GA – PROJECT NUMBER:	**101158046**
PROJECT ACRONYM:	AUTOMATA
PROJECT TITLE:	AUTOmated enriched digitisation of Archaeological liThics and cerAmics
CALL/TOPIC:	HORIZON-CL2-2023-HERITAGE-ECCCH-01-02
TYPE OF ACTION	HORIZON RIA
PRINCIPAL INVESTIGATOR	Prof Gabriele Gattiglia, UNIPI
TEL:	+39 050 2215228
E-MAIL:	gabriele.gattiglia@unipi.it

This project has received funding from the European Union's HORIZON RIA research and innovation programme under grant agreement N. 101158046

D4.1 Algorithms and procedures for automatic calibration of the robotic automation system

Version: 1.0

Work Package:	4 – Robotic automation system development - Sensing Integration and Simulation
Lead Author (Org):	Nevio, Dubbini (MIN)
Contributing Author(s) (Org):	Daniël P. van Helden (KCL), Martina Naso (UNIPI), Claudia Sciuto (UNIPI), Heeli Schechter (HUJ), Gabriele Gattiglia (UNIPI), Arthur Leck (UBM), Clement Houbert (INRIA)
Due Date:	M20
Date:	30/04/2026

Dissemination Level		
P	**Public**	**X**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Description
0.1	25/03/2026	Nevio Dubbini (MIN)	Structure and draft content
0.2	10/04/2026	Daniël van Helden (KCL)	Content added
0.3	20/04/2026	Claudia Sciuto (UNIPI)	Content added
0.4	23/04/2026	Martina Naso (UNIPI), Arthur Clement (UBM), Clement Joubert (INRIA)	Content added
0.5	24/04/2026	Daniël van Helden (KCL)	Content added
0.6	28/04/2026	Heeli Schechter (HUJ)	Content added
0.7	30/04/2026	Gabriele Gattiglia (UNIPI)	Revision of the document
0.8	05/05/2026	Nevio Dubbini (MIN), Daniël van Helden (KCL)	Content added
0.9	06/05/2026	Gabriele Gattiglia (UNIPI)	Final revision of the document

Disclaimer

Sommario

Abbreviations

WP: Work package

M: Month

UNIPI: Università di Pisa

UBM: Université Bordeaux Montaigne

UoY: University of York

INRAP: Institut National de Recherches Archéologiques Préventives

INRIA: Institut national de recherche en sciences et technologies du numérique

AMZ: Arheoloski Muzej u Zagrebu

QB: QBrobotics Srl

HUJ: The Hebrew University of Jerusalem

MIN: Miningful srls

KCL: King's College London

IIT: Fondazione Istituto Italiano di Tecnologia

UB: Universitat de Barcelona

CL: Culture Lab

Executive summary

This deliverable presents the algorithmic framework developed within Task 4.2 for calibration monitoring, quality assessment, and adaptive acquisition control in the AUTOMATA sensing workflow. The work focuses on the interpretation and validation of instrument outputs produced by the sensing modalities integrated into the project: 3D photogrammetric modelling, hyperspectral imaging, X-ray fluorescence, and Raman spectroscopy.

In this deliverable, automatic calibration is understood as an algorithmic layer that operates between standard instrument-level calibration and robotic acquisition control. It does not replace manufacturer-defined or instrument-specific calibration procedures, which remain a prerequisite for data acquisition. Instead, it assesses whether the data produced under calibrated conditions are statistically consistent, analytically informative, and suitable for integration into the enriched digitisation workflow.

The proposed pipeline follows a common structure across all sensing modalities. It includes a reference stage, a feature-space stage, exploratory analysis, classification, explainable artificial intelligence, and feedback. This architecture allows the system to compare observed acquisition behaviour with expected reference distributions, identify deviations, and associate them with possible causes such as calibration drift, acquisition inconsistencies, noise, artefact illumination, spectral distortions, or reconstruction failures.

For 3D modelling, the pipeline evaluates geometric and topological properties of photogrammetric reconstructions, including mesh integrity, boundary conditions, face-area distributions, dihedral-angle behaviour, and other descriptors related to reconstruction quality. For hyperspectral imaging, the framework supports the identification of unreliable bands, pixels, or regions affected by shadows, specular reflections, low signal, or artefact illumination. For XRF and Raman spectroscopy, the pipeline analyses spectral quality through indicators such as signal-to-noise ratio, peak detectability, baseline behaviour, anomaly scores, and model uncertainty.

A central component of the deliverable is the feedback stage. Diagnostic outputs are translated into potential operational responses, including acceptance, parameter adjustment, reacquisition, rejection, or human intervention. This mapping remains conservative: automatic correction is limited to deviations whose likely cause is identifiable and linked to controllable acquisition parameters. Ambiguous or severe deviations are escalated to reacquisition or expert review, ensuring that automation remains compatible with the complexity of archaeological artefacts and archaeometric data.

The deliverable also addresses data efficiency and resource-aware operation. Acquisitions identified as non-informative or severely degraded can be excluded from persistent storage, while marginal but still informative cases may be retained only in processed or reduced form. This reduces storage requirements, limits redundant data transfer, and decreases downstream computational load, contributing to the sustainability and scalability of the AUTOMATA workflow.

D4.1 therefore establishes the methodological and algorithmic basis of the closed-loop calibration framework. At this stage, the pipeline serves as a decision-support and baseline-validation framework, tested on available multimodal datasets. Future work will expand validation under controlled robotic acquisition conditions, refine the classification of failure modes, and connect diagnostic outputs to the robotic and sensor-control layers. In parallel, the representation of validated outputs and their integration into enriched 3D models will be further developed in connection with D5.3, Reference Enriched 3D Data. These developments will progressively transform the present framework into an increasingly automated calibration loop for enriched archaeological digitisation.

1 Introduction

This deliverable concerns the development and validation of machine learning methods for calibration monitoring and interpretation of instrument outcomes within the project framework. The work is carried out mainly under Task 4.2 (Machine Learning for sensors and lighting control and calibration), devoted to automatic calibration procedures, understanding instrument outputs and assessing their reliability in the context of archaeological artefact digitisation. The approach adopted in this task is based on the systematic exposure of artefacts to the available sensing tools, followed by rapid post-processing of the acquired data. In this perspective, automatic calibration refers to an algorithmic layer for quality assessment and adaptive acquisition control. It does not replace standard instrument-level calibration, which remains a prerequisite, but verifies whether calibrated outputs are statistically consistent, analytically informative, and suitable for the AUTOMATA workflow. Detected deviations are translated into corrective recommendations, including parameter adjustment, reacquisition, data rejection, or human intervention. Data management strategies are also implemented to avoid retaining non-informative or degraded signals. In particular, acquisitions that do not meet minimum quality criteria (e.g., low signal-to-noise ratio, strong noise contamination, spectral distortions) can be discarded, while, in other cases, only processed representations of the data (e.g., smoothed or denoised signals) are retained instead of the full raw measurements. This ensures that subsequent analysis is performed on data that preserves meaningful information. In addition, this approach contributes to improved computational efficiency and sustainability by reducing data storage requirements, limiting unnecessary data transfer, and decreasing the energy consumption associated with data processing and long-term storage. D4.1 thus establishes the methodological basis of the closed-loop calibration framework, to be progressively integrated with the robotic control architecture in subsequent tasks.

This deliverable is part of WP4, addressing the integration of sensing technologies for archaeological artefact manipulation, analysis, and digitisation. WP4 focuses on enabling robust acquisition and processing of heterogeneous sensor data within the robotic system. In this context, machine learning algorithms play a central role in supporting the interpretation and validation of sensor outputs, improving data reliability and mitigating possible calibration errors.

Within WP4, Task 4.1 addressed the development of perception, sensing, and digitisation capabilities for the robotic working cell. The outputs of Task 4.1 provided the primary data sources whose quality and reliability are further processed within Task 4.2. In parallel, Task 4.3 focuses on simulating the deployable acquisition system once the soft robotic system has been defined. Based on system specifications, simulation models will be developed to evaluate acquisition performance across a range of artefacts and configurations. These simulations provide a controlled environment for assessing sensing strategies and system behaviour, and their results are compared to experimental measurements generated in subsequent work packages. Task 4.4 will address the creation of the operative system responsible for controlling the robotic platform and coordinating the functioning of sensors and tools. This will include implementing a control architecture based on reconfigurable hardware components and modular software, supported by a skill library encoding typical scanning operations and associated equipment. Finally, Task 4.5 will focus on integrating hyperspectral scanning into the overall acquisition workflow.

Within this framework, Task 4.2 provides the methodological foundation for ensuring the reliability, consistency, and accuracy of sensor data across the entire work package. By enabling calibration and machine learning-based understanding of measurements, it supports the overall objective of achieving high-quality, efficient, and repeatable digitisation of archaeological artefacts.

1.1 State of the art

The design of automatic calibration and quality assessment procedures for multi-modal sensing systems requires an analysis of existing methodologies for data validation, error quantification, and reliability assessment. Within the scope of the AUTOMATA framework, this involves photogrammetric 3D reconstruction, spectroscopic techniques (hyperspectral imaging, HSI), X-ray fluorescence (XRF), and Raman spectroscopy, each characterised by distinct signal structures, noise sources, and calibration constraints.

The definition of "quality" for 3D models is highly dependent on their intended purpose (Bryndza et al. 2024; Montusiewicz et al. 2026). The various roles and applications of 3D models in cultural heritage and archaeology, i.e. documentation, preservation, research, education, outreach and more, create intrinsic tension between visualisation, geometric accuracy and archival needs, making the assertion of model quality case-specific or context-dependent (Costopoulos and Papaioannou 2026; Hernández-Muñoz 2023; Malik et al. 2021). Most current attempts to define a "good" 3D model in heritage contexts concentrate on comparing acquisition technologies and procedures and assessing model quality and purpose suitability along quantitative and qualitative parameters (e.g., Menna et al. 2016; Montusiewicz et al. 2026; Polo et al. 2022). While comparing mesh density and resolution, file size, metric accuracy, or texture fidelity of 3D models are common, model construction quality and integrity are barely addressed.

Photogrammetry was selected as the 3D data acquisition technology for the AUTOMATA project (see Deliverable 5.2). Assessing the quality of 3D models from photogrammetry is a complex problem, especially on the fly (see Lou *et al*. 2025 for a discussion of this in the context of Unmanned Aerial Vehicle photogrammetric acquisition) and in the absence of digital comparisons (di Filippo *et al*. 2024). No standard currently exists for quantifying errors and assessing quality in photogrammetric 3D modelling (di Filippo *et al*. 2024, Sorgente *et al*. 2023). Much of the existing literature focuses on assessing the correspondence between reconstructed 3D models and the physical reality they represent, or on benchmarking photogrammetric outputs against higher-accuracy acquisition techniques (Daneshmand et al. 2018). In parallel, a substantial body of work investigates automatic or algorithmic detection of reconstruction errors, often in conjunction with methods for repairing or improving degraded 3D models (e.g. Charton, Baek, and Kim 2021; Sfikas, Perakis, and Theoharis 2022; Zhang, Zhou, and Duan 2023).

Calibration in spectroscopic techniques (HSI, XRF, Raman) operates at two distinct but complementary levels: radiometric (or instrumental) calibration and quantification (or analytical) calibration (Frahm 2024). Radiometric calibration concerns the physical response of the instrument and the quality of the acquired signal. Its purpose is to convert raw data into physically consistent and comparable measurements by correcting for sensor effects and acquisition conditions. This includes, for example, dark and white reference corrections in HSI, energy scale alignment and detector response in XRF, and wavenumber and intensity corrections in Raman. This step ensures that spectral features are correctly positioned and that signal intensities are stable and comparable. Quantification calibration, by contrast, translates the calibrated signal into chemical or compositional information. It establishes the relationship between signal intensity and material properties, such as elemental concentrations, mineral or material composition, or molecular compounds. This step relies on empirical standards or modelling approaches and is dependent on the material system and analytical context.

