Ref. Ares(2025)6864805 - 25/08/2025



HORIZON-CL6-2024-CIRCBIO-01-5

Programmed biodegradation capability of bio-based materials and products, validated in specific environments

MAGICBIOMAT

Manufacturing Guiding Tool for Circular and Programmed Biodegradable Materials Under Open Environment Conditions

Start Date: January 1st, 2025

Duration: 48 Months

Deliverable: D3.1

Deliverable Title - AI-driven Data Quality Inspection Tool

Due date of deliverable: 31/08/2025

Actual submission date: 28/08/2025

Responsible WP: WP3, Soumya Kanti Datta, Digiotouch

Responsible Technical Leader: Soumya Kanti Datta, Digiotouch

Revision: V1.0

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Funded by the European Union

Authors

Author	Institution	Contact (e-mail, phone)
Soumya Kanti Datta	**Digiotouch**	**soumya@digiotouch.com**
Marharyta Domnich	**Digiotouch**	**marharyta@digiotouch.com**

Document History

Document version	Date DD/MM/YY	Changes
V0.1	15/07/2025	Table of content and initial content
V0.2	25/07/2025	Added state-of-the-art
V0.3	15/08/2025	Completed the first draft
V0.4	22/08/2025	Internal review completed and deliverable updated
V1.0	26/08/2025	Released for submission to the EC portal

Document Approval

Reviewers		Validation date
Work Package Leader	**Digiotouch**	**25/08/2025**
Coordinator	**TUS**	**27/08/2025**

Document Data

Keywords	Artificial Intelligence; Cloud Computing; Data Quality Inspection Tool
Point of Contact	Name: Soumya Kanti Datta Partner: Digiotouch Address: Narva mnt 5, 10117 Tallinn, Estonia Phone: +372 5854 2328 E-mail: soumya@digiotouch.com
Partner:	Digiotouch

Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them

Acronyms

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
DCAI	Data-centric AI
EU	European Commission
FAIR	Findable, Accessible, Interoperable, Reusable
GCN	Graph Convolutional Network
GDPR	General Data Protection Regulation
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
PCA	Principle Component Analysis
QSAR	Quantitative Structure Activity Relationship
WP	Work Package
XAI	Explainable AI

Executive Summary

Deliverable D3.1 “AI-driven Data Quality Inspection Tool” presents the main outcomes of Task 3.1 in WP3. It establishes the foundations for ensuring that high-quality, validated data enter the project’s biodegradability models. The deliverable reviews the state-of-the-art in data curation for AI models, anomaly detection, and trustworthy AI, and identifies relevant public datasets on biodegradability outcomes, polymer properties, and microbial degradation. It also introduces the AI-driven data quality inspection pipeline, covering data ingestion, schema validation, anomaly detection, and expert-in-the-loop validation. Compliance with GDPR, FAIR data principles, and data security standards is embedded in the design. An assessment against the ALTAI framework demonstrates alignment with European guidelines for trustworthy AI. Future tasks of WP3 (T3.2 and T3.3) will utilise the tool and enrich it with explainability and robustness features and validate it in pilot environment.

Table of Contents

Acronyms 3

Executive Summary 4

List of Figures 7

1. Introduction 8

1.1. Context and purpose of the deliverable 8

1.2. Organisation of the deliverable 8

2. State-of-the-Art 10

2.1. Data quality control and AI-driven curation 10

2.2. Anomaly detection in time-series and tabular data 11

2.3. Applications in biodegradability and chemical process data 11

2.4. Ensuring model reliability and trustworthiness 12

3. Data collection and AI-driven data quality inspection tool 14

3.1. Data requirements and compliance 14

3.1.1. Data types and scope 14

3.1.2. GDPR and ethical compliance 14

3.1.3. Data security and protection 15

3.1.4. Platform setup and infrastructure 15

3.2. Public datasets for biodegradability research 15

3.2.1. Polymer properties databases 15

3.2.2. Biodegradability datasets 16

3.2.3. Microbial and enzymatic degradation databases 16

3.2.4. QSAR biodegradation 16

3.3. Anomalous data behaviours and detection techniques 17

3.3.1. Types of anomalous behaviours 17

3.3.2. Rule-based and statistical methods 17

3.4. Overall pipeline for data collection and quality inspection 18

3.4.1. Data ingestion 18
3.4.2. Schema definition 19
3.4.3. Anomaly detection and data quality inspection tool 19
3.4.4. Human-in-the-loop validation 19
3.4.5. Continuous enrichment and iterative improvement 19
4. Trustworthy AI Evaluation Using ALTAI 20
4.1. Human agency and oversight 20
4.2. Technical robustness and safety 20
4.3. Privacy and data governance 20
4.4. Transparency and explainability 21
4.5. Diversity, non-discrimination, and fairness 21
4.6. Societal and environmental well-being 21
4.7. Accountability 21
5. Conclusion 22
6. References 23

List of Figures

Figure 1: MAGICBIOMAT data collection pipeline including the data quality inspection tool._____18

