FAIRICUBE –
F.A.I.R. INFORMATION CUBES

Project Number: 101059238

WP3 Process
D3.6 Validation of processing and ML applications

Deliverable Lead: 4SF
Deliverable due date: 30/06/2024

Version: 1.5
01/07/2024

Document Control Page

Document Control Page	
Title	Deliverable D3.6 Validation of processing and ML applications
Creator	Stefan Jetschny
Description	Deliverable D3.6 Validation of processing and ML applications
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Jaume Targa
Date of delivery	30/06/2024
Type	R — Document, report
Language	EN-GB
Rights	Copyright "FAIRICUBE – F.A.I.R. information cubes"
Audience	☒ Public ☐ Confidential ☐ Classified
Status	☐ In Progress ☐ For Review ☒ For Approval ☐ Approved

Revision History			
Version	Date	Modified by	Comments
0.1	11/12/2023	Stefan Jetschny	Setup of document
0.5	14/12/2023	Stefan Jetschny	Agreed structure in place
1.0	20/01/2024	Jaume Targa	Input on data validation chapter
1.1	23/01/2024	Jaume Targa	Draft checklist table for validation
1.2	26/01/2024	Stefan Jetschny	Update of checklist
1.3	05/02/2024	Mirko Gregor/Jaume Targa	Final Review
1.4	13/05/2024	Stefan Jetschny	Review with feedback from General Project Review Consolidated Report, disclaimer added to the checklist, future usage outlined
1.5	01/07/2024	Stefan Jetschny	Minor edit of Validation checklist, removed the AI ethics row, this is part of the AI ethics validation of D6.11

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

Disclaimer

This document is issued within the frame and for the purpose of the FAIRICUBE project. This project has received funding from the European Union's Horizon research and innovation programme under grant agreement No. 101059238. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

Table of Contents

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

List of Tables

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

1 Introduction

In the ever-evolving landscape of technology, processing, and machine learning (ML), applications have become integral components of commercial and research fields. Machine learning applications perform automating tasks, making predictions, and extracting valuable insights. However, to harness the true potential of these applications, it is essential to focus on the validation process, ensuring accuracy, reliability, and overall effectiveness.

Validation is a critical step in the development and deployment of processing and ML applications. It involves assessing the performance of algorithms and models to confirm that they meet the specified requirements and deliver accurate results. The validation process is particularly crucial as it helps identify potential issues, mitigate risks, and provide confidence to the end-users.

1.1 Overview

In general, the validation of processing and ML applications can be subdivided into data, model, performance, and ethical/bias validation which is in parts covered by other FAIRiCUBE deliverables:

- Data Validation:
    - The basis of every data processing and ML application is data and the data quality determines the quality of the process. Validating the input data ensures that it is accurate, representative, complete, consistent, relevant, and free from biases. Cleaning and preprocessing data are essential steps to enhance the overall performance of the data processing and ML application.
    - Covered in D3.1 *UC exploratory data analysis*
    - Covered in D5.3 *Validation of ingestion*
- Data Processing Validation:
    - Main element will be the validation of the correct implementation of processing steps, including benchmarking and documentation of processing steps as meta data information.
    - Covered in D3.2 *Machine learning strategy specific for each use case*
    - Covered in D4.3 *Public Listing (Catalogue) of FAIRiCUBE processing-analysis resources*
- Model Validation:
    - ML models are the core components of processing applications. Validating these models involves assessing their accuracy, precision, recall, and other performance metrics like creating a confusion matrix. Techniques such as cross-validation and split testing help ensure the robustness of the model across different datasets.
    - Covered in D3.2 *Machine learning strategy specific for each use case*
- Performance Validation:
    - Assessing the performance of processing and ML applications under various conditions is crucial. This involves testing their efficiency, scalability, and responsiveness to ensure they meet the demands of real-world scenarios.
    - Covered in D3.3 *Processing and ML applications*
- Ethical and Bias Validation:

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

- ML applications must be validated for ethical considerations and potential biases. It's important to ensure that the models are fair and unbiased, treating all individuals and groups equally. Addressing bias in algorithms is essential to avoid unintended consequences and ensure ethical use.
- Covered in D6.11 *OEI - Requirement No. 2 Ethics Board review*
- Covered in D7.1 *OEI - Requirement No. 1*

1.2 Motivation

First and foremost, validation significantly enhances the accuracy, performance and stability of processing and ML applications. By subjecting the algorithms and models to rigorous testing, developers can identify and rectify errors, resulting in improved precision and effectiveness. This improved accuracy, in turn, increases the decision-making capabilities and overall performance of the applications.

Moreover, validation fosters increased reliability. Trust is paramount in user adoption, and by validating the processing and ML applications, developers provide confidence to end-users. Reliability is not only about accurate predictions but also about consistent and trustworthy results. Users are more likely to embrace and depend on applications that consistently deliver reliable outcomes. Validation also plays an important role in risk mitigation. By proactively identifying and addressing potential issues during the validation process, developers can significantly reduce the likelihood of failures in real-world scenarios. This approach helps in creating robust applications that can withstand dynamic and challenging environments.

The benefits of validation extend to the ethical dimension and address potential biases. In today's diverse and interconnected world, it is crucial to especially validate ML applications for fairness and ethical use. Validation processes help identify and rectify biases in algorithms, ensuring that these applications treat all individuals and groups equally, thus avoiding unintended consequences. Generally, validation is instrumental in achieving regulatory compliance which can be mandatory in some critical industries with specific regulations and standards governing the use of technology, particularly in sensitive domains like healthcare and finance. The point of mandatory regulatory compliance is not yet a significant focus of the FAIRiCUBE deliverables and processes but may so in the future.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

2 Data validation

At the core of every data processing and machine learning (ML) application lies the fundamental element of data. The quality of this data is pivotal as it directly influences the efficacy of the entire process. Ensuring and validating the elements of data quality such as accuracy, representativeness, completeness, consistency, relevance, and absence of biases in the input data for each data science work is crucial. Data quality needs to be described, documented, and strictly preserved during data handling. In case of data quality issues, measures to improve data quality can be applied.

Under the FAIRiCUBE umbrella, data validation covers two key aspects: the quality of the data itself and the quality of the data after the ingestion process into FAIRiCUBE platforms.
Several methods for describing data quality and conducting quality checks are elaborated in delivery D5.3. These cover the key aspects of data ingestion into FAIRiCUBE platforms. Accuracy, for example, ensures that the data is free from errors, while representativeness guarantees that the dataset accurately mirrors the real-world scenarios it aims to model. Completeness ensures that no essential information is missing, and consistency verifies that the data adheres to predefined rules and standards. Relevance ensures that the data is pertinent to the specific problem or task at hand, and the absence of biases is crucial to prevent skewed results or unfair predictions. The documentation of data quality can be seen as mostly a descriptive procedure as it captures the state of the data as it is. Results of the data quality assessments are considered in the data exploratory phase (see *D3.1_UC exploratory data analysis*) and have consequences for the development of a processing and machine learning strategy (see *D3.2_Machine learning strategy specific for each use case*). Once data quality issues prevent the proper application of processing and ML applications, data needs to be disregarded or alternative approaches must be applied to increase data quality.

To optimize or even enable the performance of data processing and ML applications, measures to improve data quality can be applied, e.g. data cleaning and preprocessing. These steps involve identifying and rectifying inaccuracies, handling missing values, normalizing data to a consistent format, and addressing any anomalies or outliers. By meticulously refining the data through these processes, the overall performance and reliability of the data processing and ML application are significantly enhanced. This, in turn, contributes to more accurate insights, predictions, and outcomes in the subsequent stages of the application's lifecycle. Further, additional processing and or ML applications can be applied to improve data quality issues such as gap filling which in turn have to be validated as well.