For HSI, radiometric calibration ensures the physical validity of reflectance values. Radiometric calibration is performed by acquiring white and dark reference measurements. White reference calibration, typically based on certified reflectance standards (e.g., Spectralon targets), enables conversion of raw digital numbers into relative or absolute reflectance values. These steps should be repeated at regular intervals and whenever acquisition conditions change, ensuring consistency across datasets. Within the AUTOMATA workflow,

radiometric calibration procedures are adapted to the specific architecture and acquisition logic of the hyperspectral sensors employed, notably the Specim IQ and the Hinalea 4250 VNIR camera (see AUTOMATA Deliverables D2.1, D2.2 and D2.3).

For the Specim IQ, radiometric calibration is largely embedded within the instrument's internal processing chain. The sensor performs automated acquisition of dark reference data (via internal shutter) and requires the user to acquire a white reference image, typically using a calibrated reflectance panel (e.g., Spectralon). The calibration is therefore dependent on a correct reference acquisition prior to scanning. The Hinalea system operates differently, with a more modular and externally controlled calibration workflow. A "dark" is necessarily recorded before a series of analyses. The "full frame" white can be recorded at the same time or later during the analysis series. It allows calibration either with a "full frame" white, which is more precise and takes into account illumination variations for each pixel of the image, or with a white calibrated from a small calibrated reflectance panel present in each image, which is less precise but less sensitive to potential changes in lighting. Calibration is generally applied in post-processing, allowing greater flexibility in correcting for illumination heterogeneity, sensor drift, and acquisition inconsistencies, with a stronger emphasis on reproducibility and metadata completeness.

To mitigate drift effects and errors related to changes in illumination - one of the most critical sources of variability in hyperspectral acquisition - AUTOMATA adopts a harmonised strategy that can be used with both systems. In particular, the inclusion of a calibrated reflectance target (e.g., Spectralon) within the scene during acquisition is considered the most robust solution from a data reliability perspective. This approach, widely used in field-based archaeological HSI applications, ensures that reflectance values remain comparable across datasets and acquisition sessions (Sciuto et al., 2022). Once radiometric calibration and initial quality assessment, based on reference targets, have been completed, a second level of qualitative spectral evaluation can be undertaken. Even in well-calibrated hyperspectral datasets, not all pixels carry meaningful analytical information. Some spectra may be compromised by acquisition artefacts, such as specular reflections on highly reflective surfaces or low-signal regions associated with shadowed areas. This issue is particularly relevant in archaeological contexts, where objects often exhibit irregular morphologies and heterogeneous surface properties. Variations in geometry can produce localised illumination effects, while specific materials - such as certain lithic raw materials with reflective surfaces (e.g., siliceous rocks) or ceramic coatings (e.g., glazes and slips) - can introduce spectral distortions. As a result, these pixels may display anomalous reflectance values, reduced signal-to-noise ratios, or non-representative spectral signatures (Hubbard et al., 2004; Linderholm et al., 2019). Usually, this type of issue can be effectively mitigated during manual processing by selectively focusing on reliable pixels and regions of interest (ROIs). Even when calibrated datasets contain artefacts due to shadows or specular reflections, the operator can identify and exclude compromised pixels based on their spectral behaviour (e.g., anomalous reflectance values, low signal-to-noise ratio, or non-physical spectral shapes). In practice, this involves defining ROIs that correspond to homogeneous and diagnostically meaningful surface areas, avoiding zones affected by illumination inconsistencies or reflective effects (Grahn and Geladi, 2007). The manual selection of Regions of Interest (ROIs) in hyperspectral imaging (HSI) is considered a data analysis step rather than part of the acquisition or calibration process.

Radiometric calibration for XRF focuses on correcting any shift in energy bins, checking that the detectors are correctly sorting X-rays into the correct energy. The calibration procedure can vary across sensors, but it's always based on the measurement of a known standard and the automatic recalibration of peak drift at known concentrations (Frahm, 2024). For the Olympus Vanta pXRF, radiometric calibration is achieved through a combination of automated internal checks and periodic external verification procedures. A key step is the execution of the built-in CalCheck routine, typically performed using a reference material such as

stainless steel 316. The procedure is the same for the SciAps X-550, except that the energy calibration must be carried out every time the device is restarted. During this process, the instrument identifies characteristic X-ray peaks (notably Fe Kα at 6.4 keV and Mo Kα at 17.5 keV) and applies a two-point calibration to adjust the detector gain and any energy offset, ensuring correct peak positioning. In addition, the system evaluates critical performance parameters, including tube output, count rates, and spectral resolution. Low-quality XRF spectra can usually be identified by diagnostic criteria such as abnormally low total count rates, excessive noise or irregular baseline, and poorly defined or shifted characteristic peaks.

In AUTOMATA, the XRF component of the pipeline focuses on the quantitative assessment of spectral data quality by combining physically grounded indicators, statistical descriptors, and machine-learning-based classification. The objective is to characterise the variability in XRF measurements acquired from archaeological artefacts, identify acquisition and calibration issues, and enable automated discrimination between reliable and problematic spectra. Because the system is meant to process large volumes of data of variable composition automatically, the assessment is designed to function with as few assumptions about composition as possible. Where feasible, information about composition will, of course, be integrated, as this provides an indispensable foundation for calibration (see Frahm 2024). Our main focus for the assessment of spectrum quality will, however, be on analysing the spectra themselves. This also fits in a trend in the field of spectral analysis of increasingly data (or spectra) driven approaches (Andric *et al*. 2024; Al-Tameeni *et al*. 2026). Such a shift relies on clean spectra, an assumption that does not always hold in real-world contexts. For this reason, significant effort has been put into developing methods to preprocess spectra to quantify and remove noise from a range of sources from the desired signal (Yan 2025). Such techniques can be used to inform automatic distinctions between usable spectra and problematic ones.

Raman spectroscopy requires calibration at two levels: wavenumber (or frequency shift) accuracy and intensity response. Wavenumber calibration ensures that spectral peaks are correctly positioned along the shift axis, which is essential for reliable material identification. Intensity calibration corrects for variations in detector response and illumination, ensuring that relative band intensities are reproducible across sessions and instruments. Despite the existence of several guidelines and standards, no unified protocol currently covers the complete calibration process for both spectral axes (Lellinger *et al.* 2025). In practice, wavenumber calibration is commonly performed using reference materials with well-characterised Raman spectra. For example, polystyrene is a widely adopted reference material for this purpose, owing to its physical and chemical stability and the availability of well-characterised peak positions standardised in ASTM E1840 (Itoh and Hanari, 2021); similarly, calcite is commonly used for spectral resolution assessment, with its peak at 1085 cm^{-1} providing a standard benchmark across instruments (Lellinger *et al.,* 2025). The resulting spectrum is evaluated against expected peak positions and intensities: deviations beyond defined tolerances indicate calibration failure and flag the session for review or recalibration. When a systematic and constant wavenumber shift is detected, the calibration data can, in principle, be used to apply a correction offset to the acquired spectra. The choice and correct handling of reference materials is particularly critical for portable and handheld instruments, where lower spectral resolution and signal-to-noise ratios demand closer attention to calibration quality (Lellinger *et al.* 2025). As with XRF, low-quality Raman spectra can be identified through diagnostic criteria such as anomalously low signal intensity, elevated fluorescence background, or peak displacement. These criteria are well established in the applied literature, where noise (understood as spectral fluctuations unrelated to the Raman or fluorescence signal) has been shown to distort both peak height measurements and baseline definition (Madden *et al.* 2018). This supports the adoption of interval-based rather than single-point peak metrics in automated quality assessment workflows, as averaging across a narrow spectral range around a peak of interest reduces the sensitivity of measurements to local noise fluctuations.

2 Calibration pipeline

The overall calibration and quality assessment pipeline is based on a common methodological scheme that is applied across all sensing and acquisition modalities considered in this deliverable, namely 3D modelling, HSI, XRF, and Raman spectroscopy. The pipeline is designed to support the systematic evaluation of currently available acquisitions, the extraction of descriptive and discriminative statistical information, and the progressive development of automatic classification tools for quality control and calibration support.

In the context of this deliverable, the term calibration algorithms does not refer to the direct physical or instrumental calibration of the sensors, such as factory calibration, radiometric correction procedures, energy-scale calibration, wavelength calibration, or hardware-level adjustment. Instead, it refers to a set of data-driven methods aimed at verifying, monitoring, and assessing whether the system behaves as expected once standard calibration procedures have been applied. It is assumed that, prior to any acquisition session, the sensing devices have already undergone the appropriate standard calibration procedures, and that internal calibration routines provided by the instrument manufacturers (e.g., embedded radiometric corrections, energy scale alignment, or wavelength calibration) are correctly applied and functioning as intended. These elements are therefore considered as given and constitute the baseline operating condition of the system. Accordingly, the focus of this work is on calibration monitoring, signal consistency, acquisition reliability, and data-quality assessment. Quantification or analytical calibration, namely the conversion of instrumental responses into material-specific compositional or molecular information, will be addressed separately within WP9, where appropriate reference standards and material-specific validation procedures will be considered. The proposed approach is therefore best understood as an operational layer between instrument calibration and robotic acquisition control. It evaluates whether data produced under calibrated conditions are sufficiently stable, informative, and compatible with the requirements of enriched digitisation by comparing their statistical properties with reference distributions derived under nominal conditions. Calibration is thus interpreted as a consistency check between expected and observed system behaviour. When deviations are detected, through deterministic quality indicators, statistical descriptors, or machine learning-based classification, the system can infer calibration drift, acquisition inconsistencies, or suboptimal operating conditions. This is particularly relevant in AUTOMATA, where heterogeneous sensors, archaeological materials, acquisition geometries, and environmental conditions must be handled within a single automated workflow.

The acquisition process is expected to evolve in subsequent phases of the project, particularly with the introduction of robotic and more controlled acquisition strategies. This transition, together with potential improvements in sensing, calibration procedures, and reconstruction algorithms, is likely to modify the statistical properties of the generated data. As a result, the distributions of key quality indicators, as well as the relationships between features, may change over time. For this reason, the current statistical models and classification approaches should be considered as an initial baseline, valid for the presently available data but not necessarily directly transferable to future configurations. The proposed pipeline is therefore designed to be adaptive, allowing the continuous integration of new data, the update of statistical descriptors, and the re-training or fine-tuning of machine learning models. This ensures that the system remains consistent with the evolving acquisition conditions and maintains its effectiveness in identifying quality issues and supporting calibration across different operational scenarios. To operationalise this approach, the pipeline is structured into stages, each addressing a specific aspect of the calibration verification problem. This staged organisation is adopted to provide a clear, modular framework that facilitates both methodological consistency across different sensing modalities and implementation flexibility.

2.1 Reference stage

The first stage is concerned with establishing a baseline representation of the expected system behaviour under nominal acquisition conditions. This baseline constitutes the reference against which all subsequent acquisitions are evaluated and is fundamental for both quality assessment and calibration. The reference stage is initiated at the beginning of each acquisition session by measuring a predefined set of reference objects for each sensing modality. These reference objects are assumed to produce consistent outputs under correct acquisition and calibration conditions. The underlying assumption is that, within a defined tolerance, repeated acquisitions of the same reference under comparable conditions should yield statistically consistent results. The specific characteristics of these reference objects are detailed in the modality-specific sections. The acquired reference data are analysed directly at the level of raw measurements and primary quality indicators, without relying on the full-feature-space construction used for general acquisitions. Given that reference objects are expected to generate highly stable and repeatable signals, the analysis focuses on the computation of deterministic indices and low-level statistical descriptors that quantify deviations from expected behaviour. These measurements are used to define empirical distributions and tolerance bounds, which serve as the baseline model of the system. Any significant deviation from these reference distributions is interpreted as a potential indication of calibration drift, acquisition inconsistency, or sensor malfunction, and can trigger corrective actions or system reconfiguration before proceeding with the acquisition of actual artefacts.