1. Introduction

Deliverable D3.1 marks the main outcome of Task 3.1 of MAGICBIOMAT, focusing on the development of an AI-driven data quality inspection tool. The purpose of this deliverable is to establish the methodological and technical foundations for ensuring that only high-quality, validated data is used within the project's biodegradability models and manufacturing guidance systems. The work began with an assessment of requirements, including the types of datasets to be collected, GDPR and ethical compliance obligations, and the infrastructure necessary to guarantee secure, traceable, and FAIR-aligned data management. Building on this foundation, the partners involved in Task 3.1 reviewed the state-of-the-art in AI-driven data curation, anomaly detection, and trustworthiness, with a particular emphasis on biodegradability applications and industrial data inspection. Publicly available datasets on polymer properties, biodegradability test outcomes, and microbial degradation capabilities were identified as initial sources to populate and test the tool, with the anticipation that new datasets (generated in the other WPs and pilots) will be incrementally integrated as the project progresses. Furthermore, different classes of anomalous behaviours and error conditions were defined, and a layered detection framework has been designed and implemented. Finally, the deliverable incorporates a preliminary ALTAI-based assessment to ensure that the development of the inspection tool is guided by the principles of trustworthy AI from the outset.

1.1. Context and purpose of the deliverable

Deliverable D3.1 documents the technical specification and first implementation of the AI-driven data quality inspection tool in WP3. Its purpose is to enable ingestion of heterogeneous datasets, ranging from polymer descriptors and tabular biodegradability assays to time-series sensor data, into a pipeline that applies schema validation, anomaly detection, and expert review. The tool is contextualised within MAGICBIOMAT's data ecosystem, ensuring interoperability with other work packages, GDPR compliance, and alignment with FAIR principles. It acts as a safeguard so that downstream machine learning models for biodegradability prediction receive only validated, high-quality data.

1.2. Organisation of the deliverable

This deliverable is organised into several key sections.

Section 2 reviews recent advances in data quality control, anomaly detection, and trustworthy AI methods, with emphasis on biodegradability applications.

Section 3 describes requirements, compliance considerations, and the design of the pipeline, including schema validation, anomaly detection, and human-in-the-loop review. The public datasets subsection identifies external resources such as polymer property databases, biodegradability datasets, and microbial degradation repositories relevant to Task 3.1.

Section 4 provides a preliminary screening of the tool against the European framework for trustworthy AI (with respect to ALTAI framework), ensuring alignment with ethical and technical principles from the outset.

Finally, section 5 concludes the deliverable with key outcomes and next steps.

State-of-the-Art

This section provides the scientific and technological background for Task 3.1 which focuses on data collection and AI-driven quality inspection. The section surveys how recent advances in data quality control, anomaly detection, and model reliability are applied to industrial contexts, with particular emphasis on biodegradability studies. The section is structured into four parts:(a) AI-driven data curation techniques for improving tabular datasets, (b) methods for anomaly detection in time-series process data, (c) specific applications in biodegradability and chemical process monitoring, and (d) approaches for ensuring reliability and trustworthiness of AI models. The review highlights how robust data inspection enhances predictive modelling, safeguards process monitoring, and supports compliance in high-stakes environments. Ultimately, this overview demonstrates the relevance of emerging AI inspection tools for delivering accurate, reliable, and trustworthy outcomes in MAGICBIOMAT.

2.1. Data quality control and AI-driven curation

High-quality data is the foundation of reliable AI models, especially in industrial biodegradability studies. Recent work emphasizes a data-centric AI (DCAI) paradigm that focuses on systematically improving data (through cleaning, error correction, etc.) to build more robust models [1]. This shift contrasts with the traditional “model-centric” approach and ensures ML models generalize better by prioritizing data quality. In practice, modern AI-driven data quality systems use both rule-based and learning-based methods. For example, probabilistic data-repair frameworks like HoloClean [2] and interactive cleansing tools like ActiveClean [3] have revolutionized data quality management, achieving substantial accuracy gains while greatly reducing manual effort. A 2024 systematic review reported that AI-driven cleansing improved data accuracy by up to 85% and reduced manual intervention by 65% [4].

A concrete illustration comes from biodegradability prediction datasets. Körner et al. [5] applied rigorous data curation and label noise filtering to a large chemical biodegradation dataset. They identified issues like mismatched chemical identifiers and data leakage, then corrected errors and removed chemicals with inconsistent experimental results. The result was striking, balanced accuracy of a biodegradability classifier jumped from 80.9% to 94.2%. This demonstrates how AI-driven inspection of tabular data (for errors, outliers or label noise) can dramatically boost model reliability. The authors note that many previous biodegradability models had been model-centric (tuning algorithms and features) rather than data-centric, and that a stronger focus on data quality was key to surpassing prior performance limits. These findings align with the broader trend in industry: automated data quality monitoring (e.g. using streaming frameworks for real-time checks) and AI-enhanced cleansing are now recognised as critical for trustworthy AI in high-stakes domains.

