	TEK	24/02/2022

Executive Summary

The Deliverable D3.2 is a document of AI-PROFICIENT project delivered in the context of WP3 (Platform AI analytics and decision-making support), and more precisely T3.2 (Predictive AI for process quality assurance). The deliverable provides a summary of the advances made during the project related to the development of AI analytics for the identification of potential process incidents and failures in their early stages. The task is based on the existing knowledge of the processes and on empirical data for the development of ML models that allow to help the operator in the exploration of the data and offer learning capabilities.

D3.2 provides a brief description of how AI technologies are a relevant aspect to ensure the quality of the process and consequently the quality of the product in manufacturing processes. However, the deliverable is mainly focused on describing the algorithms for AI predictive analytics for specific use cases. This analysis has been based, as mentioned above, on the combination of specific knowledge of the processes, the historical data gathered from the monitoring of the processes and the use of ML algorithms.

1 Introduction

The objective of this deliverable is to collect and present the contributions that have been developed related to the development of models for the early identification of potential incidents and failures that can have an impact on the quality of the process and the product. To deploy such AI models capable of predicting the real-world anomalies and faults at the process level, this task will rely on the process-specific knowledge, as well as the empirical data from the industrial production process in the context of AI-PROFICIENT project.

This deliverable firstly explains how AI technologies can contribute to ensure process and product quality within a manufacturing process, and finally it focuses on describing the use cases, the specific technologies, and ML algorithms developed during execution within the project framework.

These technologies will provide the necessary functionalities to AI-PROFICIENT to provide the S_PRED service detailed in the deliverable D1.5:

Table 1: S_PRED service description.

Service ID	S_PRED
Service input and dependency on other services:	This service requires the processed sensor reading produced in by the pre-processing service together with the KPIs produced by the diagnostic and anomaly detection service. The computation of quality and its forthcoming development could be involved in the UCs.
Service output:	The aim of this service is to produce process quality indicators (KPIs) that reflect the goodness of the manufacturing process and predict how these indicators will evolve in the near future. This way, it allows operators to carry out maintenance actions that will prevent the loss of quality when it is foreseen, that is the possible evolution of those KPIs.
High level service description:	This service is aimed at watching over the quality of the production. As such, it needs to verify the assets are in good health conditions, and check that the various quality related measurements taken during the process are also under tolerances. Based on the recent readings, it will also provide estimations of how quality will evolve in the future to foresee possible losses of quality and act accordingly before quality drops.

These two services are intended to cover the _ PRED requirements identified and detailed in the deliverable D1.4 as result of T1.4. showed in the following table:

Table 2: Functionalities to be provided by the AI-PROFICIENT project (from D1.4).

AI-PROFICIENT Functionalities	***ID***
Monitoring	_MON
Diagnostic and anomaly detection	_DIA
Health state evaluation	_HEA
Component prognostics	_PRO
Hybrid models of production processes and digital twins	_HYB
Predictive Production quality assurance	_PRE
Root-cause identification	_ROO
Early anomaly detection	_EAR
Opportunistic maintenance decision-making	_OPP
Generative holistic optimization	_GEN
Future scenario based Lifelong self-learning system	_LSL
Human feedback	_HUM
Explainable and transparent decision making	_ETD

During the preparation of the D1.3, where the different Use Cases (UCs) were explained and a first sketch to their potential final solution was presented, five UCs were potential candidates for the development of AI technologies for the detection of failures and anomalies early to ensure the quality of both the process and the project, as shown in the following Table 1.

Table 3: Original excerpt of expected partners involvement in T3.2 for each use case (from D1.3).

WP/Task	CONTI-2	CONTI-3	CONTI-5	CONTI-7	CONTI-10	INEOS-1	INEOS-2	INEOS-3
WP3– Platform AI analytics and decision-making support								
3.2 Predictive AI for process quality assurance			TEK, INOS	INOS, UL	TEK	TEK, IBE, INEOS		INOS

However, CONTI-5 could not be used finally to validate these technologies because the installation of the camaras has been delayed due to problems with the distribution of the microchips. In the case of CONTI-10 use case, the proposed technologies and solution oriented to quality assurance required the segmentation, characterization, and pre-processing of the entire process stage by stage, considering the quality indicators of each one of them. This task could not be tackled due to the complexity of the process and for this reason, a partial dataset has been used, corresponding to the initial extrusion process (from CONTI-2 use case), to validate the proposed technologies.

Additional work has been done and other use case has been developed by TEK related to the quality prediction in additive manufacturing, with the aim of predicting the quality status of the final part. Additive manufacturing provides many advantages in the manufacture of complex parts and one of its current challenges is to avoid defects in the manufactured parts. This use case aims to detect porosity using artificial intelligence methods. Porosity, as explained below, is a defect that affects the mechanical performance of parts and is currently only detected by subsequent inspection, which leads to additional manufacturing costs.

In the following sections, a detailed description of the main _PRED components can be found, as well as their use in the different use cases.

2 AI-Technologies for process quality assurance and product quality assurance

Manufacturing converts raw material inputs into finished product outputs and value-added services through the coordination of relevant manufacturing facilities, resources, and activities. Some of the most promising applications that can be implemented during the manufacturing process include applications to enable smart design, smart planning, materials distribution and tracking, manufacturing process monitoring, quality control, and smart equipment maintenance (represented in Figure 1). Manufacturing process monitoring and production quality control are key and important applications to be covered within the AI-PROFICIENT project.

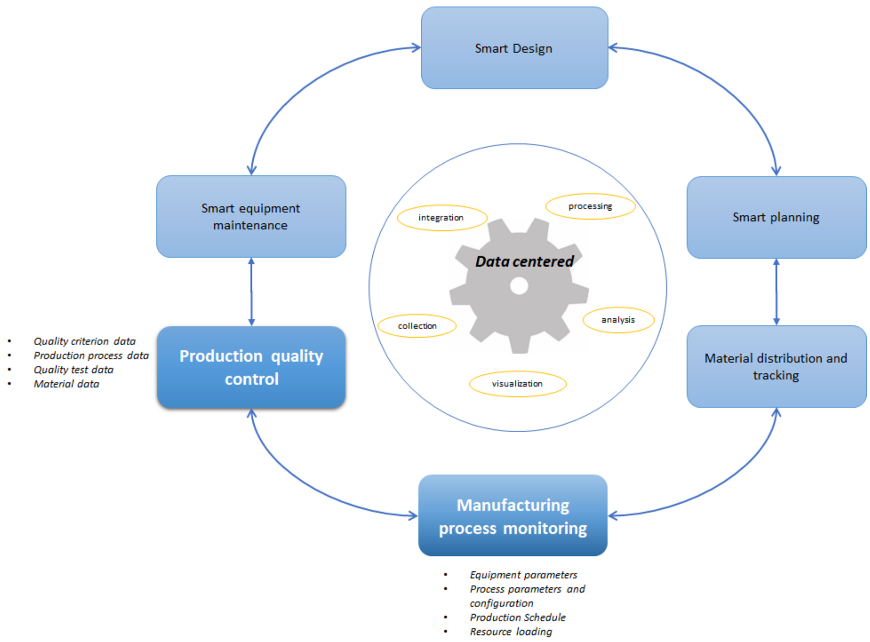


Figure 1. Data-centred manufacturing process

2.1 Process monitoring

The manufacturing process can be affected by multiple factors. These factors can affect the process and influence changes with respect to the quality of the product. In addition, they can also interact with each other and that is why it is important to monitor the different stages and/or steps of the process in real time. However, it is often difficult to trace which factors affect manufacturing processes.

Fortunately, historical data provides effective support for monitoring manufacturing processes. And with the use and predictive power of data analysis techniques, the most appropriate range could be prescribed for each factor. Statistical techniques and machine learning algorithms can be used for correlation analysis and cause-effects identification between factors and the measurable events or parameters, extracted from the pre-processing of the most influencing factors.

Once the influence of the factors and their appropriate ranges of operation are known, the problem can be marked, and alerts and recommendations can be sent to the operators, for early detection of anomalies, so that they make the appropriate adjustments. This can ensure greater compliance in the manufacturing process because it will improve the quality of the process and consequently the quality of the product. Before production anomalies occur, anomalous events often reveal certain patterns that can be captured by the variety of data (e.g., process parameters) in time series.

Since such data is mostly time dependent, static models cannot process it effectively. Furthermore, the large amount of data cannot be processed with traditional data analysis methods, which are computationally intractable. By synthesizing the timing and causality factors, an early warning model of production anomalies can be established based on relevant ML algorithms, such as decision trees, neural networks, or others. Extracting patterns of abnormal event characteristics and trends in time series, it is possible to predict, in advance, whether production anomalies will occur ahead of time. With greater flexibility and less computing time, big data analytics can handle data from multiple sources. Taking balanced use into consideration, manufacturing processes can be dynamically adjusted based on big data analytics.

2.2 Quality assurance

Today, various data-driven quality control techniques are being developed for smart manufacturing. Many sensors are used to acquire data from the machine and the process (e.g., speed, pressure, temperature, time, etc.); and machine vision applications are employed to collect product quality data such as geometric parameters (e.g., thickness, length, roughness, etc.), tolerance parameters, etc.

Data analytics are used then to generate early warning and alerts of quality defects and make a rapid diagnosis of the root causes. Based on historical data and process condition data collected from the machines and their operating environment, quality condition classification and prediction can be used to predict whether and how certain conditions are related to quality defects. Statistics and Machine Learning algorithms can be used to analyse process parameter data to identify the most influential parameters and their appropriate range with the goal of improving both process and product quality. As a result, the process factors or parameters that result in poor quality can be controlled and optimized.