2.2 Feature space stage

The second stage of the pipeline is dedicated to constructing a structured feature space by systematically extracting deterministic quality indices and statistical descriptors from each acquisition. The computation of deterministic quality indices, which quantify intrinsic properties of each acquisition, is designed to capture modality-specific characteristics of the data, such as structural consistency, signal stability and overall integrity of the measurement. Deterministic metrics provide direct and interpretable indicators of acquisition quality and constitute the primary layer of feature definition. These indices are complemented by statistical descriptors computed over the underlying measurements. The combination of deterministic indices and statistical descriptors defines a high-dimensional feature vector for each acquisition. This representation encodes both local and global properties of the data, allowing the characterisation of variability not only within individual samples but also across the dataset. A key aspect of this stage is the inclusion of both high-quality (“good”) and suboptimal or degraded (“bad”) acquisitions in the dataset. This enables the feature space to capture a good spectrum of observed behaviours, including normal operating conditions and various types of failure modes. As a result, the constructed feature space supports both discriminative and descriptive analyses, providing the foundation for subsequent classification and anomaly detection tasks.

2.3 Exploratory analysis stage

Following the construction of the feature space, an exploratory analysis stage is introduced to systematically investigate the statistical structure and behaviour of the extracted features. This stage plays a critical role in bridging feature computation and model development, providing both quantitative and qualitative insight into the data. At this stage, feature distributions are aggregated and analysed across the full dataset, including both “good” and “bad” acquisitions. Global statistical characterisation is performed to identify typical value ranges, variability patterns, and distributional properties of each feature. A key objective of this stage is to identify statistical irregularities and anomalous patterns. Outlier detection techniques are applied both at the univariate and multivariate levels, using methods such as threshold-based detection, distance-

based metrics, and density-based approaches. In addition, correlation and covariance analyses are performed to uncover dependencies between features and to identify redundant or non-informative variables. Visual analytics constitutes an essential component of the exploratory stage. Feature distributions, pairwise relationships, and low-dimensional projections (e.g., via principal component analysis or other dimensionality reduction techniques) are systematically visualised to support interpretation. These visualisations are particularly valuable in borderline or ambiguous cases, where automated classification may be uncertain. From a methodological perspective, this stage is useful to highlight features with strong discriminative power and identify those that may require transformation, normalisation, or removal. Moreover, this stage contributes to defining reference distributions and tolerance intervals, complementing the reference stage with dataset-wide statistical context. Finally, it provides diagnostic information that can guide the interpretation of model outputs and inform the design of corrective actions within the calibration loop.

2.4 Classification stage

This stage consists of building an initial classification layer on top of the extracted features. Since the objective is to detect whether something is wrong in the acquisition or reconstruction process, the problem can be formulated as a supervised machine learning classification task when labelled examples of good and bad outputs are available. However, because the availability of labelled data is currently limited, unsupervised or semi-supervised artificial intelligence approaches can also be considered, such as clustering, anomaly detection, one-class classification, or distance-based outlier identification. In supervised settings, the model is trained on statistical and engineered features extracted from the datasets and is used to produce an initial discrimination between acceptable and problematic acquisitions. These initial labels are interpreted in relation to known acquisition conditions and modality-specific diagnostic criteria, including reconstruction artefacts for 3D models, radiometric and spatial consistency for HSI, peak detectability and baseline stability for XRF, and signal quality, fluorescence background, and peak behaviour for Raman spectroscopy. In unsupervised or semi-supervised settings, the same feature space is used to identify anomalous, borderline, or internally coherent groups of acquisitions that may support subsequent labelling and model refinement. The initial classification layer is not intended to be final; rather, it provides the basis for progressive improvement as more data become available and as failure modes are better understood. Model performance is assessed through standard statistical and machine learning evaluation procedures, including confusion matrices, accuracy, precision, recall, F1-score, cross-validation, and, where relevant, ROC analysis or calibration curves. Misclassified samples can then be analysed to determine whether the error arises from insufficient features, noisy measurements, inconsistent labels, or modality-specific acquisition problems. This iterative loop supports the refinement of both the statistical descriptors and the classification strategy.

2.5 XAI stage

The Explainable Artificial Intelligence (XAI) stage focuses on the integration of XAI techniques to provide interpretability and diagnostic insights from the previously constructed feature space and classification models. While the classification stage enables automatic discrimination, it does not provide an understanding of the underlying causes driving these decisions. The XAI stage addresses this limitation by identifying the most relevant features, critical acquisition factors, and potential sources of error contributing to model outputs. From a methodological perspective, the XAI stage operates on top of the trained machine learning models and the structured feature space described in the previous sections. Given that the feature representation encodes deterministic quality indices, statistical descriptors, and distributional properties,

explainability techniques can be directly applied to quantify the contribution of each feature to the classification outcome.

A first level of interpretation is provided through global feature importance analysis, particularly suited for tree-based models, allowing the estimation of the relative importance of each feature in determining the classification decision across the entire dataset. To complement global analysis, local explanation techniques are employed to interpret individual predictions. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are used to decompose the output of the classifier into additive contributions of individual features for each specific sample. This allows the identification of the precise combination of factors that led to a given classification (e.g., why a specific acquisition has been classified as problematic), providing case-by-case diagnostic insight. In addition, feature interaction analysis is performed to capture non-linear dependencies and combined effects between variables. Given the high-dimensional, heterogeneous nature of the feature space, many failure modes are not attributable to a single indicator but rather to interactions among multiple factors. Techniques such as SHAP interaction values or partial dependence plots are used to characterise these relationships and to identify compound failure patterns.

From an operational perspective, the outputs of the XAI stage are directly linked to the calibration objectives of the pipeline. The identification of dominant features and critical factors provides a mapping between observed data deviations and potential causes within the acquisition process. These relationships enable the formulation of interpretable feedback mechanisms, where model explanations are translated into suggested corrective actions on system parameters (e.g., adjustment of exposure time, illumination intensity, acquisition geometry, or sensor configuration). In this sense, the XAI stage acts as a bridge between data-driven classification and the automatic calibration loop, enhancing the transparency and controllability of the system.

Finally, the XAI framework supports model validation and robustness analysis. By analysing feature contributions and decision patterns, it is possible to detect potential biases, spurious correlations, or over-reliance on non-informative features. Continuous monitoring of explanation stability across datasets and acquisition conditions ensures that the learned models remain consistent, interpretable, and aligned with the physical and statistical properties of the sensing modalities.

2.6 Feedback stage

The feedback stage represents the operational component of the calibration pipeline, in which the outputs of the statistical analysis, classification models, and XAI interpretation are translated into concrete system feedback. Based on predefined rules, statistical thresholds, or learned decision models, specific patterns in the feature space can be associated with adjustments of acquisition parameters. In this context, deviations can be categorised by severity, distinguishing between acceptable variations, correctable issues that can be automatically addressed, and critical conditions requiring interruption of the acquisition process and potential human intervention. At the current stage of development, the feedback stage is defined at a conceptual and methodological level and is not yet fully operational. In particular, the decision-making process and its translation into executable commands for the robotic system require further integration with the control architecture developed in subsequent tasks. As a result, feedback responses are, at present, formulated into one of the following 4 classes, for acquisitions: acceptable, automatically correctable, reacquire, or requiring human intervention acquisitions. The full implementation of automated feedback, including real-time decision-making and direct interaction with the robotic platform, will be addressed in Deliverable 7.1, where the integration between the calibration pipeline and the system control layer will be

completed. The feedback stage is designed to operate as part of an iterative closed-loop process. Following the application of corrective actions, newly acquired data are reintroduced into the pipeline, undergoing feature extraction, analysis, and classification. This iterative mechanism enables progressive refinement of acquisition conditions and convergence towards stable and optimal system behaviour. Over time, the accumulation of data from multiple sessions supports the continuous updating of statistical descriptors, thresholds, and decision rules, allowing the system to adapt to changes in instrumentation, environmental conditions, and object characteristics. In addition, the feedback stage contributes to data efficiency and resource-aware operation by integrating quality-driven data retention policies into the calibration loop. Acquisitions identified as non-informative or severely degraded can be excluded from persistent storage, while marginal cases may be retained only in processed form, such as denoised or summarised representations. This reduces storage requirements, redundant data transfer, and downstream computational load, improving the energy efficiency and scalability of the acquisition and analysis pipeline. To make the feedback mechanism operational, each diagnostic output is associated with a potential control response, translating information from feature space into control space. For instance, a low signal-to-noise ratio may trigger changes in exposure, illumination, integration time, or reacquisition; spectral saturation may lead to reduced gain or exposure; baseline instability may require filtering, recalibration, or rejection; geometric discontinuities in a 3D model may indicate incomplete coverage, object movement, or reconstruction failure; and high model uncertainty may require human review. This mapping remains conservative: only deviations linked to identifiable and controllable causes are assigned to automatic correction, while ambiguous or severe deviations are escalated to reacquisition or human intervention.

The formulation using the four operational classes (acceptable, automatically correctable, reacquire, or requiring human intervention) extends the initial binary distinction between “good” and “bad” acquisitions into a structured decision framework that reflects the needs of the calibration loop, enabling the system to differentiate between moderate and severe deviations and to determine whether autonomous correction is feasible. The decision process is based on the integration of three complementary indicators derived from previous stages of the pipeline:

- a statistical deviation score *D(x)*, measuring the distance of the acquisition from reference distributions defined under nominal conditions;
- an anomaly score *A(x)*, quantifying the degree of atypicality of the acquisition with respect to the distribution of known acceptable data;
- a prediction uncertainty score *U(x)*, capturing the confidence of the supervised classification model.

In practical terms, *D(x)* is now computed as a multivariate Mahalanobis distance in feature space, using the reference distribution constructed from acceptable acquisitions. The anomaly score *A(x)* is obtained using an Isolation Forest model trained primarily on acceptable data, allowing the detection of previously unseen or rare failure modes. The uncertainty score *U(x)* is derived from the classifier output and is computed as *1 - max(p)*, where *p* denotes the predicted class probability. These three components are normalised on the training set and combined into a global severity score:

$$S(x) = \alpha D(x) + \beta A(x) + \gamma U(x)$$

where the coefficients are estimated empirically. At the current stage, fully annotated four-class labels are not yet available. For this reason, the decision model is initially trained using a heuristic labelling strategy based on score thresholds and expert-defined rules. This weak-supervision approach allows the system to produce operational decisions from the outset while enabling progressive refinement as more data and validated outcomes become available. Specifically:

- acceptable acquisitions are characterised by low deviation, low anomaly, and low uncertainty;
- automatically correctable acquisitions exhibit moderate deviation associated with known and controllable causes;
- reacquire/reject cases correspond to degraded acquisitions where reliable correction is not feasible but the issue is likely transient or acquisition-related;
- human intervention cases are characterised by high deviation, high anomaly, high uncertainty.

Thresholds separating low, moderate, and high values are estimated from empirical score distributions, for example, using percentile-based criteria derived from acceptable reference data. A multinomial logistic regression model provides a baseline for this decision layer, although alternative models may be considered if required. A central component of the feedback stage is the integration of explainable artificial intelligence (XAI) techniques, which provide a quantitative link between model outputs and system-level actions. While the classification model determines whether an acquisition is problematic, the XAI stage identifies the underlying causes of the deviation and evaluates whether these causes can be addressed automatically. Feature attribution methods, such as SHAP, are applied to quantify the contribution of individual features to the classification outcome. For each acquisition, the dominant contributing features are identified and ranked. In deliverable 7.1, these features will be mapped to a predefined set of controllable acquisition parameters (e.g., exposure time, illumination, sensor gain, acquisition geometry). This mapping enables the definition of an actionability criterion. An acquisition is considered automatically correctable if the dominant contributing features correspond to known and controllable factors. Conversely, acquisitions are assigned to the reacquire or human-intervention classes when the contributing factors are not directly actionable, only partially actionable, or unstable across similar cases. In addition, the structure and concentration of feature contributions are used to assess explanation reliability. Cases characterised by diffuse contributions, high uncertainty, or inconsistent explanations are considered unsuitable for autonomous correction and are therefore escalated to human intervention. The proposed framework is inherently adaptive. As new acquisitions are processed, the outcomes of corrections, reacquisitions, and human interventions are recorded and used to update the training dataset. In this way, the initial heuristic labels are progressively replaced or refined with empirically validated labels, enabling the decision model to evolve from a rule-based approximation to a fully data-driven system.