2.2. Anomaly detection in time-series and tabular data

In industrial settings [6], AI-driven anomaly detection is essential for monitoring time-series data from sensors, chemical assays, and process logs. The goal is to enable early warnings of unusual or unsafe conditions. Traditional statistical process control techniques, such as univariate or multivariate control charts and principal component analysis (PCA), have long been used to flag deviations in process data. These methods remain powerful for summarizing behaviour and are often fast and interpretable. For example, PCA-based control charts can detect sudden shifts in reactor sensor streams. This concept underpins many modern techniques, essentially, a model of "normal" operation is learned from historical data, and alerts are triggered when new observations fall outside this range.

Contemporary research has greatly expanded the toolkit. Unsupervised machine learning methods can monitor many correlated signals and handle non-linear patterns or outliers better than fixed thresholds. For instance, manifold learning techniques like UMAP combined with clustering (e.g., HDBSCAN) [7] have shown superior separation of events. Simpler anomaly detectors for tabular data are also widely used, such as Isolation Forests or one-class SVMs, which learn the pattern of normal samples and then identify points that deviate significantly. Indeed, Isolation Forests have been applied to industrial datasets to detect and even help explain the sources of anomalies (e.g. pinpointing which sensor or factor is most contributory) [16].

Deep learning has also been applied to anomaly detection in recent years. Autoencoders (and recurrent variants) learn to compress and reconstruct the time-series data, and large reconstruction error indicates an anomaly. A 2024 survey notes the diffusion of autoencoders in Industry 4.0 anomaly detection [8]. Other popular architectures include LSTMs for temporal anomalies and GANs [9], which train on normal data and use the discriminator to detect out-of-distribution signals. An evaluation on the Tennessee Eastman process dataset found autoencoders were top performers, followed by GANs and forecasting models [11]. This supports the consensus that autoencoders' ability to capture a compressed signature of normal operations makes them highly effective for anomaly identification in practice.

2.3. Applications in biodegradability and chemical process data

The domain of chemical and environmental engineering has benefited from these advances. Biodegradability assessments generate both tabular assay datasets and time-series measurements (e.g. oxygen uptake). AI techniques are used to ensure the integrity of such data and to build predictive models of biodegradation outcomes. Körner et al. [5] achieved the most accurate classification model to date for aerobic biodegradation by curating training data. They also published a list of uncertain substances for re-testing. This data-centric approach improves model reliability and highlights how data inspection (checking for mislabelled or hard-to-learn instances) is vital in chemical informatics to avoid propagating errors in QSAR models.

On the modelling front, researchers have applied advanced Machine Learning (ML) to predict biodegradability from molecular structures. The authors of [11] compared a QSAR descriptor-based model with a deep Graph Convolutional Network (GCN) for classifying chemicals. They found the GCN achieved comparable or better performance with less feature engineering. Other deep learning approaches like message-passing networks and recurrent models are also being explored in similar context.

Beyond predictive modelling, real-time process monitoring in biodegradation reactors leverages anomaly detection for sensor networks (pH, oxygen, etc.). Recent surveys document deployments of anomaly detection in smart manufacturing [13] and chemical industries [14].

Sophisticated models place even greater importance on data quality shown in [5] study showed, feeding noisy or inconsistent data can obscure the benefits of advanced algorithms. Thus, in industrial biotech and environmental applications, there is a strong coupling between data inspection methods (to curate datasets, detect anomalies in sensor readings, etc.) and the success of AI models in predicting complex phenomena like biodegradation.

2.4. Ensuring model reliability and trustworthiness

The integration of AI for data inspection serves a larger goal of improving reliability and trustworthiness of AI models. One aspect is preventing models from being misled by bad data. Techniques like outlier detection and drift monitoring guard against unreliable predictions. Hybrid modelling approaches combining first-principles physics with ML have shown improved accuracy and interpretability in process systems [14]. Digital twins similarly sanity-check AI outputs against expected behaviour.

Another critical dimension is the estimation of predictive uncertainty. Uncertainty-aware models, such as Bayesian neural networks, ensemble methods, or techniques based on Monte Carlo dropout, allow users to quantify confidence intervals around predictions. In biodegradability and chemical process monitoring, this is particularly valuable: for example, a high-confidence anomaly flag can trigger immediate intervention, whereas a low-confidence output may instead prompt additional data collection or expert review. Reporting uncertainty estimates alongside predictions reduces the risk of overreliance on AI outputs and provides operators with calibrated risk information [26].

Transparency and explainability are also key enablers of trust. Modern Explainable AI (XAI) tools like SHAP can interpret tabular and time-series models to verify reasoning. In one chemical process example, SHAP analysis showed that injected noise variables had minimal impact on the yield prediction, confirming the model was focusing on the true process drivers [16]. Such insights increase confidence that the model is behaving sensibly and that any anomalies it flags are meaningful. Moreover, when an anomaly is detected, AI systems now often provide diagnostic information, e.g., identifying which sensor or feature caused the alert enabling faster root-cause analysis. This ability to explain and contextualize anomalies is paramount in industrial environments where safety and compliance are on the line.