In addition, these data analytics equip manufacturing companies with a particular type of data-based reasoning ability. The lessons learned from one quality control use case can be transferred to another to prevent a recurrence of similar problems in the future. As a result, quality management can be integrated into every step or phase of the manufacturing process, from raw materials to the finished product.

The following section describes the use cases in which data analysis technologies have been developed to predict and improve the quality of processes and the product.

3 AI-based predictive process quality assurance Demonstration scenarios

3.1 Evaluation, management, and improvement of data quality for Extrusion Process (CONTI-2)

3.1.1 Use Case description

To achieve the necessary information throughout the different processes, useful data streams are obtained to provide the Artificial Intelligence and Big data algorithms. However, strategic decision-making based on these algorithms may not be successful if they have been developed based on inadequate low-quality data. 'DQTS' package has been implemented that contains a set of techniques and tools that allow monitoring and improving the quality of the information to measure a improve Data Quality (DQ) in streaming time series. These techniques allow the early detection of problems that arise in relation to the quality of the data collected. The package has been developed by TEKNIKER, it is an open source, and it has been used to analyse and improve the historical data collecting from Continental UC 2.

3.1.2 Solution

Next figure illustrates the process and flow to improve the quality of the data. The first step after accessing the historical data is the identification of gaussian variables to assign weights to Consistency, Typicality and Moderation metrics. (A). Next, the existence of the Reference Data Set, Range Data Set, Maximum Time Difference and Unit of Time is checked (B). The third step is the computation of DQ metrics using the DQ function in the complete set or by moving windows (C). If all the metrics reach the highest quality, it is concluded that the analysis set is correct. If the quality is not 1, a list of the metrics is returned in the order in which they should be treated. Next, using the deepDQ function, the problems with the first metric in the list are analyzed in depth (D). Finally, the decision is made to rectify the data set. In that case, using the handleDQ function the metric is corrected until it reaches a value of 1 (D) and the metrics are recomputed in step C. If data are not modified in relation to this metric, they are removed from the list and if there were more items left in the list, it returns to step D using the next metric in the list. After the complete inspection of the list is done, the final modified data set is available.

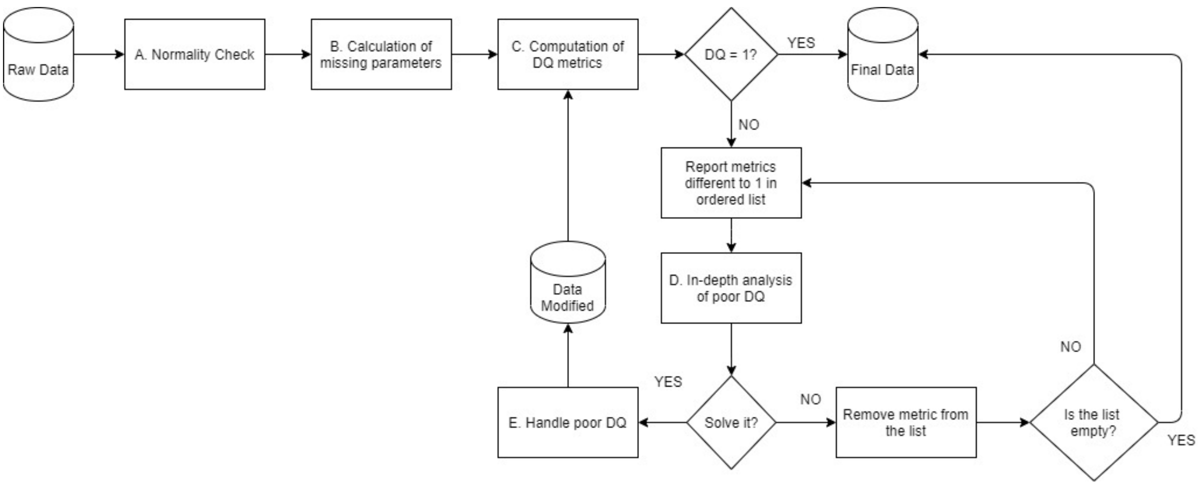


Figure 2. Flow for the detection, inspection, and resolution of poor DQ problems.

R library contains four main functions, and it is available in GitHub to be used by the entire R community:

- The DQ function performs the three steps described above. First, it checks the normality of the variables and the availability of the necessary parameters in the input arguments. If unavailable, they are estimated. Finally, the values of the DQ metrics are calculated and this function allows two ways to do so, the overall one and by windows in three different ways.

  Data Quality metrics are functions to associate a given time series with numerical scores to quantify the quality of that data. These values range from 0 to 1 where 0 is the worst quality and 1 is the higher quality. The global value of quality is a combination of all those 11 metric values:

  - Formats: Proportion of variables with the correct format (integer, date, categorical, ...).
  - Names: Proportion of variables well labelled.
  - Time Uniqueness: Proportion of unique values in the time or date variable.
  - Timeliness: Proportion of values captured in the appropriate time interval, that is, without exceeding the allowed waiting time.
  - Range: Proportion of values within the lower and upper bands that can be provided or simulated.
  - Consistency: Proportion of values within the 80% confidence interval.
  - Typicality: Proportion of values within the 95% confidence interval.
  - Moderation: Proportion of values within the 99% confidence interval.
  - Completeness: Proportion of non-missing values.
  - Completeness by variables: Proportion of non-missing variables.
  - Completeness by observations: Proportion of non-missing observations.

- The deepDQ function takes the data, the name of the metric to inspect, and the parameters required for that metric as inputs. This function returns precise information regarding the data failures in the selected metric.
- The handleDQ function estimates solutions to faults found in the data for the metric introduced as an argument. It returns the data set with the necessary changes for that metric to achieve the highest quality score.
- The plotDQ function allows visualisation of the quality of the data. The output of the DQ function is introduced as an argument. In the case that quality has been calculated in the complete data set, the plotDQ function shows a bar graph where each bar indicates the numerical value of each of the metrics with magnitudes between 0 and 1. On the other hand, if the quality of the data has been computed by windows, a scatterplot is displayed with as many lines as metrics have been calculated and the time evolution of the metrics is shown.

3.1.3 Application and validation

Information of the extrusion process that is analyzed in CONTI-2 has been measured during the last 3 years by sensors placed across the process. The data has been continuously collected and quality analysis of the data has been performed from the ‘DQTS’ package described in the previous section.

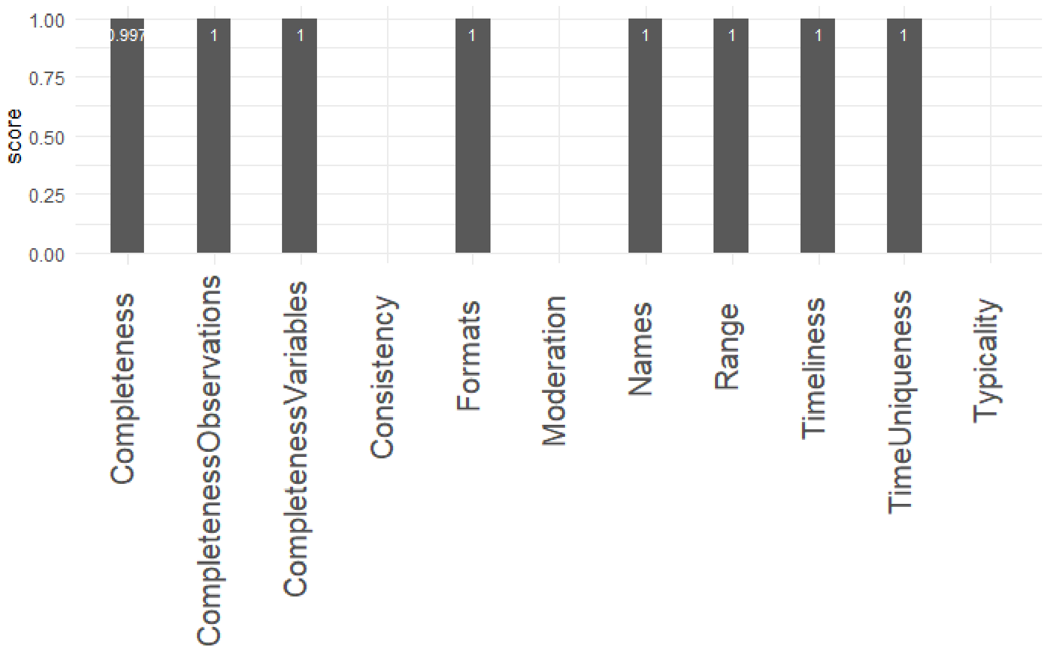


Figure 3. Quality metrics.

After the quality analysis, the global quality indicator obtained from the dataset is about 0.999. This value is very high and implies that the dataset is consistent and of good quality. It contains no major problems, shortcomings or errors that could lead to poor conclusions from data analysis techniques. There are no temporality problems in the dataset because the 99.7% of the observations are completed as can be seen in Figure 3. It is assumed that the dataset is complete and with high quality.

However, the information acquired from these sensors have not been measured in the same frequency. Some of them, such as the compounds and recipes used in the process have been measured in every change and the temperature in the extruders have been measured when some defined temperature change happens, others, such as the pressure and the speed in the extruders have been measured every second. Table 4 shows the frequencies and units of some of the measured signals. 'DQTS' package has been also used: firstly, to set the frequency to 1 second, which is the highest frequency in the original raw data, and a reasonable frequency because of the high variation of the screw speed when the extruders are in usage; secondly, for those signals with lower frequency in the original raw data, unmeasured time observations have been considered missing values; and finally, the missing values in the data have been inputted using Last Observation Carried Forward (LOCF) method.

Table 4: Summary of some of the measured signals.