3 Data Collection

The development and validation of the calibration and quality assessment methodologies described in this deliverable rely on a set of datasets acquired across multiple sensing modalities. Data have been collected from two primary sources: (i) acquisitions performed within the scope of the AUTOMATA project, using the currently available sensing infrastructure; (ii) pre-existing datasets, produced prior to the project under comparable (but not equivalent) experimental conditions. At the current stage of development, data acquisition is performed under manual conditions, rather than through the robotic system that will be developed within AUTOMATA. As a consequence, the available datasets exhibit variability arising from operator-dependent choices, acquisition setup, environmental conditions, and instrument configuration. While this introduces heterogeneity, it is intentionally preserved, as it provides a realistic representation of acquisition variability and supports the identification of robust quality indicators. This design choice is also aligned with one of the main objectives of the AUTOMATA project, namely the development of a modular system capable of integrating and operating with different sensing technologies. The current sample sizes, although limited compared to large-scale automated acquisition scenarios, are sufficient to support the definition, testing, and initial validation of the statistical descriptors and machine learning models presented.

It is important to note that the transition to fully automated acquisition through the robotic platform is expected to introduce additional sources of variability and potential failure modes, including those related to system integration, motion planning, sensor coordination, and real-time constraints. Furthermore, future deployments in different archaeological contexts, involving new materials, environmental conditions, and operational constraints (e.g., portability), will likely affect the statistical properties of the acquired data. For these reasons, the proposed calibration pipeline is explicitly designed to be adaptive and data-driven, allowing continuous integration of new data, updating of statistical models, and retraining of machine learning components. The robustness of the approach, therefore, lies not in the completeness of the current datasets but in the generality and flexibility of the methodological framework. Importantly, the currently available datasets have been structured to include a sufficiently diverse range of acquisition outcomes, encompassing both high-quality (“good”) and suboptimal or degraded (“bad”) cases. This diversity is a fundamental requirement for the methodological framework adopted in this deliverable. In particular, the presence of both classes enables the formulation of supervised learning problems. At the current stage, the labelling scheme is primarily based on a binary distinction between acceptable and problematic acquisitions. However, this represents only an initial abstraction of the underlying variability. As data collection progresses, the labelling process will be progressively refined to incorporate a more granular and semantically rich annotation of failure modes. The multi-class labelling strategy will enable the transition from binary classification to more advanced diagnostic models capable of not only detecting the presence of a problem, but also identifying its nature and potential origin. Such an extension is particularly relevant for calibration purposes, as it provides actionable information for targeted parameter adjustment and system optimisation.

The datasets used in this report are organised according to the four sensing modalities defined within AUTOMATA: 3D modelling, HSI, XRF, and Raman spectroscopy.

The 3D modelling dataset consists of photogrammetric reconstructions of archaeological ceramic and lithic artefacts. Image acquisition was performed using high-resolution photography under controlled or semi-controlled lighting conditions, followed by processing pipelines for the generation of dense point clouds and triangulated mesh models. The photos were acquired using a Nikon Z7 II equipped with a Nikon 24–70 mm f/2.8 lens. Lighting is provided by a Godox AR400 ring flash. The sample holder is a transparent plexiglass disc, 4 mm thick and 30 cm in diameter. It is connected to a motor that allows it to rotate, in order to

reproduce as closely as possible the automated acquisition conditions of Automata. Coded targets were placed on this transparent support so they are visible from both sides. Four small 3D-printed mounts were also placed on each side of the support to hold additional coded targets, ensuring they remain visible even at the most constrained viewing angles (Figure 3.1). The 3D models were then generated from these data using Agisoft Metashape. The dataset includes 13 3D models of ceramic remains and 31 lithics. A further seven models, five of metal objects and two of glass objects, were included to increase the range of photogrammetric failure conditions represented in the dataset, bringing the total to 51 models, of which 44 belong to the ceramic and lithic categories directly addressed by AUTOMATA. These additional objects are not intended to expand the archaeological scope of AUTOMATA but to expose the calibration pipeline to challenging photogrammetric conditions, such as low texture, reflective surfaces, transparency, and fine geometric detail. The current sample includes both high-quality reconstructions and models affected by common failure modes, such as incomplete coverage, non-manifold geometry, noise, misalignment, artificial flat surfaces, poor local mesh quality, and disruption caused by object movement. These failure modes are particularly relevant for photogrammetric acquisition in an automated workflow, where object rotation, camera position, illumination, and support transparency may all affect reconstruction reliability.

***Figure 3.1**: Photogrammetry setup with the rotating support and coded targets*

The HSI dataset consists of spectral image cubes acquired using a Specim IQ push-broom hyperspectral camera (SPECIM Spectral Imaging Ltd., Oulu, Finland), which covers the visible light and near-infrared (Vis–NIR) spectral range from 400 to 1000 nm, with 204 spectral bands across the entire wavelength range. The camera was mounted on a tripod at the Archéosciences laboratory (UBM), and the objects were illuminated by two halogen lamps, oriented at 45°. Three software have been used to read the data once they were acquired: Orange Quasar, Evince, and Envi. White and dark references have been used when available to support radiometric normalisation. The HSI dataset includes 513 artefacts in total, including 421 ceramics and 92 lithics. The dataset includes both high-quality acquisitions and cases affected by noise, illumination non-uniformity, spectral distortions, or sensor artefacts, supporting the construction of feature

representations for machine learning-based quality assessment. The XRF dataset consists of elemental spectra acquired using Evident (Olympus) Vanta C-series portable X-ray fluorescence devices. Each measurement produces a spectrum representing photon counts as a function of X-ray's energy, corresponding to the characteristic emission lines of the elements present in the analysed material. The dataset includes 159 artefacts in total (941 measures), including 119 ceramics and 40 lithics. The dataset includes repeated measurements and spectra of varying quality, reflecting differences in signal-to-noise ratio, peak detectability, and baseline stability. This variability enables the definition of quality indicators for spectral properties and supports the development of statistical and machine learning models for the automatic classification of reliable and problematic spectra. The Raman dataset consists of vibrational spectra acquired using a portable i-Raman Plus 785S portable spectrometer (Metrohm), equipped with a 785 nm excitation laser, a thermoelectrically cooled CCD detector, and a fibre-optic probe. The instrument covers a spectral range of 65–3350 cm^{-1} with a spectral resolution of less than 4.5 cm^{-1}. Each spectrum represents intensity as a function of Raman shift and provides information about the molecular composition of the analysed material. The Raman dataset includes 56 artefacts (201 measures), in total, 46 ceramics and 10 lithics. The dataset includes spectra with varying signal quality, including those affected by fluorescence background, noise, peak broadening, or spectral misalignment.

4 3D modelling

The 3D modelling component of our pipeline focuses on quantitative assessment of reconstruction quality through the combined use of deterministic geometric metrics, statistical descriptors, and machine-learning-based classification. Each reconstruction is evaluated through a set of deterministic quality indices that directly quantify geometric, topological, and photometric properties of the resulting models. These indices are used to monitor the following aspects of the model that affect its quality.

- Mesh density is correlated with model resolution. Higher density meshes (i.e. meshes with more faces per surface area) can more faithfully model an object's geometry. As model resolution is likely to vary according to the material being investigated and the scientific questions by which this investigation proceeds, face area, while computed (as half the cross product of the edge vectors), is not a primary indicator for error detection. To capture variation in face area, indicating perhaps areas of lower coverage or artificial mesh, the mean face area is supplemented with its standard deviation. This is a useful parameter for detecting deviations in density.
- Edge nature (manifold, boundary, or non-manifold) is an indicator of mesh quality. In a perfect 3D model, each edge is shared by exactly two triangle faces. If an edge is used by only a single face, it indicates a hole in the mesh. If an edge is shared by more than two faces, the mesh is non-manifold. We therefore gather the total number of edges and compute the number of boundary, manifold and non-manifold edges by counting the number of faces shared by each edge.
- The number of vertex fans is a measure of mesh manifoldness. In an ideal 3D model, the faces meeting at each vertex should form a single unbroken surface that one could traverse without gaps or breaks. We therefore ascertain, for each vertex, whether every other vertex that shares an edge with it is connected in a single chain. If this is the case, this vertex has a single fan. We report the number of vertices with multiple fans, and the maximum number of fans per vertex.
- Face triangle shape. Poorly aligned 3D models can often be identified by spurious flat areas where vertices that should have been adjacent to one another end up far away from each other. This results in face triangles being stretched out very thinly. To detect this, we look at the length of edges and the aspect ratios of the face triangles. Across the entire model, we compute the ratio between the maximum edge length and the median edge length. We also compute the local edge length ratio. To do this, we establish an 'expected length' for the adjoining vertices of each edge by taking the mean of the length of the other edges sharing this vertex. The length of every edge is then compared to the mean of the two expected lengths at each of its vertices. Finally, for each triangle face, we compute the ratio between its longest edge and its shortest.
- Dihedral angle measurements. Another metric for finding spurious flat areas is by looking at dihedral angles. By comparing the face normals of the two faces joined by each edge, we can compute the angle between the two faces. Angles near the extremes of 0 and 180 degrees indicate near co-planarity. We output the minimum, mean, and maximum angles per mesh as well as the mean, max, and 95th percentile of the per-vertex variance in dihedral angle (across each edge that shares this vertex). These are supplemented by the flat dihedral ratio, which is the ratio between 'flat' dihedral angles and the total number of dihedral angles. Dihedral angles are considered flat if they are within 0.1 radians of 0 or π.

Beyond deterministic metrics, statistical descriptors are computed over these quantities to capture their distributional properties across the dataset. For each reconstruction, a structured and high-dimensional feature vector is constructed by combining deterministic geometric indices with detailed statistical and

distributional descriptors computed over the underlying measurements. The feature extraction process includes the following variables.

- Classical summary statistics. Mean, median, variance, standard deviation, skewness, and kurtosis are computed as baseline descriptors of central tendency, dispersion, and distribution shape.
- Quantile-based descriptors. A set of percentiles (e.g., 5th, 10th, 25th, 50th, 75th, 90th, 95th) is extracted to capture the distribution more robustly. Interquartile range (IQR) and percentile spreads (e.g., P90–P10) are included to quantify dispersion.
- Histogram-based features. The empirical distribution of each metric is discretised into fixed bins, producing histogram representations that capture the shape of the distribution. These can be used directly as features or further summarised (e.g., entropy, peak location, modality).
- Density and distribution modelling. Kernel density estimation (KDE) is used to approximate the continuous probability distribution of each metric. From these, additional descriptors such as distribution peaks, bandwidth, and modality indicators are derived.
- Outlier-related features. The proportion and magnitude of outliers are explicitly quantified, for example, by counting samples beyond statistical thresholds (e.g., >3σ, outside IQR-based fences). Maximum and minimum values, as well as extreme quantiles, are included to capture worst-case behaviour.
- Spatially-aware statistics. Instead of computing only global descriptors, the metrics are also evaluated over local neighbourhoods (e.g., patches, regions, or voxel partitions). From these, statistics of local variability are derived, such as the mean and variance of local standard deviations, or the distribution of local extrema.
- Correlation and cross-metric features. Relationships between different metrics are explicitly modelled. Covariance matrices or pairwise correlation coefficients are included as features, along with derived indicators such as principal components from PCA applied to the feature space.