The literature highlights a clear evolution from simple statistical monitors to advanced machine learning models capable of capturing subtle multidimensional patterns of normal and faulty behaviour. These AI-driven inspection methods are now widely applied in real-time industrial settings, including biodegradability processes, where they range from automated data cleaning of experimental logs to unsupervised detection of anomalous events. By combining insights from general-purpose AI with domain-specific needs, the field is advancing rapidly toward more accurate models, resilient processes, and improved safety.

3. Data collection and AI-driven data quality inspection tool

Task 3.1 partners, with DIGI's leadership, undertook a structured approach to design and develop the AI-driven data quality inspection tool. The first step was to assess the requirements for the tool, focusing on the types of data that need to be managed, the regulatory and ethical compliance obligations, and the necessary infrastructure for secure storage and processing. This assessment ensured that the tool is aligned with both the scientific objectives of MAGICBIOMAT and the operational constraints of collaborative research. In parallel, a systematic review of publicly available datasets was carried out, covering polymer property databases, curated biodegradability datasets, and microbial and enzymatic degradation repositories. These datasets form the initial foundation for the tool and will be incrementally expanded as new data sources (generated in this project in other work packages, pilots, and any new public datasets found) are identified during the project. Building on this foundation, the work then focused on investigating forms of anomalous behaviours in such datasets and selecting appropriate rule-based, statistical, and AI-driven methods to recognise and flag them. This step is critical to ensure that only high-quality, validated data enters subsequent AI pipelines. By integrating these elements, Task 3.1 establishes both the methodological basis and the practical implementation performed for the MAGICBIOMAT data quality inspection tool.

3.1. Data requirements and compliance

3.1.1. Data types and scope

For biodegradable plastics, MAGICBIPMAT requires a diverse range of datasets covering both raw experimental measurements and pre-processed, analysis-ready data. This includes (a) polymer descriptors (chemical structures, monomer composition, molecular weights), (b) material properties (thermal stability, tensile strength, mechanical behaviour, barrier properties), as well as (c) biodegradability metrics such as degradation rates, mass loss curves, or CO_2 evolution under standardised conditions (e.g., OECD 301 tests [17]). Both tabular data (assay results, chemical descriptors) and time-series data (sensor logs of degradation experiments, oxygen uptake) are relevant, ensuring models can capture dynamics across different modalities. To maximize utility, these datasets should include metadata such as test conditions (temperature, pH, microbial inoculum, etc.), experimental protocols, and uncertainty estimates, which are crucial for reproducibility and for contextualising anomalies in later inspections.

3.1.2. GDPR and ethical compliance

Even though most biodegradability datasets are scientific and not directly personal, compliance with GDPR is deemed to be mandatory for any information that could indirectly identify individuals, such as data linked to laboratory personnel or proprietary industrial partners. GDPR requires principles of data minimization, pseudonymisation, and anonymisation to ensure no personal information can be traced back to individuals. In practice, this means removing metadata of names or identifiers, carefully anonymising datasets before sharing,

and documenting all data-handling procedures. For industry-linked datasets under non-disclosure agreements or research agreements, additional contractual obligations may apply. Research data management guidelines highlight the need to balance openness with confidentiality, ensuring that data protection requirements are met while still enabling scientific reuse [18]. These are described in detailed in the initial data management plan deliverable D13.2.

3.1.3. Data security and protection

Ensuring data security is fundamental when dealing with sensitive or high-value industrial biodegradability data. This includes encrypting data both in storage and during transfer, setting up strict access control (role-based permissions), and maintaining audit trails to record data access and modifications. Loss of research data can have significant consequences for MAGICBIOMAT execution, and hence, redundant backup systems and disaster recovery protocols must be in place. For multi-partner collaborations, formal agreements govern the secure handling of proprietary information and prevent misuse. In addition to that, data should be stored and processed in infrastructure that complies with standards such as ISO/IEC 27001 (information security management), particularly when hosting sensitive datasets on cloud platforms or shared servers.

3.1.4. Platform setup and infrastructure

The project requires a robust data management platform capable of storing heterogeneous datasets (tabular, time-series, image, and chemical structure formats) and enforcing data validation rules automatically. Open-source infrastructures like Kadi4Mat [19], developed for materials science, demonstrate how collaborative research can integrate structured data storage with FAIR principles. Such platforms enable automated workflows: data upload, metadata tagging, schema validation, and anomaly checks before release. Importantly, they allow different partners to contribute data under a shared governance model while preserving data provenance and versioning. In MAGICBIOMAT, a similar setup would ensure that all collected biodegradability and polymer datasets are securely stored, consistently annotated, and validated before being used in downstream AI models. This alignment with best practices in open science and FAIR data stewardship guarantees both scientific rigour and any regulatory, ethical compliance.