ID	Unit of measure	Frequency
EX EX2 Compound Name Setpoint	No unit	Each change
EX EX2 Temperature Feeding Zone Actual	ºC	Each 1ºC change
EX EX2 Pressure Mass Screw2 Actual	bar	Each second
EX EX2 Speed Screw Actual	rpm	Each second

3.2 Extrusion process quality assurance (CONTI-2)

3.2.1 Use Case description

During the extrusion process, one of the first stages of the tire production line, specific deviations occur that mean that the product, in this case the material, is not of the desired quality. The extrusion stage is where the different elements or materials are mixed (recipes) and that will make up the final product (rubber). It is an important stage because it results in a perfect sheet that will be used to make the tread and the sidewalls of the tire. However, it is not a continuous process.

The need to produce different types of recipes, as well as other programmed or unscheduled repairs and replacements, led to successive interruptions and restarts of the production line, which are unavoidable, and they negatively influenced the quality of the tire tread.

Because of production stops, when restarting, it is necessary to bring the production line to the optimal condition of productive performance, for which adjustments are made (now, manual control of set points). Until this point of production is reached, the tread that is produced tends to be of low quality and it is sent back to the extruders as it is of no use. The duration of the setup process determines the amount of rework that is created and returned to the extruders (also known as reintroduction) so as not to waste raw materials.

In this use case and for the sake of extrusion quality insurance, IA technologies are enablers for the development of various solutions based on statistical techniques, prediction, and optimization algorithms to guarantee that the material obtained from extrusion is of good quality. In this sense, two solutions or approaches have been developed during AI-PROFICIENT project:

1. Solution 1: Improving and optimizing the quality of the extrusion (section 3.2.2). The development of a surrogate model that will provide the operator with the best configuration so that the extrusion is optimal and minimizes the amount of 'rework' generated. That is, minimize the low quality of the material generated during the extrusion stage.

2. Solution 2: Quality analysis tool (section 3.2.3). A support tool for the analysis and automatic quality assurance that will help the quality team in making decisions about the extrusion stage to improve the quality of the process and as consequence the quality of the material generated.

Both solutions start from the pre-processing of the historical data set, 3 years of historical data obtained from the continuous acquisition of data (real executions, stops, initializations for maintenance, compounds...). As shown in the figure below, the data has been collected from the process through a set of sensors (temperature, pressure, speed, material configuration, speed configuration, profilometer measurement, profilometer objective…). Part of this pre-processing task has been explained in section 3.1.3, where the 'DQTS' package has been used to analyse the quality of the data and to improve it. Subsequently, an automatic segmentation and characterization of the data has been carried out by calculating a set of statistical and descriptive variables. These variables characterise the process. Then, the measure of the profilometer has been used to detect and label when the extruded material reaches good enough quality standards. And finally, this dataset has been used later by the ML algorithms to find the optimal setups to maximize the quality of the rubber.

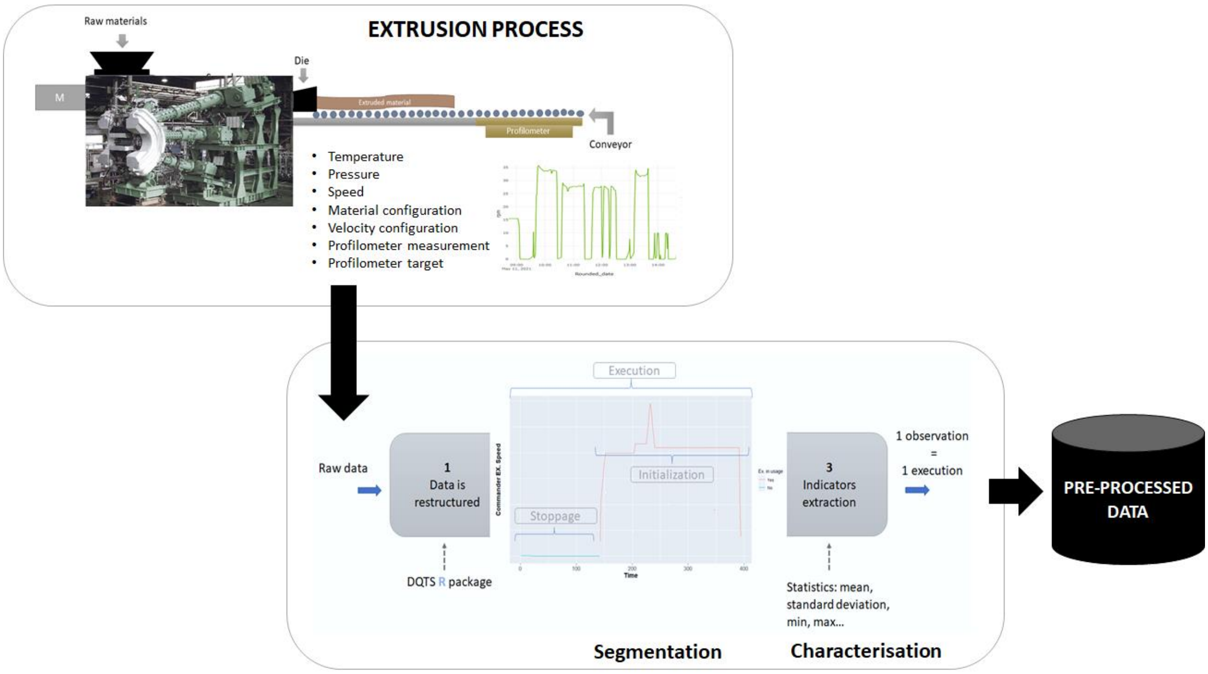


Figure 4. Pre-processing data (workflow)

To deal with this UC, two services to cover Solution 1 and Solution 2, with their functionalities, are being developed in the AI-PROFICIENT project (see

Figure 5) and work is still ongoing on the necessary tasks to complete its deployment in CONTINENTAL's infrastructure.

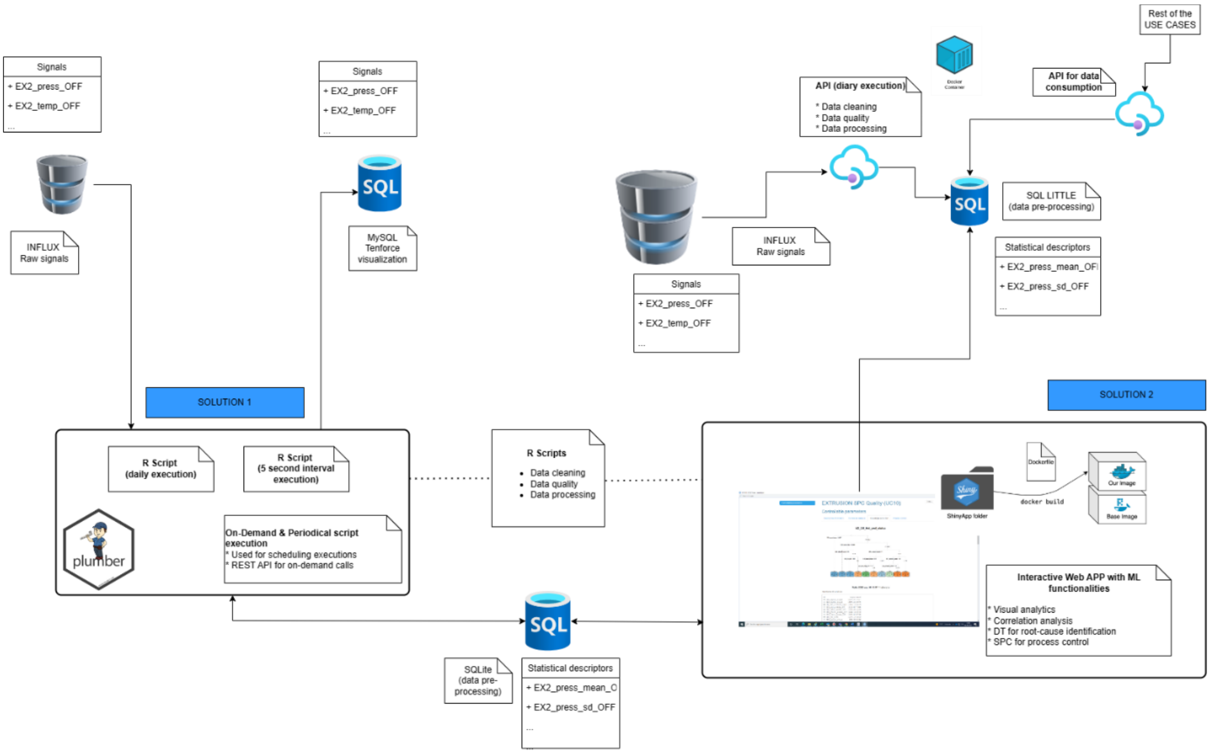


Figure 5. Diagram of the deployment of both solutions in CONTINENTAL's infrastructure

3.2.2 Solution 1: Improving and optimizing the quality of the extrusion

The objective of this solution is to provide the operator with the optimal extrusion setups when the extrusion process is starting. The AI module fed with real-time signals of the extruders and some suggestions are provided to get as fast as possible best quality of extruded materials based on the AI's knowledge.

Optimization of the extrusion process has been done by a population-based optimization algorithm (Wu, Mallipeddi & Suganthan, 2019) on which the objective function is defined by Machine Learning algorithms based on historical data. Population-based optimization algorithms are a kind of optimization solvers that approximates the optimal solution instead of finding the optimal mathematical solution. A population is defined as a set of possible solutions which are evaluated by a fitness function and then transformed by the algorithm with the aim of finding the optimal solution.