Further, validation of input data entails the data quality preservation aspect. Data errors have the potential to arise at various stages of a data processing procedure from provision to transformation, manipulation, and processing to final delivery. The FAIRiCUBE data ingestion process is currently seen as one of the most critical steps for FAIRiCUBE data science work as this is under the direct control of FAIRiCUBE whereas the original earth observation raw data and data products provisioning is outside of FAIRiCUBE scope. Consequently, when considering data validation in the following we especially address the verification and validation of the data ingestion step as foreseen and to be implemented and executed as part of the FAIRiCUBE data ingestion pipeline. In short, during the data validation process, we aim to ensure that the provision data to the FAIRiCUBE use cases as identical as possible

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

to the original raw data and products after ingestion. For direct access to data from its original source (through e.g. APIs) which is not ingested through FAIRiCUBE work, the data validation step may be omitted.

For data cubes that do not require transformations, near-binary identity of data before and after data ingestion can be assumed. For transformations like re-gridding or CRS conversations, numerical differences are expected, data validation assessment during the conversion from vector data to gridded data is not trivial and require extra attention.

In summary, once data descriptive and statistical quality measures are in place and well documented as part of meta data and further documentation of the data science work, data can be validated across the data analysis and processing chain as well as be used to evaluate the improvement of data quality.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

3 Data processing validation

Data processing validation including algorithm implementation validation are crucial steps in ensuring the reliability and correctness of data processing and machine learning systems. There are several techniques and practices we can employ to validate data processing and algorithm implementations.
Firstly, we can implement unit tests for individual components of our data processing pipeline and algorithms. This helps verify that each unit of code behaves as expected in isolation. Following that, conduct integration tests to assess the interactions between different components of the system, verifying that data flows smoothly between processing stages and that the overall pipeline functions correctly. For a more holistic evaluation, we can perform end-to-end testing to assess the entire data processing workflow. This involves testing the complete pipeline with controlled (i.e. synthetic) but representative data to ensure that it produces the expected results. Additionally, if we are working on a machine learning model, we can use techniques like cross-validation to assess the model's performance across different subsets of the data, ensuring it generalizes well to new, unseen data.

Benchmarking is another essential practice. We can compare the performance of your algorithm against established benchmarks or alternative implementations to provide context for understanding whether our implementation is competitive and effective. We can further create visualizations and plots to better understand the intermediate and final results of our data processing steps. Visualization can reveal patterns, trends, and potential issues in the data. To evaluate the robustness of your algorithm we can test processing steps with variations in input data, such as introducing noise, outliers, or missing values. Thereby, we assess how well the system handles unexpected or challenging scenarios.

A key element of data processing validation is comprehensive documentation which should be tightly linked to the data itself. We will document the rationale behind design choices, assumptions, and dependencies to facilitate understanding and future updates as well as indicate at which stage of the data pipeline which processing step was applied to the data. Our vision for FAIRiCUBE is to make the whole chdain of data processing steps including machine learning as FAIR as possible, i.e. the documentation of the final results includes direct links to meta data resources, technical documentation including ML model testing and parameters, and if needed data pre-processing steps as part of the data ingestion process.

4 Machine learning validation

Machine learning model validation is a crucial phase in the development of robust and reliable models. The primary goal of model validation is to ensure that the trained model generalizes well to new, unseen data, and its performance aligns with the intended objectives. Several key practices contribute to effective model validation which are briefly explained in the following.

One fundamental strategy in model validation is the division of the dataset into distinct training and test sets. This separation allows developers to measure how well the model performs on data it hasn't encountered during training, assessing its ability to generalize. Cross-validation is another powerful technique employed in model validation. By dividing the dataset into *n* subsets (batches) and iteratively using different subsets for training and testing, cross-validation provides a more robust evaluation of the model's performance. The average performance across all batches helps in understanding how well the model generalizes to diverse subsets of the data.