The resulting feature vector is therefore not a simple collection of scalar summaries, but a multi-level representation that captures distributional, spatial, and relational properties of the reconstruction.

4.1 3D algorithmic calibration overview

Reference stage

In the case of 3D modelling, no standardised artefact can reliably represent the variability of archaeological materials. The reference stage, therefore, relies on deterministic geometric metrics and statistical descriptors rather than on a fixed external standard. In future iterations, this empirical strategy will be strengthened, as robotic acquisition conditions become available, through repeatable benchmarks, calibrated targets, comparison with higher-accuracy acquisition methods, and repeated acquisitions under controlled conditions to assess scale accuracy, mesh completeness, geometric distortion, and reconstruction stability.

Feature space stage

The 3D models are represented through a high-dimensional feature vector that combines the previously described deterministic geometric and topological metrics with a comprehensive set of statistical descriptors computed over them. Its role is twofold: on the one hand, it provides a rich and interpretable description of model quality, supporting the identification of geometric inconsistencies, reconstruction artefacts, and failure modes; on the other hand, it serves as the input for subsequent statistical analyses and machine learning models used for classification and anomaly detection. The plots included in Figure 4.1, related to the

feature-state stage, highlight the most informative aspects of the constructed feature space. Additional visualisations, not included for reasons of space, include: (i) distributions of the individual features composing the feature space; (ii) correlation analyses capturing dependencies between features; and (iii) exploratory clustering results. The current results show that several features form correlated groups, particularly those associated with boundary edges, vertices with multiple fans, face-area metrics, and dihedral-angle variance. The cumulative distributions show that selected individual features, such as the triangle aspect ratio, appear to exhibit different behaviour in the good and bad subgroups. As the dataset grows, the statistical characterisation of these features will be progressively refined, improving the robustness and discriminative power of the feature space, enabling increasingly accurate identification and classification of usable versus problematic 3D reconstructions.

Exploratory stage

The 3D modelling exploratory stage focuses on statistical analysis of the constructed feature space to identify patterns, anomalies, and discriminative structures within the dataset. The figures for this stage, presented in Figure 4.1, including dimensionality reduction projections, outlier detection analyses, and feature ranking based on statistical separability between “good” and “bad” models, provide insight into the clustering behaviour of acquisitions, highlight potential failure modes, and identify the most informative features contributing to quality discrimination. The current exploratory analyses show only partial separation between acceptable and problematic reconstructions when clustering-based methods are used. Anomaly-based methods appear more promising. One-Class Support Vector Machine analysis identifies the top anomalous cases, which align with the ‘bad’ label. However, separation remains less clear for intermediate cases, where both labels overlap. This indicates that the present feature space is already sensitive to severe reconstruction failures, while additional data and a clearer characterisation of borderline cases will be required to improve discrimination in intermediate-quality reconstructions. This stage supports the validation of the feature engineering process and provides guidance for the subsequent design and refinement of classification models.

Classification stage

The classification stage builds upon the extracted feature vectors to train and evaluate machine learning models for the automatic discrimination between reliable and problematic 3D reconstructions. The problem is formulated as a supervised classification task, using Logistic Regression and Random Forests. The figures included in Figure 4.1, related to the classification stage, highlight the performance and behaviour of the trained models, illustrated through standard evaluation metrics, such as confusion matrices, probability distributions of predictions, ROC curves, and calibration plots. For space constraints, it is not possible to include all generated visualisations, which, in general, comprise: (i) performance metrics across different models and validation strategies, (ii) probability and score distributions, and (iii) diagnostic plots supporting model evaluation and comparison. The included plots indicate that while both Random Forest and Logistic Regression capture part of the structure in the data, Logistic Regression performs better than Random Forest, suggesting that the structure is likely to be relatively low-dimensional or of a linear nature. While the current implementation focuses on detection and classification, the outcomes of this stage provide a foundation for the integration of feedback mechanisms.

XAI stage

The XAI stage introduces interpretability into the classification framework by analysing and visualising the contribution of individual features to model predictions. The figures reported for this stage in Figure 4.1

include global feature importance rankings, partial dependence plots, and local explanation methods such as SHAP values. These analyses highlight the most influential features driving the classification decisions and reveal interactions between variables. The analysis indicates that the triangle aspect ratio, identified as likely to be important in earlier stages, is currently contributes substantially to label separability. Such insights enable the identification of causes underlying acquisition issues and support the definition of targeted corrective actions on system parameters.

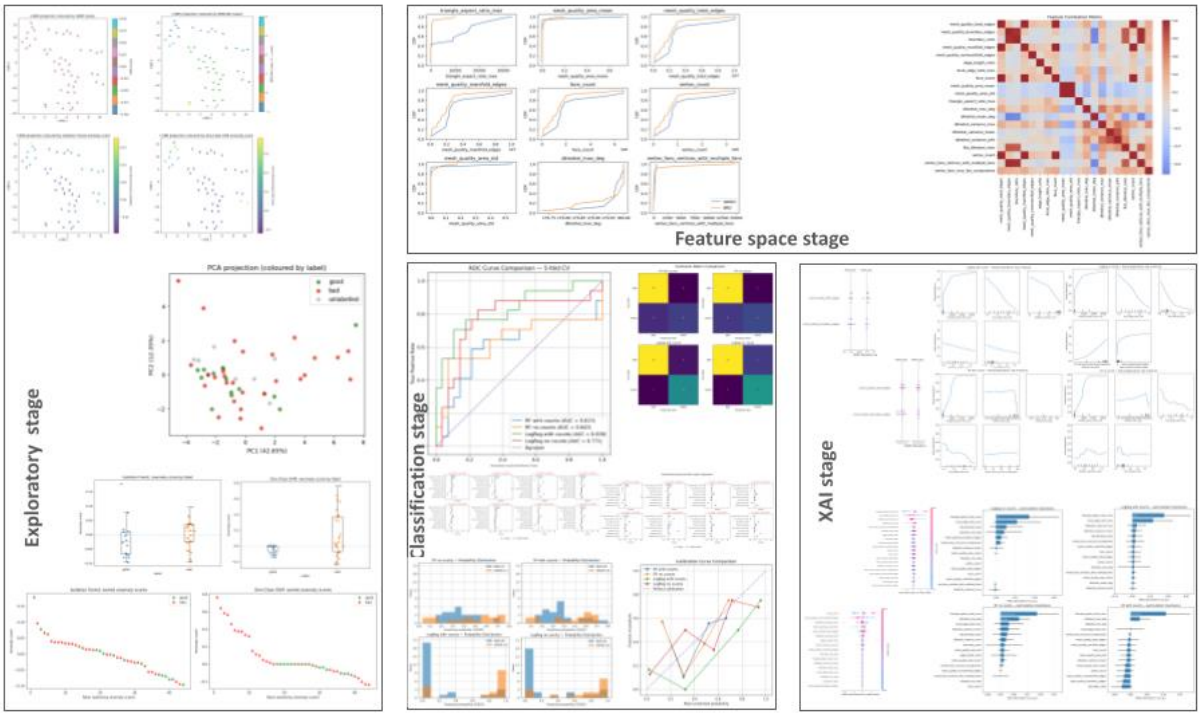

***Figure 4.1**: Overview of the 3D modelling calibration pipeline across its four main stages. Representative plots illustrate the progression from feature space construction to exploratory analysis, classification, and explainability. The pipeline captures the statistical, geometric and topological complexity of the 3D models, enabling separation between acceptable and problematic models, as suggested by the classification performance of Logistic Regression (AUC ≈ 0.83) even on a small dataset. The integration of XAI methods identifies the most influential features, linking classification outcomes to physically meaningful causes.*

Metric	Accuracy	Precision	Recall	F1	AUC
Value	0.8	0.79	0.79	0.79	0.83

***Table 4.1**: Performance evaluation of the 3D Logistic Regression classification model. The metrics provided include Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC), collectively showing the model's capability in quality assessment, with currently available data.*

Feedback stage

The feedback stage represents the operational layer of the calibration pipeline, in which the outputs of the statistical analysis, classification models, and, in particular, the XAI stage are translated into recommendations. At this stage, these recommendations remain at a decision-support level, since the automatic photogrammetric acquisition process and the model-generation workflow are still being integrated. As shown in Figure 4.2, the current formulation defines the decision logic of the feedback stage, while the mapping between feature patterns and corrective actions remains provisional. Nevertheless, this mapping is already informative: boundary or non-manifold edges may indicate incomplete coverage or reconstruction instability; abnormal triangle aspect ratios may suggest misalignment, insufficient image overlap, or artificial surface generation; and dihedral-angle or face-area anomalies may point to poor local mesh reconstruction. As the robotic acquisition workflow becomes operational, these diagnostic outputs will be progressively linked to controllable parameters, such as camera viewpoint, object rotation, image coverage, lighting configuration, exposure settings, and reacquisition. The feedback classes will therefore evolve from the present accept/problematic distinction towards operational flags such as accept, adjust, reacquire, reject, or request human review. This more nuanced guidance will also support resource-efficient data management by informing the level of detail at which the 3D models will need to be retained. IIt may also improve the efficiency of the operation of the AUTOMATA acquisition process. If a 3D model is inadequate for certain post-acquisition analyses but is sufficient for robotic data acquisition to continue, and the issues can be resolved in post-processing, it can be flagged as needing post-acquisition treatment without delaying the acquisition process.

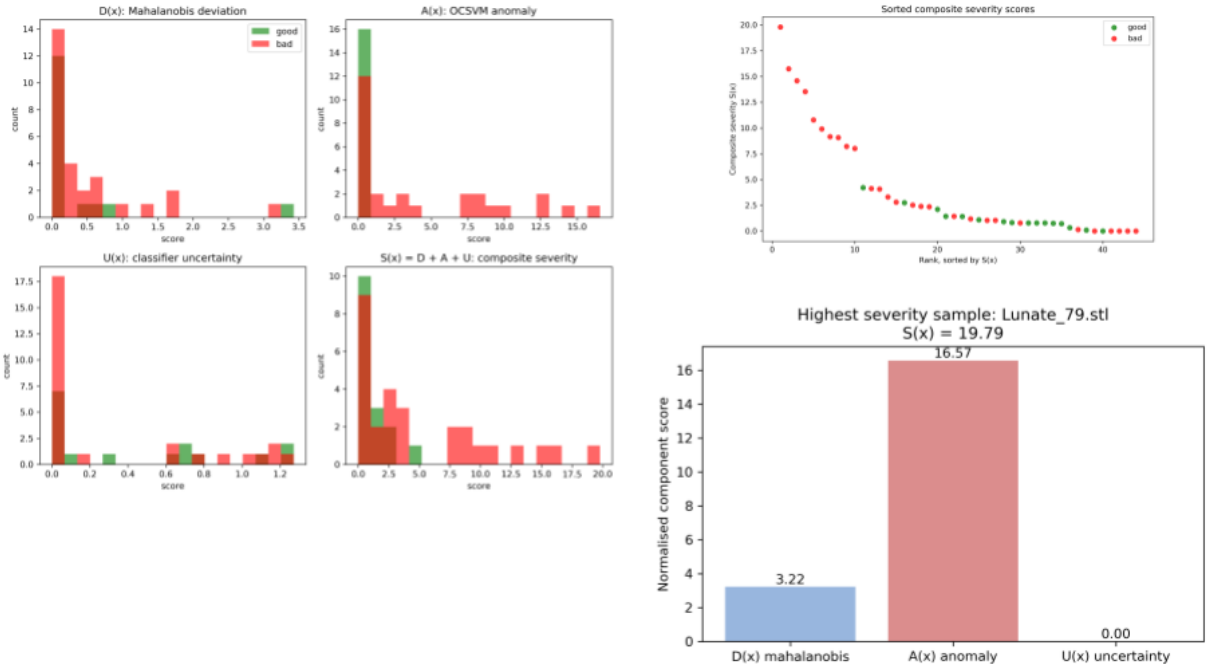

***Figure 4.2**: Feedback stage of the 3D modelling calibration pipeline. The figure illustrates the computation of the composite severity score, informing the separation between the two current decision classes. The bottom right panel presents a breakdown of the components to the highest severity scoring model.*

5 Hyperspectral imaging

The HSI component of the pipeline focuses on the quantitative assessment of spectral data quality by combining physically grounded indicators, statistical descriptors, and machine-learning-based classification. The analysis is conducted on a dataset of already available hyperspectral acquisitions, including both "good" and "bad" cases. The datasets consist of hyperspectral cubes, where each pixel is associated with a spectral signature across a defined wavelength range. The acquisitions are assumed to be performed under controlled conditions, where instrumental parameters such as illumination and sensor configuration are kept approximately constant within a single session but may vary across sessions, while object properties inherently vary across the analysed samples. Each hyperspectral dataset is evaluated using a set of deterministic quality indices that quantify its radiometric, spectral, and spatial characteristics. The following deterministic quality indices are considered.