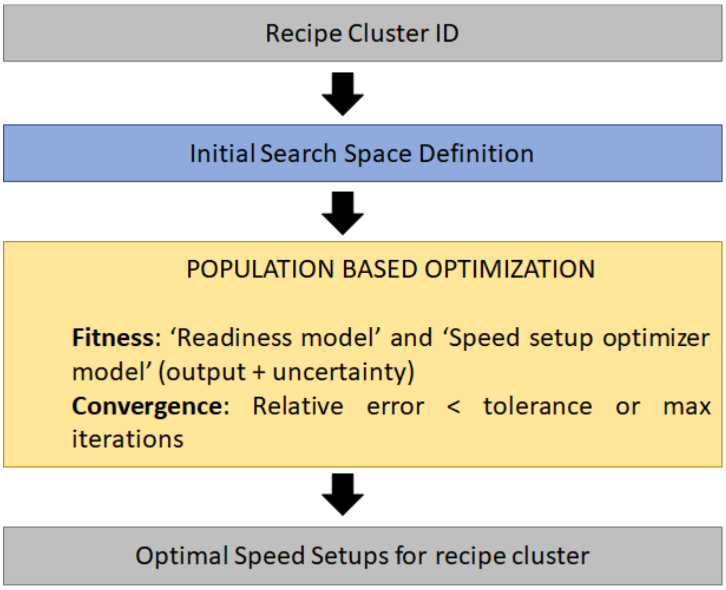


Figure 6. Optimization process approach

Thereby, the objective function, i.e., that is used as the fitness function, is an ensemble of two models.

The first model (called '**Readiness model'**) that has been developed is a *Random Forest Classifier* which predicts the steadiness of the process as either (1) rejected extrusion (material is not good enough and has not been measured in the profilometer), (2) not steady extrusion (material's profilometer value has not became steady), or (3) steady extrusion (material's profilometer value has become steady).

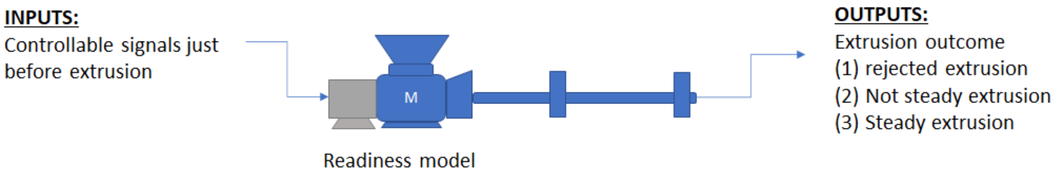


Figure 7. Readiness model. Real-time quality prediction to restrict the input space.

For the cases with a steady prediction label in Random Forest model, a second model (called '**Speed setup optimizer model'**) based on Gradient Boosting Regressor (XGBoost) algorithm, has been trained to predict the error associated with the profilometer's target value as a measure of the quality of the extruded materials. Therefore, each individual (represented by the controllable variables from an extrusion) in the population-based optimization algorithm has been evaluated firstly using the Readiness model.

If the predicted label of the Readiness model is the rejected class or the not steady class, the individual has an associated large fitness. Instead, if the predicted label of the Readiness model is the steady class, the associated fitness of the individual is given by the predicted value of the Speed setup optimizer model. Thus, the problem is a minimization problem.

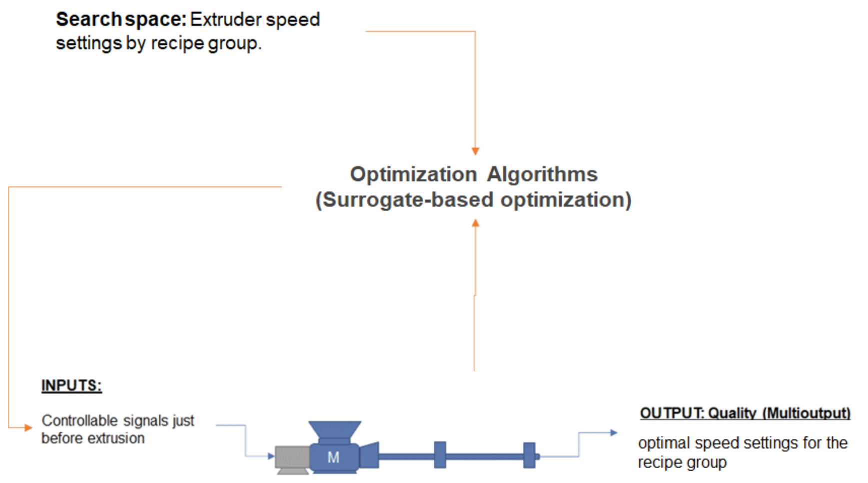


Figure 8. M2.1: Search algorithm (Population based algorithm) for optimization based on surrogate modelling.

3.2.3 Solution 2: Quality analysis tool

The solution proposed here relies on the exploitation of the data from the use of AI technologies to identify and predict the correlation and cause-effect relationship among parameters and quality factors for extrusion process from the results obtained after the pre-processing of the historical dataset explained in section 3.2.1 .

The pre-processing of historical data set and evaluating the quality of signal has been the starting point for this solution. From this dataset, the 'Solution 2: Quality analysis tool' offers the implementation of a software application (a set of functionalities) for the study of correlations and cause-effect relationships between the parameters of the process and the desired quality characteristics.

As it was previously stated, among numerous control parameters, generally influencing factors, some subset of them mainly affects the quality of the final product. This software and its functionalities will provide support for automatic quality analysis and will help the end-users, mainly quality managers, to identify process parameters, focused on those parameters that are controllable and that could be modified by the operator, that affect the quality of the process and the quality of the material obtained after the extrusion process. In the UC2, the quality of the tread is affected by the process stability which is determined from the profilometer measurement. The proposed solution incorporates a set of functionalities to automatically analyse and evaluate the correlation and effects of the extrusion signals (controllable parameters) with the stability of the process and the profilometer measure. In addition, it allows to perform a statistical control of the process to predict and compare its current behaviour against its nominal behaviour.

These functionalities, statistical and machine learning algorithms, will support a set of phases for the automatic analysis of the pre-processed historical data: (1) the correlation analysis to identify the relationship between parameters, (2) the cause-effect analysis to identify the relationships that lead to a state of quality or to other, and (3) the early identification of anomalies or deviations in the process in relation to the nominal behaviour (when quality deviations do not occur).

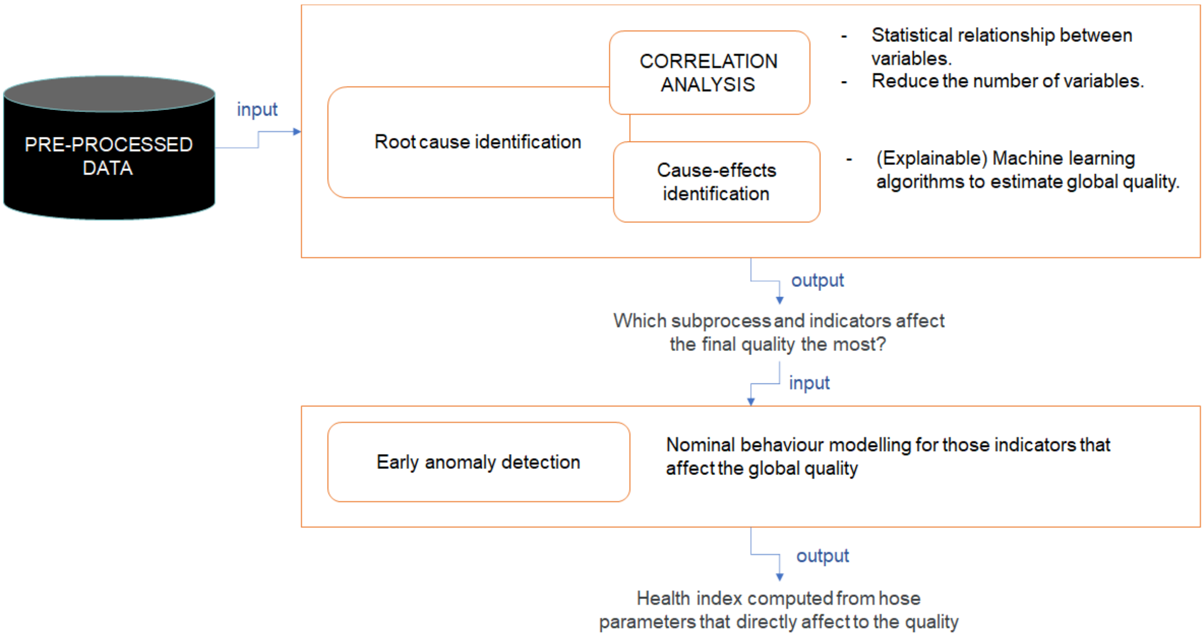


Figure 9. Phases for the development of the functionalities for correlation analysis, cause effect identification and early anomaly detection.

The types of algorithms, statistical and Machine Learning, are numerous and within this UC some techniques have been examined and finally implemented, to extract the most influential parameters that affect the characteristics of the product or process, from simple algorithms, interpretable and easily transferable to rules that can be later easily interpretable for the end-user.

The software application (interactive web application) and its functionalities for the correlation analysis based on statistical techniques, cause-effect identification through ML, and early anomaly detection for quality assurance have been developed in R code and they are being deployed as a service in CONTINENTAL's infrastructure (see

Figure 5). So, the service will be a support for the decision-making of the end-users (quality managers or operators in charge of the quality of processes and products) who are not analysts but final consumers of the technology.