3.2. Public datasets for biodegradability research

3.2.1. Polymer properties databases

A major resource for polymer research is the PoLyInfo database [20] maintained by National Institute for Materials Science (NIMS), Japan. PoLyInfo is one of the largest polymer property databases, containing information on more than 400,000 polymers (including homopolymers, copolymers, and blends) and over 100 property types, such as thermal (glass transition, melting point), electrical, and mechanical data. These datasets provide an invaluable baseline for the MAGICBIOMAT project, as they allow researchers to compare the properties of new biodegradable

plastics against known materials. Such comparisons are essential both for validating experimental measurements and for training AI models on structure–property relationships.

3.2.2. Biodegradability datasets

Several curated datasets are available which focus specifically on biodegradability outcomes. For example, the authors of [5] assembled and cleaned a dataset of thousands of organic chemicals with OECD 301 biodegradability test results (classified as readily vs. not readily biodegradable) to improve machine learning models.

MIT researchers published a dataset of 642 polyester and polycarbonate samples, each annotated with biodegradation outcomes measured through microbial assays, which was used to train machine learning models capable of predicting polymer biodegradability directly from chemical structure [21].

These datasets are being investigated for MAGICBIOMAT's predictive modelling tasks which can assist in both the QSAR-ready descriptors and raw biodegradation test outcomes.

3.2.3. Microbial and enzymatic degradation databases

Biodegradability is not only a material's property but also a function of microbial and enzymatic activity. Databases like PlasticDB catalogue [22] information on microbial species and enzymes known to degrade different types of plastics. The database currently has 824 species of microorganisms that were reported in the scientific literature to have plastic-degrading capabilities and 230 proteins described to breakdown plastics. In addition to that, we have investigated the Plastics-Active Enzymes (PAZy) [24] database compiles experimentally validated enzymes capable of breaking down synthetic polymer [23]. These datasets are valuable for cross-linking material degradation rates with potential biological mechanisms. For MAGICBIOMAT, they are also candidates for anomaly detection (e.g., checking if a reported biodegradation rate aligns with known microbial capabilities) and for guiding innovation in enzyme-assisted degradation strategies in the successive tasks of WP3.

3.2.4. QSAR biodegradation

The QSAR biodegradation dataset was built in the Milano Chemometrics and QSAR Research Group. It contains values for 41 attributes (molecular descriptors) used to classify 1055 chemicals into 2 classes: readily biodegradable and not readily biodegradable [25].

At the outset, it is emphasised that the datasets referenced in this section are those currently available in the public domain. In parallel, the project activities will generate new datasets through experimental work and partner contributions, which will be systematically integrated into the data inspection tool to enhance its coverage and performance. Furthermore, it is anticipated that during the course of WP3, MAGICBIOMAT partners may identify additional datasets of relevance to the objectives. In such cases, the data inspection tool will be iteratively updated and enriched, ensuring that it remains comprehensive and aligned with the evolving data landscape of the project.

This continuous enrichment process will safeguard the integrity of the data pipeline, ensuring that only high-quality, validated datasets are fed into the subsequent AI models and solutions.

3.3. Anomalous data behaviours and detection techniques

This subsection presents the strategies integrated into the AI-driven data quality inspection tool for detecting anomalous behaviours in datasets used in MAGICBIOMAT. The tool incorporates a layered approach, combining rule-based filters, statistical checks, and advanced machine learning techniques to identify outliers, contextual errors, and collective anomalies. These methods have been embedded to ensure that data irregularities are captured early and corrected before entering downstream AI models. By implementing this functionality, the tool directly safeguards the integrity and reliability of the project's data pipeline.

3.3.1. Types of anomalous behaviours

Anomalous behaviours in biodegradability datasets can manifest in several forms. Point anomalies (single outliers) occur when a measurement falls far outside the expected range — for example, a polymer reported with an implausibly high melting temperature. Contextual anomalies arise when values are abnormal in a given context, such as a biodegradation curve showing a sudden increase in mass (impossible biologically, likely due to instrumentation error). Collective anomalies involve sets of values that together contradict known material behaviour, e.g., property combinations that are physically unrealistic. In data management practice, anomalies also include missing values, duplicate entries, schema inconsistencies, and unit mismatches that distort downstream analysis. The developed data inspection tool is capable of detecting these classes of anomalies. It is essential to prevent contamination of AI training data with faulty records.

3.3.2. Rule-based and statistical methods

In addition to the above, the developed tool supports fast, interpretable, and simpler inspection methods relying on domain rules and statistical checks. Rule-based filters can automatically flag impossible values (e.g., negative biodegradation rates, or tensile strength higher than theoretical limits) or enforce consistency in units. Statistical techniques such as Z-scores and interquartile range (IQR) are commonly used to detect point anomalies by identifying values deviating significantly from the distribution. For multivariate data (e.g., polymer descriptors combined with assay results), principal component analysis (PCA) is incorporated to detect abnormal samples in reduced feature space.