- Signal-to-noise ratio (SNR) is estimated from the acquired hyperspectral cube based on local statistical analysis, where the signal is computed as the mean intensity within local spatial or spectral neighbourhoods, and the noise is approximated through local variance, high-frequency spectral components, or residuals obtained after local smoothing (e.g., Savitzky–Golay filtering). SNR can be computed per spectral band and aggregated across the spectrum.
- Spectral smoothness and continuity is evaluated by analysing the regularity of spectral signatures across wavelengths using numerical derivatives and signal processing techniques. First and second-order spectral derivatives, total variation, and residuals from local polynomial fitting are used to quantify spectral irregularities.
- Spatial uniformity is assessed by analysing spectral intensity variability across spatial regions using local statistics. The hyperspectral cube is partitioned into patches or superpixels, and within each region, statistical measures such as variance, coefficient of variation, and entropy are computed. Additionally, spatial autocorrelation metrics (e.g., Moran's I or Geary's C) are used to quantify the degree of spatial consistency.
- Illumination consistency is evaluated by analysing the spatial distribution of intensity across the field of view, without requiring explicit reference targets. Global and local intensity gradients are computed per band, and statistical descriptors are used to detect systematic illumination patterns. Additionally, low-frequency components of the image (obtained via filtering or Fourier analysis) are analysed to identify illumination gradients.
- Band correlation and redundancy are computed using covariance and correlation matrices across spectral bands, either globally or within spatial regions. PCA is also used to assess redundancy and the intrinsic dimensionality of the data. Indicators include explained variance ratios and the decay of eigenvalues.
- Dead or saturated bands are identified automatically by analysing the distribution of intensity values per spectral band. Bands with near-zero variance (dead bands) or with values consistently at or near the sensor limits (saturation) are detected through threshold-based analysis of histograms, variance, and dynamic range.

Statistical descriptors are computed over both spatial and spectral domains to capture the distributional properties of the data. Each hyperspectral acquisition is represented as a high-dimensional feature vector combining spectral statistics, spatial statistics, and cross-domain descriptors. The following statistics are computed:

- classical statistics (mean, variance, skewness, kurtosis) of reflectance or radiance values;

- quantile-based descriptors to capture distribution shape and robustness to outliers;
- histogram representations of intensity distributions to identify multimodal behaviour;
- entropy measures to quantify information content and signal variability;

In addition, spectral-specific feature engineering is applied:

- spectral derivatives (first and second order) to capture local spectral variations;
- continuum removal and normalisation to isolate absorption features;
- band ratios and spectral indices (e.g., normalised difference indices) to enhance discriminative properties.

Spatially-aware descriptors are also included:

- local variance and texture measures (e.g., using grey-level co-occurrence matrices, GLCM);
- spatial autocorrelation metrics (e.g., Moran's I);
- statistics over segmented regions or superpixels to capture local consistency.

Cross-domain features combine spatial and spectral information:

- covariance between spectral bands across spatial regions;
- distribution of spectral signatures within clusters or regions;
- variability of spectral features across the image plane;

Outlier detection features are derived by identifying pixels or regions with anomalous spectral behaviour, using statistical thresholds or distance-based measures (e.g., Mahalanobis distance in spectral space).

5.1 HSI algorithmic calibration overview

Reference stage

The reference stage provides an initial validation of the hyperspectral acquisition system by establishing a baseline representation of expected behaviour under nominal conditions. The reference targets, typically white and dark references, are acquired and analysed by computing deterministic quality indicators and low-level statistical descriptors. These measurements are then compared against historical reference distributions to assess consistency and detect potential calibration drift. The upper-left panel of Figure 5.1 presents a selected visualisation of the reference stage, in which the system's baseline behaviour is established using reference measurements. The plots show average spectra and distributions of selected indicators, highlight the variability under nominal conditions, and allow the definition of empirical tolerance ranges. The reference stage supports the definition of tolerance bounds and confidence intervals, which serve as a benchmark for evaluating all subsequent acquisitions. For space constraints, not all visualisations are reported; these generally include distribution plots of key indicators, variability analyses, and comparisons with historical reference data. Variability and confidence measures derived from the newly acquired reference data are used to determine whether the system is correctly calibrated. If the observed measurements fall within predefined statistical bounds, the system is considered ready for acquisition. Otherwise, a calibration issue is detected, and a warning is raised, requiring verification of the acquisition setup or recalibration before proceeding.

Feature space stage

Following successful reference validation, each hyperspectral acquisition is transformed into a structured feature representation through the systematic extraction of the described quality indices and statistical descriptors. The lower-left panel of Figure 5.1 corresponds to the feature space stage. It includes a feature distribution plot, showing the variability of individual descriptors across samples, where differences between acceptable and problematic acquisitions are visible. A correlation matrix, highlighting dependencies among features. Strong correlations (both positive and negative) can be observed, indicating redundancy among some descriptors and the presence of structured relationships in the data. These results confirm that the feature space is high-dimensional but structured, enabling both statistical interpretation and downstream machine learning tasks. These visualisations demonstrate how both high-quality and degraded datasets are embedded within a unified representation, forming the basis for subsequent statistical analysis and machine learning tasks. For space constraints, not all generated visualisations are reported; these generally include (i) distributions of individual spectral and spatial features, (ii) correlation analyses capturing dependencies between variables, and (iii) aggregated representations supporting the interpretation of the feature space structure.

Exploratory stage

The exploratory stage aims to characterise the statistical structure of the feature space and to identify patterns, anomalies, and discriminative behaviours within the dataset. At this stage, both univariate and multivariate analyses are performed to assess feature distributions, detect outliers, and investigate dependencies between variables. The upper-right panel of Figure 5.1 shows selected exploratory plots. Multiple cumulative distribution plots and statistical comparisons (good vs bad) are shown, together with a PCA scatter plot, coloured by “good” and “bad” acquisitions. tSNE projection reveals a good separation between the two classes, with “good” samples forming a compact cluster and “bad” samples dispersed in a different region of the feature space. This indicates that, already with the available data, the extracted features capture meaningful structure and provide discriminative power for quality assessment. The cumulative plots further show systematic differences in feature distributions between classes. These visualisations provide insight into the separability between “good” and “bad” acquisitions and highlight the most informative features contributing to this distinction. This stage plays a critical role in validating the relevance of the extracted features and in guiding the design of the classification models.

Classification stage

In the classification stage, the extracted feature vectors are used to train and evaluate machine learning models for the automatic discrimination between acceptable and problematic hyperspectral acquisitions. The problem is formulated as a supervised classification task, where labelled examples are used to learn decision boundaries in the feature space. The central lower panel of Figure 5.1 represents the classification stage, including examples of misclassified sample profiles, which provide insight into borderline or ambiguous cases; a ROC curve showing classification performance. In this case, the curve indicates strong discriminative performance between good and bad acquisitions (see also Table 5.1). This result demonstrates that the selected features and model are highly effective in separating classes, although the presence of misclassified examples suggests some overlap or noise in specific cases. The availability of labelled datasets allows the formulation of a supervised machine learning problem where XGBoost and CatBoost are employed. In parallel, unsupervised learning techniques are used to explore the data's structure. Clustering approaches such as Gaussian Mixture Models, DBSCAN/HDBSCAN, and spectral clustering are applied to group acquisitions or regions with similar spectral characteristics. Dimensionality reduction techniques such as PCA,

t-SNE, or UMAP are used for visualisation and to support exploratory analysis. Anomaly detection methods, including Isolation Forest and One-Class SVM, are used to identify outliers and rare failure modes.

XAI stage

The XAI stage introduces interpretability into the classification framework by analysing the contribution of individual features to model predictions. While the classification stage provides a decision outcome, this stage aims to explain *why* a given acquisition is classified as "good" or "bad", linking model outputs to physically meaningful indicators. The lower-right panel represents the XAI stage, where model interpretability is addressed. A global feature importance plot ranks the most influential variables driving classification decisions (top plot), while multiple local explanation plots (e.g., feature contribution bars) illustrate how individual features contribute positively or negatively to specific predictions. These visualisations reveal that a subset of features dominates the decision process, providing a direct link between statistical descriptors and physical acquisition factors. This enables diagnostic interpretation of failure modes and supports targeted corrective actions.

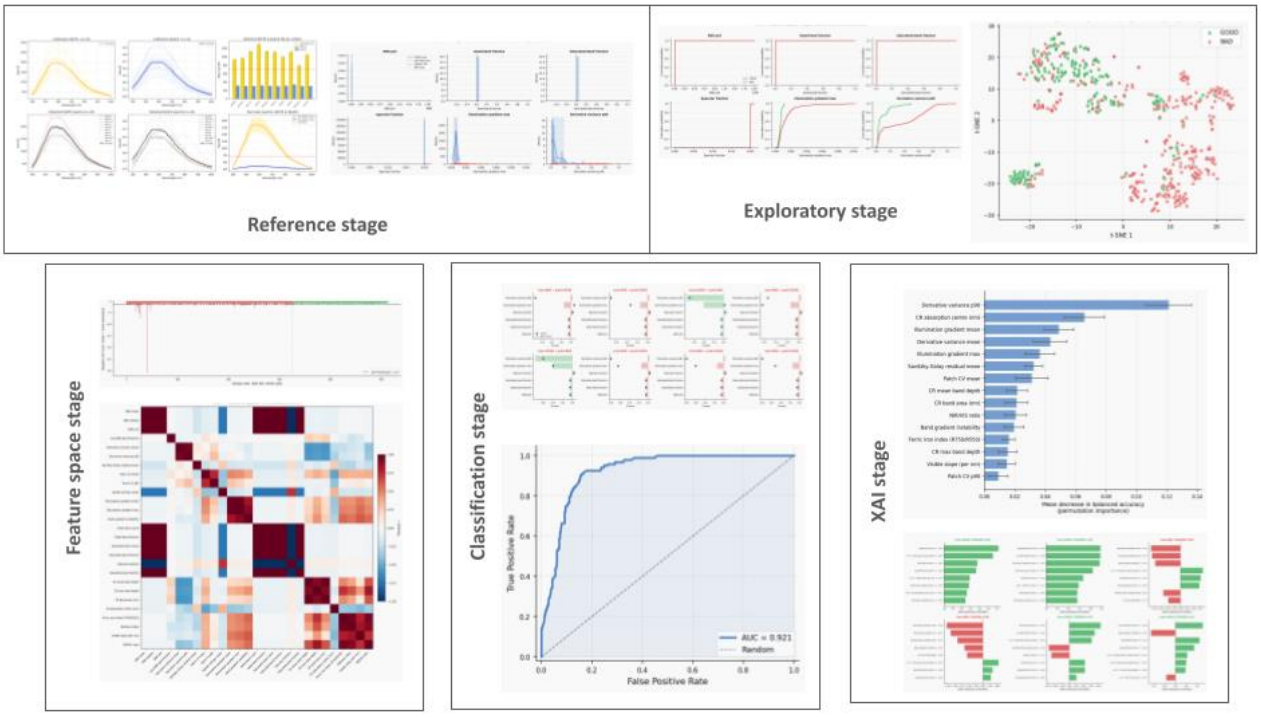

***Figure 5.1**: Overview of the HSI calibration pipeline across its main stages. Representative plots for each stage are shown, and a description is provided in the text. The pipeline constructs a structured and discriminative feature space from hyperspectral data, enabling clear separation between high-quality and degraded acquisitions. The integration of XAI methods further enhances interpretability, allowing the system not only to detect anomalies but also to identify their underlying causes and support targeted corrective actions.*