3.2.4 Application and validation of solution 1: Improving and optimizing the quality of the extrusion

The validation of the models that have been fitted and used as objective function in the optimization of the extrusion process has been done if the following way. For the Readiness model, which is a Random Forest Classifier as mentioned above a Stratified Cross Validation method has been used, because of the imbalance in the output variable, and F1 and accuracy metrics have been considered. Several experimentations have been conducted to analyse the variability in the prediction. The Random Forest algorithm has been trained with three data sets: one containing all the signals from the extrusion, a second containing only the uncontrollable variables and a third with the controllable ones. The results from each of them have been significantly similar. Therefore, it was decided to use only the controllable variables because they are variables on which actions can potentially be taken. Results are shown in Figure 10.

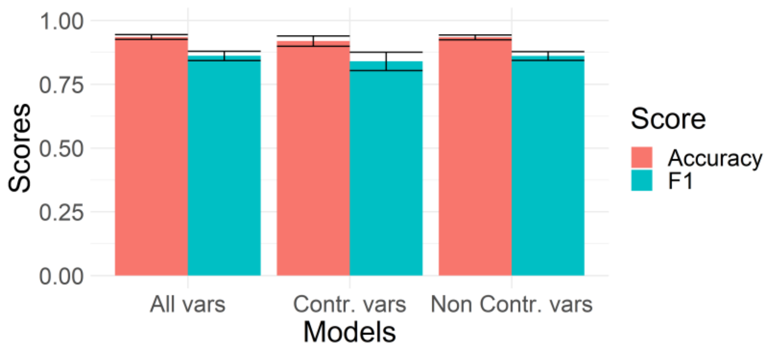


Figure 10. Readiness model results for the three datasets (All variables, Controllable variables only and Non controllable variables only).

Besides, the Random Forest impurity, a common measure used in decision tree algorithms to decide the optimal split from a root node and subsequent splits (Louppe, Wehenkel, Sutera, & Geurts, 2013), has been used to analyse the importance of the features in the solution, which has shown that the most important variables to consider in the steadiness of the process were related to the speed setups used in the restart.

For the validation of the Speed Setup optimizer model, a random train/test split has been conducted, with the 80-20% of the samples and the well-known R squared metric. For the search of the optimal parameters a simple Cross Validation method has been used to optimize the max depth and the number of the estimators on the trees using the training set and three folds.

As the objective of this use case has been the optimization of the extrusion process, a robust quality parameter has been needed which could characterize the final quality of each extrusion. For that, a 5 samples windowing Root Mean Square Error (RMSE) has been computed on which the desired target value and the real value in the profilometer. Then, this RMSE is used to determine whether the real value is steady near the target for the first time. It has been considered as a robust measure after some analysis and it has been used as the output for the fitness function. Then, a final model has been achieved with the optimal parameters set to max depth equal to 2 and the number of estimators equal to 10. The obtained mean R squared value in the training Cross Validation has been 0.51 and the R squared obtained in the test has been 0.41. Besides, when the outliers were discarded these values have been improved to 0.54 and 0.64 respectively.

3.2.5 Application and validation Solution 2: Quality analysis tool for the extrusion

In this solution two main steps are considered including 3 stages represented in Figure 9 and described below in more detail:

- Correlation analysis based on Statistical techniques.

Correlation is a measure of the strength of the relationship between 2 variables. Among the several available correlation statistics, Pearson correlation (abbreviated as “r”) and Spearman rank correlation (abbreviated as “ρ” or rho) are probably most widely used. Their coefficients quantify the strength of a linear (Pearson) or monotonic (Spearman) relationship. A relationship is monotonic when the value of one variable consistently increases (positive correlation) or decreases (negative correlation) as the value of the other variable increases. A linear relationship is a special case of a monotonic relationship, in which the rate of change is constant.

Pearson and Spearman correlation coefficients range from –1 to +1, with absolute values increasingly closer to 1 indicating an increasingly stronger relationship. Various somewhat arbitrary cut-points have been proposed to categorize the strength of the relationship using descriptors like “weak” (e.g., $r < 0.40$), “moderate” (e.g., $r = 0.40$ to 0.69), or “strong” (e.g., $r \geq 0.70$). The interpretation should also take into account the confidence interval of the observed coefficient as an estimate of what the correlation could plausibly be in the population from which the data were sampled.

Spearman correlation is recommended when either at least one variable is not normally distributed, or the relationship between the variables is not linear, or there are relevant outliers. Spearman correlation is based on the ranks of the values of each variable instead of their actual values, and it can basically be used for all data that can be ranked, including ordinal and nonnormally distributed continuous data.

These techniques have been implemented as functionalities for the ‘Solution 2: Quality analysis tool’, to compute correlation coefficients, and they are being deployed as a service. From this correlation analysis and its visualization, the end-user of the service will be able to have some first indications of which controllable parameters of the extrusion process can be related and affect its stability and consequently to the quality of the thread.

For UC2, from the pre-processed data[1] (explained earlier in section 3.2.1.), direct correlations can be calculated and observed using these functionalities as shown in the following figure. It represents the visualization of the calculation of Spearman’s correlation coefficient from the data and it shows that there are significant correlations between the type of extrusion determined based on the profilometer measure: (1) – rejected, (2) - not steady, (3) – steady, and Extruder 2. It is necessary to highlighted that this label is directly related to the quality of the thread.

[1] data quality analysis, segmentation of the signals for each extrusion, characterization of the signals by means of statistical descriptors and having labelled the extrusion type as (1) – rejected, (2) - not steady, (3) – steady from the measure of the profilometer

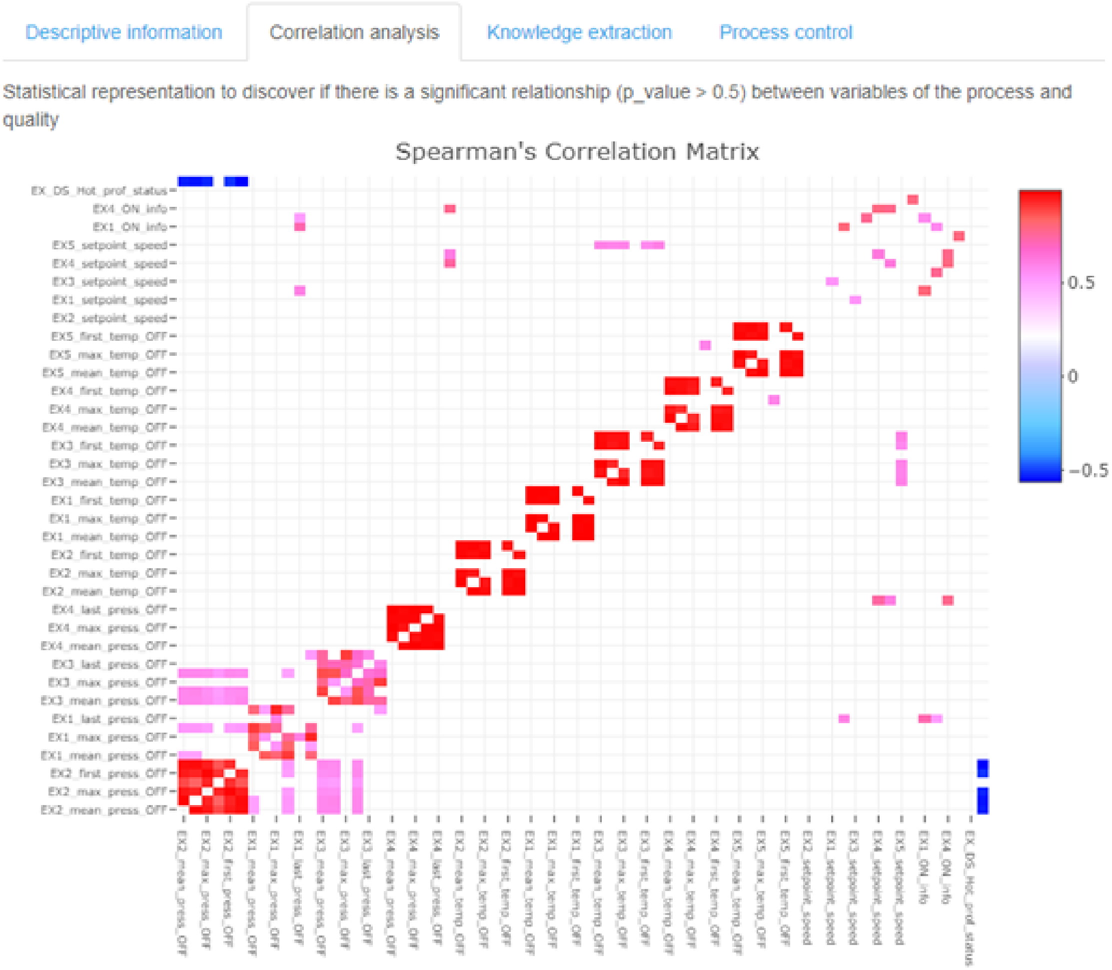


Figure 11. Correlation analysis.

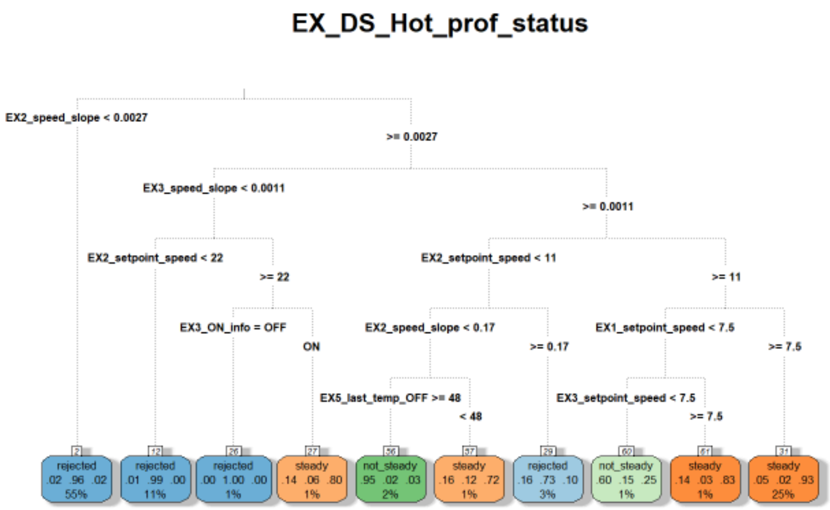
- Cause-effect identification.