Defining appropriate validation metrics is essential for evaluating the model's performance accurately. Metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve are chosen based on the nature of the problem at hand. For instance, in applications like medical diagnosis, sensitivity (recall) might be more critical than overall accuracy. Hyperparameter tuning is a critical aspect of model validation. Many machine learning models have hyperparameters that are not learned from the data but significantly impact the model's performance. Techniques like grid search or random search are commonly employed to identify the optimal set of hyperparameter values.

Guarding against overfitting and underfitting is a continuous concern in model validation. Overfitting occurs when a model performs exceptionally well on the training data but poorly on new data, while underfitting happens when the model is too simple to capture underlying patterns. Validation techniques, along with regularization and adjustments to model complexity, help strike the right balance.

Beyond performance metrics, model validation also considers ethical considerations such as bias and fairness. Ensuring that models are unbiased, and fair is critical, especially in applications that impact individuals or groups. Techniques like fairness-aware machine learning are employed to address fairness concerns.

Model validation doesn't end with development—it extends to the deployment phase. Continuous monitoring of the model's performance in production is crucial to identify and address issues that may arise as the data distribution evolves over time.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

5 Performance validation

Assessing the performance of processing and ML applications under various conditions is crucial. Monitoring computational resources offers three significant advantages. Initially, it enables the main user to furnish insights and estimates to other users engaging in similar tasks on comparable hardware.

Secondly, obtaining the insights into hardware prerequisites and resource utilisation could influence the cost of cloud resources, which often remains uncertain during project planning. Lastly, gathering data on the actual performance and demands of computational tasks serves as a starting point for numerical optimisation, particularly when expectations are not met by current measures.

Assessing the performance of processing and machine learning (ML) applications under various conditions is a critical aspect of their development and deployment. This evaluation process involves testing several key aspects to ensure that the applications meet the demands of real-world scenarios:

- **Efficiency -** Assessing efficiency involves measuring factors like CPU utilisation, memory usage, and disk I/O operations.
- **Scalability -** Testing scalability involves measuring how well an application performs as the workload or dataset size grows.
- **Responsiveness -** responsiveness involves measuring factors such as inference latency or model prediction time. For processing applications, responsiveness may include metrics like task completion time or event processing latency.
- **Robustness -** Testing robustness involves intentionally introducing faults or errors into the system and evaluating how well it handles them.

Overall, assessing the performance of processing and ML applications involves a comprehensive evaluation of their efficiency, scalability, responsiveness, and robustness under various conditions. By conducting thorough performance testing, developers can identify and address performance bottlenecks, optimize resource utilization, and ensure that the applications meet the requirements of real-world deployment.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

6 Ethical and Bias validation

Ethical validation comprises several key considerations, with a primary goal of ensuring fair treatment across different demographic groups. This involves a thorough examination of the application's impact on variables such as race, gender, age, and other sensitive attributes to prevent discriminatory outcomes. Privacy considerations are also paramount, necessitating an evaluation of how the application handles user data in terms of collection, storage, and processing, in compliance with data protection regulations. Moreover, ethical validation delves into the crucial aspect of obtaining informed consent from users. This involves ensuring that users are adequately informed about how their data will be used and that they have transparent options to consent or opt-out based on clear information provided by the application.

Bias validation begins with a comprehensive assessment of biases within the training data as already discussed in chapter 4. Identification and mitigation of biases are essential to prevent skewed model outcomes, ensuring that the model does not inadvertently discriminate against certain groups. Techniques such as fairness-aware machine learning are employed to measure and address biases within the model itself, ensuring that predictions remain impartial and unbiased. While bias validation can be important for the numerical stability and meaningfulness of results from an ML application, from the ethical assessment point of view bias validation is of great importance as the interpretation of a bias ML application can lead to discrimination and can have severe consequences for humans, the environment, goods, and services, etc. Validating a balanced distribution of the input data and derived features should have a high priority.