When data from other work packages become available, the developed tool will be progressively enriched with more advanced inspection methods that leverage unsupervised machine learning to learn the normal distribution of data and flag deviations. Two promising approaches under consideration are: (a) clustering-based methods (e.g., k-means, DBSCAN, HDBSCAN), which can effectively separate normal from anomalous data clusters in material property space, and (b) one-class SVMs, which are designed to learn a boundary around normal samples

in high-dimensional descriptor space. These techniques are particularly suited to datasets generated within MAGICBIOMAT, where the majority of entries are expected to be valid and anomalies are rare but critical to detect. In addition, post-hoc explainability methods can be applied to identify out-of-distribution instances by highlighting atypical feature contributions compared to the training data. Such methods provide diagnostic insight into why a sample is flagged as unusual and supports expert validation. Furthermore, the tool is envisioned as an **evolving system**: as additional datasets are identified through project activities or sourced from public repositories, new AI-driven methods will be integrated to expand its detection capabilities. In this way, the data inspection tool will not remain static but will mature into a hybrid framework that combines domain-specific rules, statistical checks, and machine learning–based detectors, ensuring sustained robustness and adaptability throughout the WP3 lifecycle.

3.4. Overall pipeline for data collection and quality inspection

The overall pipeline for data collection and quality inspection is shown in Figure 1, which illustrates the integration of the AI-driven anomaly detection and data quality inspection tool.



Figure 1: MAGICBIOMAT data collection pipeline including the data quality inspection tool.

3.4.1. Data ingestion

The first stage of the pipeline is data ingestion, where datasets from multiple sources mentioned before are collected and transferred into a central storage in Digiotouch's cloud-based Paradise platform for raw data. This

step involves setting up connectors to handle different file formats (e.g., tabular CSVs, time-series sensor logs, image files, chemical structure representations) and ensuring all data enters the pipeline in a controlled manner. This raw data storage functions as a buffer zone where the collected raw data is held before quality checks, providing a clear boundary between external sources and the validated project dataset.

3.4.2. Schema definition

Once ingested, the data is subjected to **schema validation and constraint enforcement**. Here, each dataset is aligned with predefined template that specify allowable types, units, ranges, and expected completeness. For example, a schema might require that tensile strength values are expressed in MPa within plausible physical limits, or that biodegradation rates are reported as percentages over a defined time horizon. Tools such as TensorFlow Data Validation (TFDV), Great Expectations can automate these checks, flagging missing values, unit mismatches, or schema violations. Enforcing these constraints at an early stage helps to prevent the accumulation of structural inconsistencies that could undermine downstream AI modelling.

3.4.3. Anomaly detection and data quality inspection tool

Following schema validation, the pipeline integrates **AI-driven anomaly detection** to capture more subtle data issues. The tool currently supports detection of behaviour mentioned in Section 3.3.1 and methods mentioned in Section 3.3.2.

3.4.4. Human-in-the-loop validation

While automation is central to the pipeline and for MAGICBIOMAT, human expertise remains critical for adjudicating ambiguous cases. All anomalies flagged by the automated methods are routed into a human-in-the-loop review process, where domain experts (from TUS, UNIPD, OK) can classify them as genuine anomalies, experimental noise, or acceptable variation. Their feedback is captured and stored as training data for active learning, allowing the AI-driven inspection models to continually improve their accuracy over time. This interaction between automated detection and expert review ensures that false positives are minimised and that domain knowledge is systematically embedded in the tool.

3.4.5. Continuous enrichment and iterative improvement

The pipeline is not static, rather it is designed to evolve over the lifespan of WP3 of MAGICBIOMAT. As partners generate new experimental data or identify additional public datasets, the ingestion stage will be expanded and the anomaly detection models retrained with more hybrid strategies being adopted. This iterative process ensures that the inspection tool matures alongside the project, incorporating increasingly diverse and representative data. In the later stages, the pipeline will evolve into a robust, adaptive system that not only validates the quality of biodegradability datasets but also serves as a blueprint for AI-driven data inspection in broader similar applications.

4. Trustworthy AI Evaluation Using ALTAI

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) provides a practical framework for evaluating whether AI systems are aligned with the principles of trustworthy AI set out by the European Commission's High-Level Expert Group on AI. Conducting a preliminary screening at this stage is important for MAGICBIOMAT, as it ensures that the development of the data inspection tool is guided by considerations of human oversight, robustness, transparency, privacy, and societal well-being from the outset. By embedding these principles early in Task 3.1, the project safeguards against risks of bias, misuse, or opacity, and builds confidence that the AI solutions developed under WP3 will remain trustworthy and aligned with EU policy priorities.

4.1. Human agency and oversight

The tool is designed with a human-in-the-loop mechanism, where anomalies flagged by automated inspection methods are reviewed by domain experts. This ensures that decision-making authority remains with humans and that there are clear override and appeal pathways. Such oversight will be further supported by maintaining audit trails of expert interventions, which enhances accountability and transparency of the process for the entire WP3.

4.2. Technical robustness and safety

While the developed tool is technically robust and does not post any safety issue to humans or industrial processes, initial findings indicate that the main risks could potentially lie in distribution shifts (new polymers, new test conditions arising) and sensor drift during biodegradability experiments. Mitigation strategies already foreseen include version control of datasets, drift monitoring, and the use of redundant anomaly detection methods to reduce false positives or negatives. This layered approach strengthens resilience and improves the reliability of the tool.