Metric	Accuracy	Precision	Recall	F1	AUC
Value	0.86	0.77	0.89	0.82	0.92

Table 5.1*: performance evaluation of the HSI classification model. The metrics provided include Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC), collectively showing the model's capability in quality assessment, with currently available data.*

.**Feedback stage**

The feedback stage converts outputs from feature extraction, classification, and XAI analysis into concrete feedback for the hyperspectral acquisition system. The decision process is driven by the modality-specific quality indicators. Based on these indicators, deviations are mapped to corrective actions through rule-based logic or threshold-driven policies. Typical mappings include:

- low SNR → increase exposure time, increase illumination intensity, reduce scanning speed;
- illumination non-uniformity → adjust lighting configuration, apply flat-field correction, reposition light sources;
- spectral noise or discontinuities → trigger recalibration (dark/white reference), apply smoothing or filtering;
- band saturation → reduce sensor gain or exposure time;
- presence of dead or low-variance bands → exclude affected bands from analysis or trigger sensor diagnostics.

Corrective actions are prioritised according to severity. At this stage, the feedback stage operates at a decision-support level, as direct integration with the hyperspectral sensor control and robotic system has not yet been implemented. The output of this stage consists of a structured set of parameter updates and acquisition flags (e.g., accept, adjust, reacquire, reject), which can be used by the operator or external control modules. Figure 5.2 includes the composite score components in the left panel, showing the distributions of the three key components used to compute the global severity score, i.e. D(x), A(x) and S(x). Mahalanobis distance indicates that acceptable acquisitions are concentrated at low values, while reacquire and human-intervention cases extend towards higher distances, reflecting stronger deviation from reference distributions. The anomaly score also shows a clearer separation between classes is visible, with acceptable samples clustered at low anomaly values and problematic acquisitions (especially reacquire and human intervention) shifted towards higher scores. The classifier uncertainty displays broader distributions shifted towards higher values, indicating reduced model confidence. Overall, the combined severity score provides effective separation between decision classes, with a clear progression from acceptable to human-intervention cases. This confirms that combining the three components improves robustness and discriminative power. In the top-right panel, the violin plots represent the distribution of actionability scores across decision classes, highlighting the role of actionability in distinguishing between correctable and non-correctable issues. The bar chart summarises the dominant causes of non-acceptable decisions, providing the basis for mapping feature-space deviations to physical acquisition problems. The bottom panel presents a single-sample decision example, where the individual scores (D(x), A(x), U(x)) are shown as a bar chart, with the anomaly component dominating the severity. The resulting global score leads to a “Reacquire” decision,

flagged as rule-based. The system identifies SNR-related issues as the dominant cause and suggests corrective actions (e.g., adjusting exposure time). An actionability score of ~40% indicates partial feasibility of automatic correction, supporting the decision to reacquire rather than directly correct. The HSI feedback stage supports data reduction and resource-efficient operation by enforcing quality-driven selection at both spectral and spatial levels. Spectral bands identified as unreliable can be dynamically masked and excluded from further processing, while pixels or regions affected by shadows, specular reflections, or illumination, artefacts are filtered out based on statistical and spatial consistency criteria. This aspect will be further refined in D5.3, particularly with respect to integrating validated spectral information into the enriched 3D model. In the AUTOMATA workflow, this selection is particularly relevant because archaeological artefacts often present irregular morphologies, reflective lithic surfaces, ceramic slips, glazes, coatings, and heterogeneous preservation conditions, all of which may locally alter spectral behaviour. For acquisitions that marginally meet quality thresholds, only pre-processed representations of the hyperspectral data are retained instead of the full hyperspectral cube. This approach reduces the dimensionality of the data and limits the propagation of noise into downstream analysis.

Figure 5.2: *Feedback stage of the HSI calibration pipeline. The figure illustrates the computation and interpretation of the composite severity score based on statistical deviation, anomaly detection, and classification uncertainty. The combined score enables separation between decision classes, supporting the assignment of acquisitions to operational categories (acceptable, auto-correctable, reacquire, human intervention). The bottom panel provides a representative example of the decision process, showing how quantitative indicators are translated into interpretable feedback and targeted corrective actions.*

6 XRF

The AUTOMATA system is intended to process large volumes of XRF data from artefacts of variable and often only partially known composition. Therefore, the spectral quality assessment is designed to function with as few assumptions about composition as possible, focusing on signal-level indicators rather than compositional validation. Where reference materials of known composition are available, they can be integrated into the pipeline to support quantitative validation, that is, to verify that reported elemental concentrations are accurate and consistent across sessions. This step, however, is complementary to spectral quality assessment and depends on the availability of appropriate certified standards for the material class under analysis (Frahm, 2024). Its integration into the automated pipeline is therefore foreseen as a subsequent development, once the material-specific requirements of the archaeological collections processed within AUTOMATA have been established. Preliminary analyses are conducted on a dataset of already available XRF acquisitions, including both “good” and “bad” cases. Each acquisition consists of one or more spectra representing photon counts as a function of energy, typically corresponding to elemental emission lines. The acquisitions are assumed to be performed under partially controlled conditions, with variability in acquisition time, detector settings, geometry, and environmental factors. Each XRF dataset is evaluated through a set of deterministic quality indices that quantify spectral, statistical, and signal-related properties of the measurements. The following deterministic quality indices are considered.

- Signal-to-noise ratio (SNR). SNR is automatically estimated from the XRF spectra by analysing the ratio between signal intensity at characteristic peaks and background noise. The signal component is derived from peak intensities, while noise is estimated from local background fluctuations or from spectral regions without expected peaks. SNR can be computed per peak or aggregated across the spectrum.
- Total count rate and minimum count threshold. The total number of X-ray counts recorded across a spectrum is a direct indicator of signal statistical quality. Spectra falling below an empirically defined minimum threshold (e.g., the bottom 5% quantile of the distribution) are considered unreliable, as low counts lead to noisy peaks, high elemental uncertainties, and a large proportion of elements reported below the limit of detection. Low count rates (if higher is expected) can result from poor sample positioning, irregular or porous surfaces, insufficient acquisition time, or obstruction between the detector and the sample.
- Argon peak intensity. When a gap exists between the instrument window and the sample surface, X-rays pass through air and excite argon atoms, producing a characteristic fluorescence peak at approximately 2.96 keV. The intensity of this peak provides a direct indicator of measurement geometry: a prominent argon peak indicates a significant air gap and signal attenuation, while its absence or a minimal presence confirms direct contact between the instrument and the sample. It is important to note that atmospheric pressure can also play a significant role in detecting argon, particularly at different altitudes. It is, therefore, a parameter that must be taken into account and is often measured by the pXRFs themselves. It should be noted that this criterion cannot be applied when the instrument uses a silver (Ag) anode, as the L-lines of Ag overlap with the Ar Kα peak, making the two signals indistinguishable.
- Peak detectability and prominence evaluates the presence and clarity of spectral peaks corresponding to elemental emissions. Peak detection is performed automatically using methods such as local maxima detection, derivative-based approaches, or fitting with parametric models (e.g., Gaussian or Voigt profiles). Peak prominence, width, and height relative to the background are quantified.

- Spectral baseline stability is assessed by analysing the low-frequency component of the spectrum after peak removal or smoothing. Techniques such as polynomial fitting, morphological filtering, and asymmetric least-squares (ALS) baseline correction are used to estimate the background. Statistical descriptors of the baseline (e.g., variance, drift) are then computed.
- Energy calibration consistency is evaluated automatically by analysing the relative positions of detected peaks and their consistency across spectra. Instead of relying on external references, internal consistency is assessed by comparing peak positions across repeated measurements or across the dataset.
- Spectral entropy and information content. Entropy-based measures are computed from the spectral distribution to quantify the signal's information content and variability.

Beyond deterministic metrics, statistical descriptors are computed over the spectral domain to capture the distributional properties of the data. Each XRF acquisition is represented as a high-dimensional feature vector combining peak-based features, statistical descriptors, and distributional characteristics. For each spectrum and for detected peaks, the following are computed:

- classical statistics (mean, variance, skewness, kurtosis) of spectral intensities;
- quantile-based descriptors to capture distribution shape and robustness to outliers;
- histogram-based features of count distributions across energy bins;
- entropy measures to quantify signal complexity.

In addition, XRF-specific feature engineering is applied:

- peak fitting parameters (e.g., peak height, area, FWHM, residual error);
- peak ratios between characteristic emission lines (useful for relative composition consistency);
- background-subtracted spectra and baseline-corrected features.

Outlier-related features are derived by identifying spectra with anomalous peak structures or statistical properties, using distance-based measures (e.g., Mahalanobis distance in feature space) or density-based approaches. The resulting feature vector captures both global spectral characteristics and localised anomalies, enabling robust discrimination between high-quality and degraded XRF measurements.

6.1 XRF algorithmic calibration overview

Reference stage

Before statistical quality assessment of the acquisitions, instrument-level calibration must be ensured. For the Olympus Vanta pXRF, energy calibration is largely automated: the instrument continuously performs internal checks using a simulated X-ray photon to verify the energy scale before each measurement. Periodic external verification is carried out through the built-in CalCheck routine, which uses a stainless steel 316 reference to confirm the positions of the Fe Kα (6.4 keV) and Mo Kα (17.5 keV) peaks and adjust detector gain and energy offset accordingly (Frahm, 2024). This two-point calibration ensures that X-ray peaks are correctly positioned along the energy axis. CalCheck also evaluates tube output, count rates, and spectral resolution, providing a set of performance indicators that can be logged and monitored over time. At the beginning of each acquisition session, spectral quality indicators are computed on a set of initial acquisitions to establish a session-level statistical baseline of expected instrument behaviour. The resulting descriptors are compared with those obtained in previous sessions. If the measured spectra are consistent with expected statistical ranges, the instrument is considered to be operating correctly; otherwise, deviations are flagged

and require review before proceeding. The top-left panel of Figure 6.1 shows selected plots of the Reference stage, displaying representative XRF spectra and associated statistical distributions obtained from baseline acquisitions. In this case, the spectra show well-defined peaks corresponding to elemental emission lines, with low noise and stable baselines. Accompanying histograms and distributions of selected indicators demonstrate tight clustering, indicating consistent instrument behaviour. These results define the reference distributions and tolerance bounds, against which subsequent acquisitions are evaluated for calibration consistency.

Feature space stage

Each spectrum is transformed into a structured feature vector combining the spectral and statistical indicators described above. This representation encodes both global spectral behaviour and localised irregularities, enabling a comprehensive description of acquisition quality. The selected plots for the feature space stage are shown in the top-centre pane of Figure 6.2, including distribution plots and cumulative curves that highlight the variability of extracted features across the dataset; differences between classes (good vs bad spectra) begin to emerge. The feature space captures both global spectral properties and local irregularities, forming a structured representation suitable for downstream analysis. As the dataset evolves, the statistical characterisation of spectral features is progressively updated, allowing the system to adapt to long-term variations in acquisition conditions and to refine its ability to distinguish between acceptable and problematic measurements.