Decision tree algorithms can be used to represent decision making visually and explicitly. The resulting model can be an input for decision making. They are among the most popular machine learning algorithms given their intelligibility and simplicity, especially when pruning technique is used to reduce the size of decision trees by removing parts of the tree that do not provide power to predict. They are easy to understand and interpret and can be display graphically in a way that is easy for non-experts or analysts to interpret them. In addition, they work well with large data sets, and large amounts of data can be analyzed using standard computing resources in a reasonable amount of time.

A Recursive Partitioning and Regression Tree (RPART) has been implemented- rpart: Recursive Partitioning and Regression Trees (r-project.org). The rpart programs build classification or regression models of a very general structure using a two-stage procedure; the resulting models can be represented as binary trees. The tree is built by the following process: first the single variable is found which best splits the data into two groups (‘best’ will be defined later). The data is separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size (5 for this data) or until no improvement can be made. The resultant model is, with a certainty, too complex, and the question arises as it does with all stepwise procedures of when to stop. The second stage of the procedure consists of using cross-validation to trim back the full tree.

From the visualization of the model trained and generated by the RPART algorithm, the end-user will be able to obtain some insights about which process parameters (and from which extruder) are affecting the stability of the extrusion process and consequently the quality of the material at the end of the extrusion. It is feasible to plot the decision tree or even the rule generated and the importance of each of the variables used for the decision tree in order to make the prediction (Figure 12).

Descriptive information | Correlation analysis | Knowledge extraction | Process control



Importance of variables:

```
##                   importance
## EX2_speed_slope     3257.0747027
## EX3_speed_slope     1497.2014874
## EX_INF_restart_type 1352.1048315
## EX2_min_press_OFF   1313.4077944
## EX2_mean_press_OFF  1303.6412285
## EX2_first_press_OFF 1264.1843019
## EX2_last_press_OFF  1263.4029766
## EX3_ON_info         1207.1072889
## EX1_speed_slope      932.3361096
## EX1_ON_info          889.6412701
## EX1_last_press_OFF   730.5893310
## EX2_setpoint_speed   566.5945228
## EX5_last_temp_OFF    340.6049397
## EX5_mean_temp_OFF    328.7455877
## EX5_min_temp_OFF     313.2572716
## EX5_max_temp_OFF     297.4983501
## EX5_first_temp_OFF   275.6434298
## EX1_setpoint_speed   146.3622397
## EX3_setpoint_speed    98.6612356
## EX4_last_temp_OFF     79.9147212
## EX1_max_press_OFF     75.9970703
## EX3_last_press_OFF    70.0150826
## EX4_mean_temp_OFF     37.1256368
## EX4_max_temp_OFF      16.9959984
## EX4_min_temp_OFF      14.4780727
## EX2_sd_temp_OFF        0.3949498
```

Figure 12. Cause-effect identification and importance of variables through RPART decision tree algorithm

As we can see in the generated tree, there are three rules that are generated to determine that an extrusion is to be rejected (Figure 13):

1. when EX2_speed_slope < 0.0027
2. when EX2_speed_slope >= 0.0027 & EX3_speed_slope < 0.0011 & EX2_setpoint_speed < 22
3. when EX2_speed_slope >= 0.0027 & EX3_speed_slope < 0.0011 & EX2_setpoint_speed >= 22 & EX3_ON_info is OFF

```
##  EX_DS_Hot_prof_status  not_ reje stea


##             not_steady [ .60  .15  .25] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed >= 11 & EX1_setpoint_speed <  7.5
              & EX3_setpoint_speed <  7.5
##             not_steady [ .95  .02  .03] when EX2_speed_slope is 0.0027 to 0.1712 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed <  11                                               & EX5_last_t
emp_OFF >= 48
##               rejected [ .16  .73  .10] when EX2_speed_slope >=          0.1712 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed <  11

##               rejected [ .02  .96  .02] when EX2_speed_slope <  0.0027


##               rejected [ .01  .99  .00] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e <  0.0011 & EX2_setpoint_speed <  22

##               rejected [ .00 1.00  .00] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e <  0.0011 & EX2_setpoint_speed >= 22                            & EX3_ON_info is OFF

##                 steady [ .16  .12  .72] when EX2_speed_slope is 0.0027 to 0.1712 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed <  11                                               & EX5_last_t
emp_OFF <  48
##                 steady [ .14  .06  .80] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e <  0.0011 & EX2_setpoint_speed >= 22                            & EX3_ON_info is  ON

##                 steady [ .14  .03  .83] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed >= 11 & EX1_setpoint_speed <  7.5
              & EX3_setpoint_speed >= 7.5
##                 steady [ .05  .02  .93] when EX2_speed_slope >=          0.0027 & EX3_speed_slop
e >= 0.0011 & EX2_setpoint_speed >= 11 & EX1_setpoint_speed >= 7.5
```

Figure 13. Rules computed by RPART

Its main disadvantage is that the result can be ‘weak’, i.e., be submitted to variability, because the results can vary depending on the data used to train the model. Nevertheless, as the historical data set will increase daily as new extrusion processes are conducted, the decision tree will refine its knowledge over time. Furthermore, Random Forest ensemble learning method has been also implemented to overcome the problem of RPART.

Both models can be used to compare results and make better decision-making. As can be seen in the next figure, the decision tree shows that certain conditions in the extrusion signals can lead to a rejection of the extruded material. However, the final decisions and actions will depend on the team and quality managers.

- Early anomaly detection to guarantee the quality of the process and material.

The statistical tool called "Statistical Process Control" (SPC) has been implemented and it is presented below, which will help the end-user responsible for quality to maintain an attitude of continuous improvement for the extrusion process, to minimize the rework generated, and to compare production with respect to the required specifications.

It is a basic tool to study the variation and use the information obtained by taking data from the process and its critical characteristics and treat them appropriately. This information allows the end-user to monitor with respect to its “natural” variability to detected abnormal deviation. The knowledge and use of this tool will allow to evaluate and maintain the stability of a process. Likewise, in a stable manufacturing process, the calculation of the process capacity will contribute to achieving the required quality level.

Again, to develop and validate these statistical technologies, the starting point is the pre-processed dataset from the UC2. Statistical control has been performed using the Shewhart chart, which has been selected for the calculation and visualization of the statistical control because it does not assume or require a normal distribution in the data for the calculation of the limits. This makes it a very robust technique, as demonstrated in Wheeler's work using real data with non-normal distributions (Wheeler, 2009). For the calculation of the control limits, the data of the stabilised extrusions (well-finished extrusions with good quality of tread) are calculated as follows:

$$UCL = \bar{x} - 3MR$$

$$LCL = \bar{x} + 3MR$$

$$MR = \frac{\sum_{i=2}^{m} |x_i - x_{i-1}|}{m-1}$$

Subsequently, new individuals are traced from the extrusions made, and those that are within the control limits indicates that everything is working as expected. Any variation within the control limits is probably due to a common cause: the natural variation that is expected as part of the process. If the data is outside the control limits, this indicates that an assignable cause is likely the source of the product variation, and something must be changed within the process to fix the problem before defects occur.

The implementation of these statistical techniques can help to reduce the scrap, react to process drift, and make decisions. Now, the pre-processing of the data obtained from the extrusion process is intended to be executed daily and for that reason this statistical control cannot be used in real time because the required pre-processing cannot be carried out in real time. However, if in a near future the preprocessing of the data can be computed in real time, these techniques would allow to react almost instantly to such changes in the process and decisions could be made in real time on the shop floor.

The following figure shows the use and validation of this statistical control technique using the data from UC2 and it is seemed that certain extrusions produce a variation (almost all of them correspond with extrusions that were rejected) that is not within the normal behaviour of the process and therefore they can be identified as outliers that produce a bad quality output in the extruded material. Early detection of these situations will allow the team to decide whether to stop the process or to take certain actions to mitigate the deviation from normality that can lead to poor quality in the final extruded material.

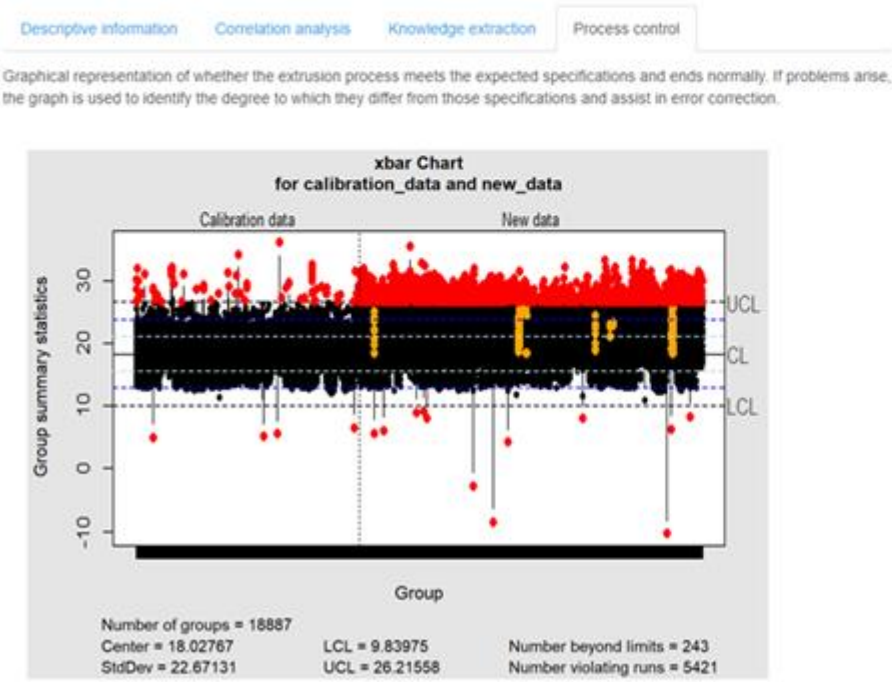


Figure 14. Cause-effect identification

3.3 Porosity prediction in wire Laser Metal Deposition Processes (TEK-LMD-ADDI)

This use case has already been presented in Deliverable D2.5: Local automated control for quality assurance. However, this task has focused on the resolution of another problem which is also related to the same additive manufacturing process: the creation of porosities.