4.3. Privacy and data governance

Although no personal data is expected to be processed, the project adopts a precautionary stance by applying GDPR principles such as pseudonymisation and minimisation to any potentially sensitive information (e.g. laboratory metadata linked to individuals). Licensing and data-sharing agreements are enforced for external datasets (where needed), ensuring proper data governance and compliance with legal obligations. Each dataset's origin, transformations, and usage history will be documented to ensure traceability, compliance, and reproducibility. Storage and processing will take place in infrastructures that comply with international standards such as ISO/IEC 27001 (information security management). Formal agreements across partners govern secure handling of proprietary information, while backup systems and disaster recovery protocols further strengthen data governance.

4.4. Transparency and explainability

Transparency is supported through the creation of dataset cards and model cards that document inputs, methods, limitations, and performance of the inspection tool. Explainability is not addressed yet but is anticipated to be covered in Task 3.2 and Task 3.3.

4.5. Diversity, non-discrimination, and fairness

Not applicable.

4.6. Societal and environmental well-being

By ensuring higher quality biodegradability data, the tool directly contributes to the project's environmental objectives, enabling more reliable assessments of sustainable materials. The tool also aligns with broader EU sustainability objectives by supporting the reliable assessment of biodegradable materials in line with the Green Deal and the Circular Economy Action Plan. The additional computational overhead of anomaly detection is considered minimal compared to its positive contribution to reducing waste, improving data quality, and supporting sustainability targets. Furthermore, the environmental footprint will be calculated within WP8 and WP9 which dedicates tasks for environmental impact analysis.

4.7. Accountability

Clear accountability mechanisms are in place through logging of all data processing steps, anomaly detections, and human reviews. This will pave way for a transparent audit trail and facilitates regular re-assessment. The responsibility for maintaining and improving the tool is shared within WP3, ensuring distributed accountability across technical and ethical dimensions.

This preliminary ALTAI based screening demonstrates that the AI-driven data quality inspection tool has been developed in line with the principles of trustworthy AI. However, it must be stressed that this is an initial assessment carried out early in the project lifecycle. The tool is anticipated to undergo formal ethical review along with the rest of the AI tools during the second reporting period of MAGICBIOMAT. Further refinements as new datasets and methods are highly likely to be integrated as WP3 progresses. Subsequent deliverables will therefore provide an updated ALTAI evaluation, ensuring that ethical, technical, and societal considerations remain integral throughout the project.

5. Conclusion

Deliverable D3.1 accomplishes a key task in MAGICBIOMAT by delivering a functioning AI-driven data quality inspection tool that transforms how biodegradability datasets are curated and validated in WP3. The tool already integrates multiple capabilities covering data ingestion, schema validation, anomaly detection, and expert-in-the-loop review into a coherent pipeline that enhances reliability and accelerates decision-making. In this deliverable, publicly available datasets were reviewed and mapped as candidate sources for integration, ensuring that the tool is designed to handle heterogeneous data types while maintaining interoperability with both public and project-generated datasets. The design explicitly incorporates compliance with regulatory and ethical standards. From a business perspective, this achievement reduces risks of poor-quality data propagating into downstream models, ultimately safeguarding the credibility of project outputs. It also positions the consortium with a scalable, adaptable solution that can extend to other industrial applications beyond biodegradable plastics. Looking ahead, the tool will continue to evolve with additional datasets and advanced methods, reinforcing its role as a cornerstone for trustworthy and impactful AI-driven solution for MAGICBIOMAT.

6. References

[1] Veluru, Suhasnadh Reddy and Erukude, Sai Teja and Marella, Viswa Chaitanya, Data-Centric AI: A Systematic Review of Methods, Challenges, and Future Directions (April 01, 2025). Multidisciplinary International Journal of Research and Development 4(4) 121-130 2025. Doi: http://dx.doi.org/10.2139/ssrn.5267741.

[2] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: holistic data repairs with probabilistic inference. Proc. VLDB Endow. 10, 11 (August 2017), 1190–1201. https://doi.org/10.14778/3137628.3137631.

[3] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. 2016. ActiveClean: interactive data cleaning for statistical modeling. Proc. VLDB Endow. 9, 12 (August 2016), 948–959. https://doi.org/10.14778/2994509.2994514.

[4] Anupkumar Ghogare, “AI-DRIVEN DATA QUALITY MANAGEMENT: A SYSTEMATIC REVIEW OF AUTOMATED DETECTION AND CLEANSING METHODOLOGIES”, *IJCET*, vol. 15, no. 6, pp. 442–453, Nov. 2024, Accessed: Aug. 24, 2025. [Online]. Available: https://lib-index.com/index.php/IJCET/article/view/IJCET_15_06_037.