Not necessarily directly related to the ethics validation, however nevertheless an enabler for general, ethical and performance validation is the introduction of Explainable AI (XAI) which focuses on making the decision-making process of ML models transparent and understandable to users. Transparent models not only enhance user trust but also empower users to comprehend the reasoning behind specific decisions. Techniques such as feature importance analysis and model-agnostic interpretability methods contribute to the transparency of ML models, ensuring that users can grasp how and why a particular decision was made.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

7 Validation Checklist

To provide clear and structured guidance for users who apply processing and machine learning methods to their data, we have developed a checklist as shown in Table 1. Following the table will create awareness of available methods and will assure the proper assessment of all validation aspects. Further efforts will be invested to document the validation results, integrate them into appropriate documentation and meta-data information as well as make them available to other users of the data. As for now, this checklist has only been conceptionally verified with the use cases under FAIRiCUBE and is mostly meant to provide guidance and concrete processing and ML application validation steps to follow. Further, the guidance shall be applicable to all current and future use cases executed under FAIRiCUBE and in principle every data science framework. No use case specifics are therefore included and the checklist as the simplified and potentially main output of this deliverable will not contain results of the validation of the processing and ML applications. It is foreseen to make a webservice form from Table 1 with an optional field for comments for each checkbox item and the guidance results with comments can be harvested and provisioned as part of the Knowledge Base services.

Some elements of the table are matched against the 7 key requirements of AI systems as specified in the Ethics guidelines for trustworthy AI [1]. Our overview however aims to address validation in a broader context and includes additional validation checks while also suggests concrete steps for use cases to consider.

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

Table 1: Validation checklist of processing and ML applications

Process	Check type	Feature	Description	Check
Data Validation for Ingestion	General checks	Feature List Generation	Create a list of features that need to be validated for the ingested dataset including: - data completeness, correctness, consistency, and - conformity to predefined standards.	☐
		Descriptive Statistics Calculation	Describe statistics computed for each feature, including measures such as - Mean and standard deviation, - minimum, maximum, and - other relevant statistical indicators.	☒
		Anomaly detection	Describe process applied to identify any significant deviations or anomalies in the newly ingested data.	☒
		Error Labelling and Data Incorporation	Based on the results of above, the ingested data is labelled either as acceptable or erroneous.	☒
		Spatial Validation	Verify that the spatial attributes have been preserved accurately during the ingestion process.	☒
		Reporting and Logging	Validation of outcomes, including any detected errors or anomalies.	☒
	Data validation checks	Duplicates	Check for duplicate entries.	☒
		Date overlap	Check for overlaps in the date column, i.e., repeated dates with different values (only for timeseries).	☒
		Date gaps	Check for missing dates between start and end date (only for timeseries).	☒
		No data values	Verification of the correct use of no data.	☒
		Value types	Check if data types are correct (string, integer, float, datetime format).	☒
		Value encoding	Check if the encoding of the data is correct (e.g., character encoding is utf-8; point (.) is used as decimal separator).	☒
		Completeness	Calculate and evaluate the ratio of not-NULL values.	☒

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

	Feature vector validation	Count of distinctive values	Count and evaluate the number of distinct values in the dataset.	☒
		Ratio of the most frequent value	Determine the number of occurrences for the most frequently repeated value, normalized by the batch size.	☒
		Maximum	Maximum value of the dataset	☒
		Mean	Mean value of the dataset	☒
		Minimum	Minimum value of the dataset	☒
		Standard deviation	Standard deviation of the dataset	☒
		Number of records	Number of rows in the dataset (e.g., number of polygons)	☒
		Date range	Start and end date (only for timeseries)	☒
	Feature spatial validation	Grid boundaries	Top-left and bottom-right coordinates (only for gridded datasets)	☒
		Data completeness	Verify data set is complete (total area, total number of features or pixel)	☒
		Projection/CRS	Verify correct use of the projection/CRS.	☒
		Pixel size	Verify that the pixel size is correct	☒
		Number of bands	Verify all channels have been transmitted correctly.	☒
		Number of attributes	Verify all attributes of the table or vector dataset have been transmitted correctly.	☒
		Datatype	Check if the data type is OK.	☒
		Data format	Check if the ingested data follows the desired standard format (e.g., for raster, cloud optimized Geo Tiff)	☒
		Centre coordinates and total area	For vector data, check if centre coordinates of polygons match (to find shifted polygons or duplicate polygons)	☒