3.3.1 Use Case description

Additive manufacturing (AM) is an interesting solution for many companies that produce geometrically complex parts. This process consists in the deposition of material layer by layer following a sliced CAD geometry. It brings several benefits to manufacturing capabilities, such as design freedom, reduced material waste, and short-run customization.

However, one of the current challenges faced by users of the process, mainly in wire laser metal deposition (wLMD), is to avoid defects in the manufactured part (Figure 15), especially the porosity. This defect is caused by the extreme conditions and metallurgical transformations of the process. And not only does it directly affect the mechanical performance of the parts, especially the fatigue properties, but it also means an increase in costs due to the inspection tasks to which the manufactured parts must be subjected.

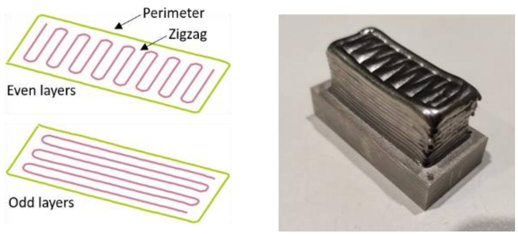


Figure 15. Sample of wLDM manufacturing part.

The monitoring of AM processes provides large amounts of data containing high-frequency information of the status of both the manufacturing process and the built parts. Different types of signals are acquired at different stages of the process, so that time-referenced and spatially referenced signals can be distinguished. A pre-processing and data fusion step is necessary to create the most exploitable data set to properly train machine learning algorithms.

As represented in the first step of Figure 16, the position over time is recorded, which defines the material deposition trajectory performed by the robot. The signal is composed of the three univariate variables X, Y, Z and the Time. Both variables serve as the reference to merge all data acquired, as they include the relation between space and time information. During the material deposition process, two types of signals are acquired with reference to the process time. On the one hand, the signals that control the laser power and the wire feed speed are recorded by the PLC. On the other hand, images of the area where the material is being added are acquired by coaxial monitoring of the melt pool. The uniformity of the growth throughout the manufacturing is acquired by geometric scanning. As the manufacturing is done layer by layer, after the deposition of each layer, the resulting surface geometry is scanned. The obtained 3D point cloud representing the surface is spatially referenced to the trajectory and the manufactured part. The porosity is measured offline when the manufacturing of each part is finished. The data provided by the 3D Computerized Tomography (CT) is the location of the center of each pore and its volume and shape, so it is spatially referenced to the part geometry. That information is used to label the recorded data sets and discern the pore observations.

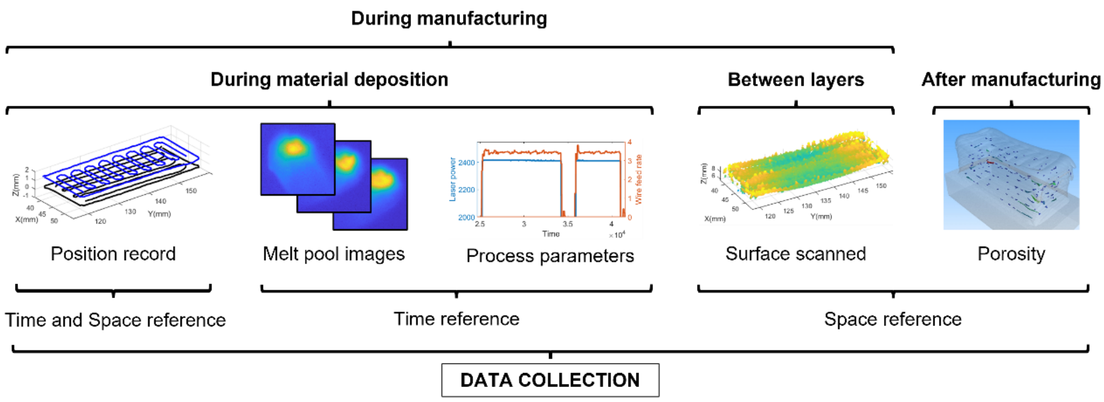


Figure 16. Data collection.

The data is obtained from the manufacturing of three 3D geometry parts. These are formed by 12 rectangular layers with dimensions of 16 by 40 mm approximately. Each layer is manufactured using a predefined material deposition trajectory consisting of a perimeter strategy and a zigzag for the interior filling. The time spent building them ranges between 18.2 and 18.8 seconds.

A set of optimal process parameters has been used, adjusted in a previous experimental process. The commands that control these parameters are constant throughout the manufacturing process, the main ones being laser power, wire feed speed and movement velocity. Regarding the trajectory, an optimal distance between beads that provides uniform growth is determined. Finally, from the experimental tests of the selected set of parameters, the theoretical growth of the layers is estimated.

3.3.2 Solution for improving and optimizing the quality of the porosity

Regarding Artificial Intelligence technologies, data-driven techniques for automated data analysis and decision making are remarkable. Currently, the efforts in additive manufacturing, and particularly in wLMD, are focused on the development of alternatives based on data and AI algorithms to create systems capable of detecting defective parts in real time and subsequently stopping the process, saving effort and economic costs. The solution proposed for this UC is based on Topological Data Analysis and Machine Learning algorithm to predict the defect of porosity for a wLMD process.

- Experimental dataset and pre-processing

Once the signals are acquired, a data processing and fusion method has been implemented to generate the data set (presented in the previous section). The data fusion is based on the trajectory information, which enables independent signals of different typologies to be converted into a spatially and temporally aligned data set. Therefore, the variables extracted from the signal processing can be linked to a moment in the manufacturing process and to a location in the part geometry. Finally, a multi variate signal has been elaborated thanks to the concatenation of all the signals.

Table 5: Summary of the measured signals.

Name	Description	Type
Z distortion	Vertical distortion of the deposited layer	time-series values
Base distortion	Vertical distortion of the base surface. The combination of this data and the vertical movement allows the calculation of the variation of the working distance.	time-series values
Overlap C	Factor indicating the degree of spacing (or NON-overlap) between strands.	time-series values
Vertical Movement	Vertical displacement (Z-axis) of the robot with respect to theoretical height	time-series values

Three parts or pieces (identified as 13, 16 and 18) have been manufactured and the signals were acquired during the process for constituting the data set. Afterwards, a CT analysis was performed to know the characteristics of the porosity in the samples. Once the data was labelled, 3 different data pre-processing have been compared:

1. No pre-processing
2. TDA + Pixelwise features
3. TDA + Colour moment feature extraction

(1) No pre-processing: process parameters captured during the manufacturing process, observation by observation

Data captured directly during the manufacturing process (Figure 17) observation by observation has been used to train the models.

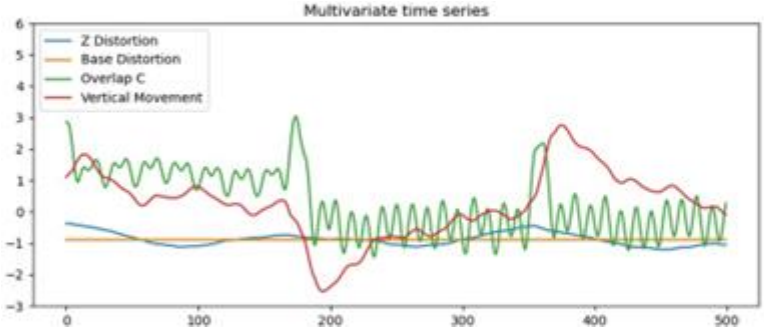


Figure 17. Data capturing from the process (Multivariate Time Series).

(2) Topological data analysis (TDA) and Persistence Images Creation

A rolling window of size 100 has been applied in the multivariate time series skipping 50 values to avoid unnecessary overlaps as can be seen in Figure 18. Notice that the values achieved by Base Distortion variable in the first layer are equal to 0 due to its definition. At each execution, a different multivariate time series of length T = 100 has been obtained and then, TDA has been applied in these time series. The persistence image of each of the available variables has been extracted and saved as jpg format. The images are set to be 20 by 20 pixels dimension.