[5] P. Körner, J. Glüge, S. Glüge, and M. Scheringer, “Critical insights into data curation and label noise for accurate prediction of aerobic biodegradability of organic chemicals,” Environ. Sci.: Processes Impacts, vol. 26, no. 10, pp. 1780–1795, 2024, doi: 10.1039/d4em00431k.

[6] Philip Stahmann, Maximilian Nebel, Christian Janiesch, “AI-based real-time anomaly detection in industrial engineering: A structured literature review, taxonomy, and research agenda”, Computers & Industrial Engineering, Volume 207, 2025, 111236, ISSN 0360-8352. Doi: https://doi.org/10.1016/j.cie.2025.111236.

[7] C. Pealat, G. Bouleux and V. Cheutet, "Improved Time-Series Clustering with UMAP dimension reduction method," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 5658-5665, doi: 10.1109/ICPR48806.2021.9412261.

[8] A. A. Neloy and M. Turgeon, “A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs,” Machine Learning with Applications, vol. 17, p. 100572, Sept. 2024, doi: 10.1016/j.mlwa.2024.100572.

[9] G. Zhu, H. Zhao, H. Liu and H. Sun, "A Novel LSTM-GAN Algorithm for Time Series Anomaly Detection," 2019 Prognostics and System Health Management Conference (PHM-Qingdao), Qingdao, China, 2019, pp. 1-6, doi: 10.1109/PHM-Qingdao46334.2019.8942842.

[10] I. Lomov, M. Lyubimov, I. Makarov, and L. E. Zhukov, “Fault detection in Tennessee Eastman process with temporal deep learning models,” Journal of Industrial Information Integration, vol. 23, p. 100216, Sept. 2021, doi: 10.1016/j.jii.2021.100216.

[11] M. Lee and K. Min, “A Comparative Study of the Performance for Predicting Biodegradability Classification: The Quantitative Structure–Activity Relationship Model vs the Graph Convolutional Network,” ACS Omega, vol. 7, no. 4, pp. 3649–3655, Jan. 2022, doi: 10.1021/acsomega.1c06274.

[12] I. Elía and M. Pagola, “Anomaly detection in Smart-manufacturing era: A review,” Engineering Applications of Artificial Intelligence, vol. 139, p. 109578, Jan. 2025, doi: 10.1016/j.engappai.2024.109578.

[13] I. Monroy, G. Escudero, and M. Graells, “Anomaly detection in batch chemical processes,” Computer Aided Chemical Engineering. Elsevier, pp. 255–260, 2009. doi: 10.1016/s1570-7946(09)70043-4.

[14] C. D. Hubbs, C. Li, N. V. Sahinidis, I. E. Grossmann, and J. M. Wassick, “A deep reinforcement learning approach for chemical production scheduling,” Computers & Chemical Engineering, vol. 141, p. 106982, Oct. 2020, doi: 10.1016/j.compchemeng.2020.106982.

[15] E. J. Braham, J. M. Ruddock, and J. O. Hardin, “Generating and leveraging explanations of AI/ML models in materials and manufacturing research,” Patterns, p. 101340, Aug. 2025, doi: 10.1016/j.patter.2025.101340.

[16] M. Mowbray, M. Vallerio, C. Perez-Galvan, D. Zhang, A. Del Rio Chanona, and F. J. Navarro-Brull, “Industrial data science – a review of machine learning applications for chemical and process industries,” React. Chem. Eng., vol. 7, no. 7, pp. 1471–1509, 2022, doi: 10.1039/d1re00541c.

[17] OECD (1992), *Test No. 301: Ready Biodegradability*, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, https://doi.org/10.1787/9789264070349-en.

[18] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[19] Kadi4Mat is the Karlsruhe Data Infrastructure for Materials Science, an open source software for managing research data. Online - https://kadi4mat.readthedocs.io/en/stable/.

[20] Polymer Database "PoLyInfo" systematically provides various data required for polymeric material design, available online - https://polymer.nims.go.jp/

[21] K. A. Fransen et al., “High-throughput experimentation for discovery of biodegradable polyesters,” Proc. Natl. Acad. Sci. U.S.A., vol. 120, no. 23, May 2023, doi: 10.1073/pnas.2220021120.

[22] An online database - https://plasticdb.org/

[23] P. C. F. Buchholz et al., “Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database—PAZy,” Proteins, vol. 90, no. 7, pp. 1443–1456, Feb. 2022, doi: 10.1002/prot.26325.

[24] PAZy online - https://pazy.eu/doku.php?id=start

[25] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni. "QSAR biodegradation," UCI Machine Learning Repository, 2013. [Online]. Available: https://doi.org/10.24432/C5H60M. Dataset available at https://archive.ics.uci.edu/dataset/254/qsar+biodegradation.

[26] Deuschel, A. Foltyn, K. Roscher, and S. Scheele, “The role of uncertainty quantification for trustworthy AI,” in Unlocking Artificial Intelligence, C. Mutschler, C. Münzenmayer, N. Uhlmann, and A. Martin, Eds. Cham: Springer, 2024, pp. 65–79, doi: 10.1007/978-3-031-64832-8_5.