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

Data processing Validation	Algorithm implementation validation	Technical Robustness and safety	Ensuring robustness and safety of the implementation through e.g. unit tests that can verify that unit of codes behaves as expected in isolation (inc. individual components of our data processing pipeline and algorithms).	☒
		Assess the interactions	Conduct integration tests to assess the interactions between different components of the system, verifying that data flows smoothly between processing stages and that the overall pipeline functions correctly.	☒
		End-to-end testing	Assess the entire data processing workflow. This involves testing the complete pipeline with controlled (i.e. synthetic) but representative data to ensure that it produces the expected results.	☒
		Cross-validation	Cross-validation to assess the model's performance across different subsets of the data, ensuring it generalizes well to new, unseen data	☒
	Benchmarking	Monitor compute resources	Monitoring and storing the consumption of computational resources as defined and described in the FAIRiCUBE GitHub repository.	☒
		Re-run and compare	If performance is in question, re-run of an application can be advised. The monitoring results of a compute task can be compared with other similar tasks as listed in the FAIRiCUBE knowledge base (https://fairicube-kb.dev.epsilon-italia.it/).	☒
	Comprehensive documentation	Documentation and transparency	Document the rationale behind design choices, assumptions, and dependencies to allow transparency of the processing and ML application methods.	☒
		Meta-data	Update, complete, and maintain meta-data records associated to the data set. This applies to the FAIRiCUBE analysis / processing meta-data.	☒
Machine learning validation		Dataset preparation for training	Create subsets of data separate for training, testing (and validation), check the selection method (random, by consecutive index).	☒
		Define appropriate validation metrics	Select and document appropriate metrics such as total accuracy, precision, recall, F1 score, area under the ROC curve to evaluate the performance of your ML method. Define expectations first, establish baseline methods to compare against.	☒

		Prevent/Test overfitting and underfitting	Comparing performance / accuracy metrics from applying an ML model to different datasets which have not been included in the training of the ML model will give insights on over or underfitting.	☒
		Statistical bias validation	Checking the statistical distribution of the input feature both within one feature space and across features will avoid unwanted biases in the training of the ML model. Some ML methods require certain statistical distributions (Gaussian or even distribution, etc.), some other methods require scaling of the data features.	☒
		Human agency and oversight	Implementation of human oversight mechanisms such as human-in-the-loop, human-on-the-loop, and human-in-command approaches. At each step of the ML application, user feedback and interaction should be foreseen.	☒

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

8 Conclusion and Outlook

In the fast-paced world of technology, the validation of processing and ML applications is paramount. By thoroughly assessing data, models, performance, and ethical considerations, developers and organizations can build applications that are not only accurate and reliable but also ethical and compliant with regulatory standards. As we continue to integrate these technologies into various aspects of our lives, the emphasis on validation becomes increasingly crucial for creating a positive and trustworthy technological future.

A holistic approach to machine learning model validation involves careful consideration of data splitting, cross-validation, appropriate metrics, hyperparameter tuning, addressing overfitting and underfitting, and ensuring fairness. This comprehensive strategy ensures that ML models not only perform well during development but also maintain their reliability and accuracy when deployed in practical applications. To make the user aware of all validation checks possible and to record validation progress, we have created an intuitive overview table that can be followed and ticked off.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.