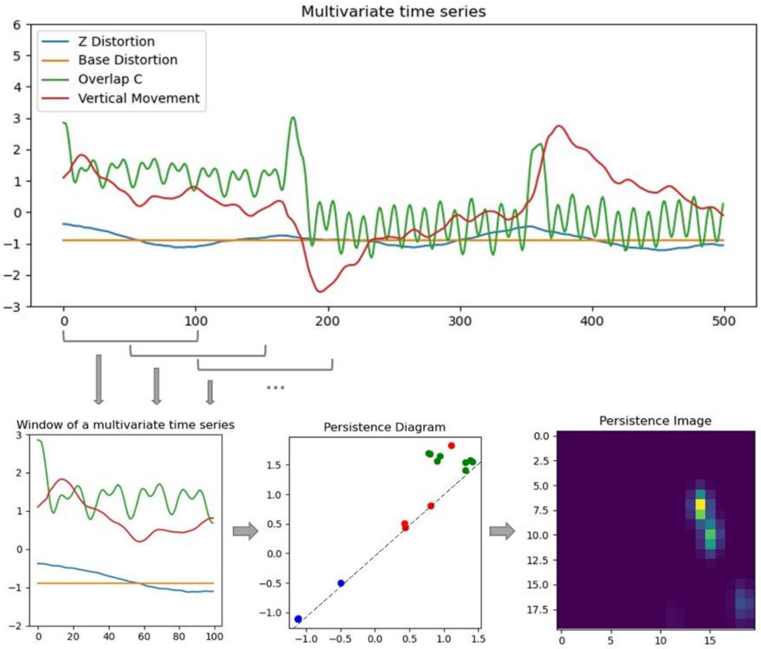


Figure 18. Creation of Persistence Images in Multivariate Time Series.

(3) TDA + Feature extraction from the images

The third compared pre-processing adds a feature extraction technique on the top of TDA images. Common Moments has been the selected colour feature extraction technique because it is one of the simplest and most effective strategy. It has been applied in each persistence image of 20x20 pixels, that is, in each colour value set {p1, p2, ..., p400}.

- Machine Learning algorithms

Five of the most used supervised classification algorithms in the literature has been implemented and compared: the k-Nearest Neighbours (KNN), the Support Vector Machine (SVM), the Decision Tree (DT), the Random Forest (RF) and the Extreme Gradient Boosting Ensemble (XGBoost).

The strategy used to find the most suitable hyperparameters for each algorithm has been the same in all the cases. First, a random search provides the knowledge of the parameter space on which the optimisation should be done. Once that result is obtained, a grid search is used to assess the different combinations of hyperparameters. These searches are done by 3-fold cross validation repeated 30 times and the classification errors are compared to choose the optimal hyperparameters. The 3-fold cross validation randomly divides the entire train set into three subsets, and each subset is tested using the model trained with the other two subsets. The performance is assumed to be the average of the error metrics obtained in each iteration. Other possibilities of number of folds exist but it is decided not to use the classic 10-fold cross validation by agreement with the size of the validation set. In the implementation of this solution, the porosity of a whole part must be detected, so the size of the validation set will be the amount of data recorded in the entire manufacturing process of that part. Using a third of the training data to validate on each iteration is considered more robust than using only a tenth.

- Balancing Train Sets

After splitting the sets into training and validation sets, the decision has been made to balance the classes in the training set. It is a common problem in troubleshooting to have a low number of records of the faulty class. However, in this case, the difference between the number of observations with each label is extremely high as seen in the Figure 19. There is a risk of missing relevant information if subsampling techniques are applied directly on the raw dataset and the resulting dataset may be too small for the model to learn. Similarly, little relevant data can be added unnecessarily with oversampling techniques. For this reason, it has been decided to randomly subsample the majority class.

Part ID	Class 0	Class 1	Proportion of 0	Proportion of 1
13	19087	109	99.43 %	0.57 %
16	18926	144	99.24 %	0.76 %
18	19245	36	99.81 %	0.19 %

Figure 19. Class imbalance

3.3.3 Application and validation in improving and optimizing the quality of the porosity of wLMD process

These algorithms have finally been used to learn from the data sets that have been obtained from the three transformations mentioned above, after the balancing of the data:

(1) **No pre-processing.** From the data captured directly during the manufacturing process observation by observation. In this case, the most accurate classification in this process has been obtained by the Random Forest algorithm.

(2) **Topological data analysis (TDA) and Persistence Images Creation.** The 20x20 values of the persistence images are written as a vector of length 400. Each component of that array was used as an input variable of the classification model. The best fit was reached in both parts by KNN classifier.

(3) TDA + Feature extraction from the images. Colour moments - the most accurate classification was obtained with RF but hey cannot be considered good enough to generate a porosity diagnosis model.

Following the results obtained, the porosity is considered a local fault that is unrelated to the relative position of its occurrence. For that reason, data from the three parts can be mixed without loss of generality. That is, each observation can be considered independent, without considering the values around it or the piece to which it belongs. So, data from parts 13, 16 and 18 are combined and a RF classification is done in the new data set. The confusion matrix of this classification is shown in Figure 20. In that case, the Accuracy value was 0.962 and the Recall metric got 0.853. This result shows that the RF model allows a correct detection of the pores in the manufactured parts. Therefore, it is possible to estimate the existence of pores in the parts and it implies a potential substitute for the classical tomography process.

		Predicted	
		0	1
Actual	0	18175	714
	1	15	87

Figure 20. Confusion matrix from classification of three parts combined- RF results.

There are some lines of activity and research to continue with the work in the future. The first one, as it is expected to have a larger number of monitored and tomographed parts, is to train the classification models with data from several parts and then validate the performance of the fitted models using data from an entire part. At this point, it is important to emphasise the importance of data coming from parts manufactured under the same conditions, since any variation in the configuration of the input parameters could significantly modify the behaviour of some data, adding variability to the data set and avoiding models from capturing those behaviours.

The following figure shows the position of the estimated pores (red points) by the Random Forest model in one part. In this case, 273 pores have been estimated in a part of 12025 points recorded. So, the porosity proportion is 0.023. It is assumed that this part has enough quality, and it is not required to send it to analyse by a tomography to identify defects.

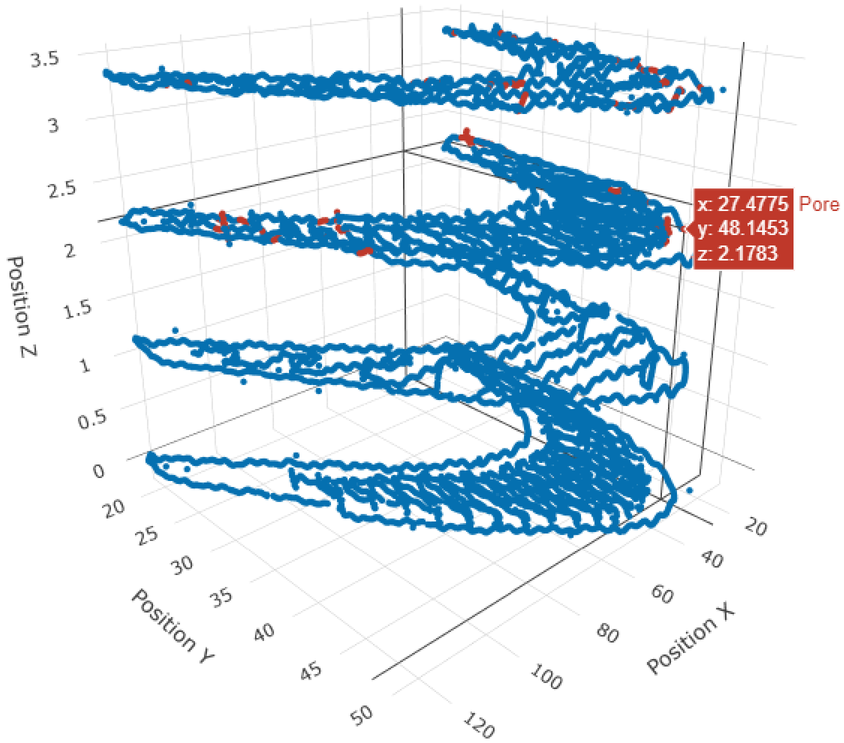


Figure 21. Porosity prediction over the manufacturing time of the part.

During this use case, a service has been developed during the AI_PROFICIENT project. The model for the prediction of pores, based on the Random Forest algorithm, has been deployed in the infrastructure of Tekniker. After each manufacturing of a part, the operator can verify the quality of the part without the need of computed tomography. Automatically, on-demand, an .html report is generated with the predictions made by the model.

4 Conclusions

At present, ensuring the quality of processes and products performs a very important role worldwide. It is the key issue to success for industrial sectors, where it is necessary for the final product to have a quality control, normally, based on stringent specifications. Especially in industrial manufacturing, companies strive to improve their processes and products by ensuring quality and to reduce the costs; identifying and controlling any aspect or situation that may cause the final product to be unacceptable, working on finding innovative and profitable solutions and technology to prevent errors and defects in advance.

At the same time, digital technologies and industrial IoT systems, sensors and devices that collect data from the infrastructure of the factories and companies, improve opportunities for the use of data and AI technologies for data analysis. These artificial intelligence technologies allow the data obtained to be combined and put it in value, from the extraction and transformation of the data and the use of machine learning algorithms.

Thanks to the increasing volume of data held today by the companies, these approaches based on Artificial Intelligence, the information extraction and data analysis make it possible to detect problems automatic and friendly which also provides added technical support to the operator. They will not replace the knowledge of the operator, gain over the years, but they can provide the necessary tools to support the decision-making to improve processes and products.

Considering all this, the deliverable has presented different industrial use cases with problems still pending to be solved, in which the use and combination of IA technologies allows to improve the quality of processes and products in an extrusion process and in additive manufacturing process.

5 References

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, a. P. (2013). Understanding variable importances in Forests of randomized trees. *Advances in Neural Information Processing Systems.*

Wheeler, D. J. (06 de 07 de 2009). *QualityDigest.* Obtenido de https://www.qualitydigest.com/inside/health-care-column/good-limits-bad-data.html

Wu, G., Mallipeddi, R., & Suganthan, P. N. (2019). Ensemble strategies for population-based optimization algorithms–A survey. *Swarm and Evolutionary Computation*, 695-711.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 957391.