Exploratory stage

The exploratory stage focuses on the statistical analysis of the constructed feature space in order to identify patterns, anomalies, and discriminative structures within the dataset. At this stage, both univariate and multivariate analyses are performed to investigate the distribution of spectral features, detect outliers, and assess relationships between variables. Feature correlation and variability are shown in the bottom-left panel of Figure 6.1. This panel presents: a feature distribution plot showing variability across samples; a correlation matrix, revealing strong dependencies among groups of features. The presence of correlated blocks indicates redundancy and structured relationships within the feature space, which can inform feature selection and model design. This stage supports the validation of the feature engineering process and provides guidance for the subsequent design and refinement of classification models.

Classification stage

In the classification stage, the extracted feature vectors are used to train and evaluate machine learning models for the automatic discrimination between reliable and problematic XRF spectra. The problem is formulated as a supervised classification task, using XGB and Catboost algorithms. Clustering approaches such as Gaussian Mixture Models, DBSCAN/HDBSCAN, and spectral clustering are applied to group spectra with similar characteristics. Anomaly detection methods, including Isolation Forest and One-Class SVM, are used to identify outliers and rare failure modes. In Figure 6.1, the classification stage is shown in the bottom-centre panel, summarised through: misclassification analysis plots highlighting borderline cases; a ROC curve, showing strong discriminative performance, although Precision, Recall and F1-score remain moderate (see Table 6.1). This indicates that the model achieves high discrimination, reflecting effective feature representation. At the current stage, XRF data have been acquired under heterogeneous conditions, with variability in acquisition parameters and setup. This variability is beneficial for constructing a representative dataset but may not reflect future acquisition scenarios, particularly with the integration of automated or robotic systems. As a result, the statistical properties of the data may evolve over time. The current models,

therefore, represent an initial baseline. The pipeline is designed to be adaptive, allowing continuous integration of new data, updating of statistical descriptors, and retraining of machine learning models to maintain robustness and reliability.

XAI stage

The XAI stage introduces interpretability into the classification framework by analysing the contribution of individual spectral and spatial features to the model predictions, with the objective of linking data-driven decisions to physically meaningful properties of the hyperspectral data. The XAI stage (bottom-right panel) is shown in Figure 6.1, providing interpretability through feature importance rankings that identify the most influential variables driving classification decisions, and local explanation plots (e.g., SHAP values) that show how individual features contribute positively or negatively to predictions. The results highlight that a subset of features dominates the decision process, and that both spectral quality indicators and statistical descriptors contribute to classification, reflecting the multi-faceted nature of XRF data quality. For space constraints, not all generated visualisations are reported; these generally comprise (i) global importance analyses, (ii) local explanations for representative samples, and (iii) feature interaction analyses capturing combined effects between variables.

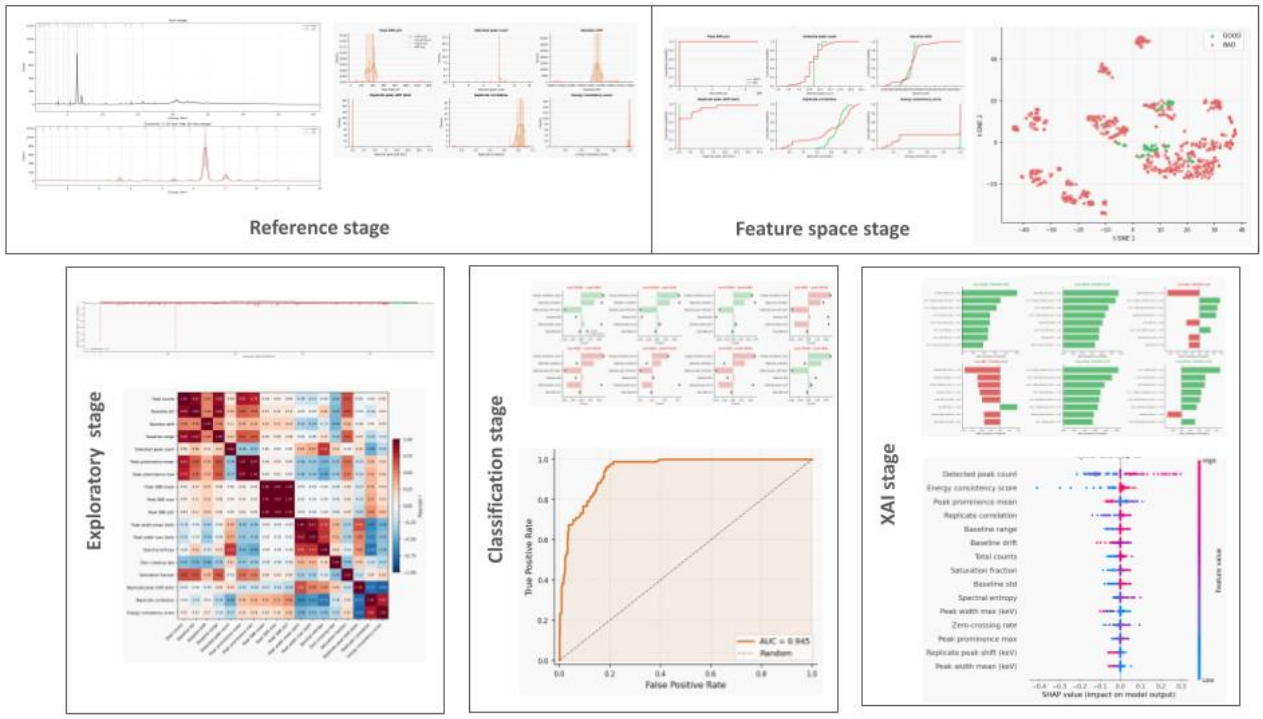

***Figure 6.1**: Overview of the XRF calibration pipeline across its main stages. Representative plots illustrate the progression from reference validation to feature space construction, exploratory analysis, classification, and explainability. The pipeline captures the statistical and spectral structure of XRF data, enabling meaningful separation between acceptable and degraded spectra, as evidenced by classification performance. The integration of XAI methods highlights the most influential features and provides insight into the underlying causes of quality degradation, supporting diagnostic interpretation and calibration refinement.*

Metric	Accuracy	Precision	Recall	F1	AUC
Value	0.94	0.63	0.65	0.64	0.93

***Table 6.1**: Performance evaluation of the XRF classification model (CatBoost). While the model demonstrates strong discriminative ability (AUC) and high accuracy, the moderate Precision, Recall, and F1-score suggest room for improvement in handling class imbalance and prediction consistency.*

Feedback stage

The feedback stage translates the outputs of feature extraction, classification, and XAI analysis into concrete parameter adjustments for the XRF acquisition system. The decision process is driven by the modality-specific quality indicators described. Typical mappings include:

- low signal-to-noise ratio → increase acquisition time, improve excitation conditions, ensure stable contact between sensor and sample;
- weak or poorly defined peaks → increase acquisition time, optimise measurement geometry, verify detector settings;
- spectral baseline instability → adjust acquisition environment, reduce external interference, apply baseline correction procedures;
- peak shifts or energy misalignment → trigger recalibration (e.g., reference measurement or internal calibration routine);
- spectral noise or irregular fluctuations → apply smoothing or filtering, verify detector stability;
- saturation or excessive count rates → reduce acquisition time or detector exposure;
- absence of expected peaks or anomalous spectral structure → verify measurement positioning, repeat acquisition under controlled conditions.

Moderate deviations can be addressed by parameter tuning within predefined operational bounds, whereas severe deviations trigger acquisition rejection and require human intervention. In addition, the XRF feedback stage supports data efficiency and resource-aware operations: spectra identified as unreliable or severely degraded can be excluded from further analysis, while marginal cases may be retained only in processed form (e.g., baseline-corrected or denoised spectra). Figure 6.2 illustrates the feedback stage of the XRF calibration pipeline. The upper-left panel presents the distributions of the three components contributing to the global severity score, i.e. D(x), A(x) and S(x). The combined severity score provides a clear separation between decision classes, showing a consistent progression from acceptable to human-intervention cases, confirming the effectiveness of combining statistical deviation, anomaly detection, and uncertainty into a single decision metric. The violin plots on the top-right show the distribution of actionability scores across decision classes, highlighting that actionability is not strictly monotonic with severity, but depends on the nature of the underlying issue. The bar chart identifies the most frequent causes of non-acceptable decisions, providing a direct link between feature-space deviations and physical or instrumental causes. The bottom panel presents a single-sample decision example, highlighting the contributions of D(x), A(x), and U(x), with the anomaly score dominating the overall severity. The actionability score is relatively high, suggesting that

the issue is technically correctable but requiring controlled intervention rather than automatic adjustment.

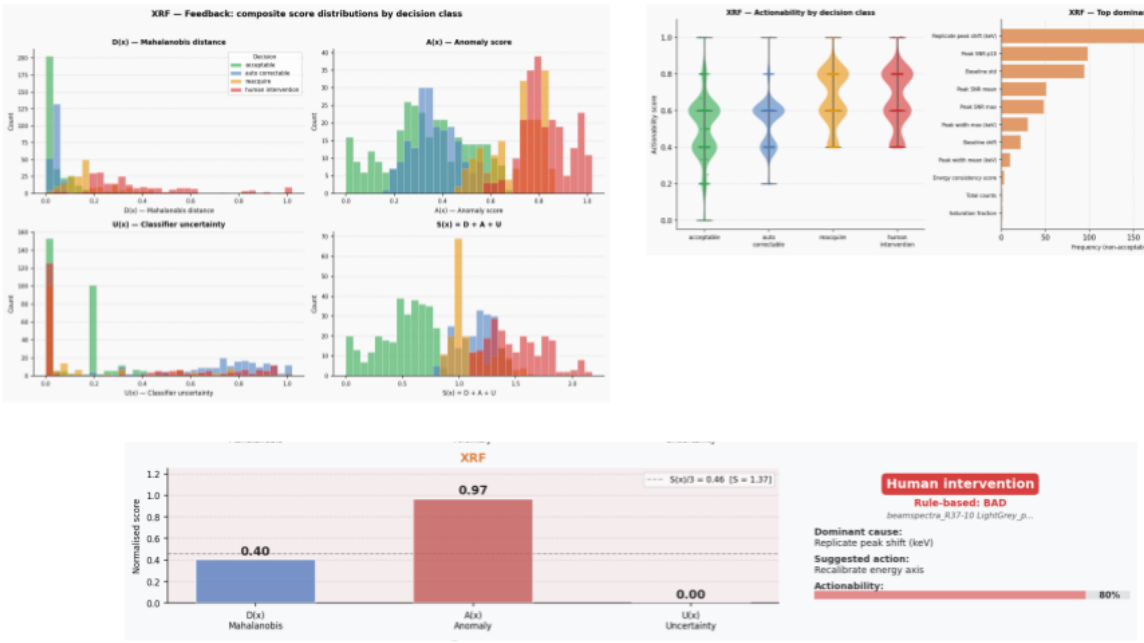

Figure 6.2 *: Feedback stage of the XRF calibration pipeline. The figure illustrates the computation of the composite severity score, enabling the separation between operational decision classes. Actionability analysis highlights the extent to which deviations can be corrected automatically, while feature attribution identifies the dominant causes of spectral degradation. The bottom panel presents a representative decision case, demonstrating how the system can translate quantitative indicators into interpretable diagnostics and targeted corrective actions.*

7 Raman spectroscopy

The Raman spectroscopy component of the pipeline focuses on the quantitative assessment of vibrational spectral data quality by combining physically grounded indicators, statistical descriptors, and machine-learning-based classification. The objective is to characterise the variability in Raman spectra acquired from archaeological artefacts, identify acquisition and calibration issues, and enable automated discrimination between reliable and problematic measurements. Each acquisition consists of one or more spectra representing intensity as a function of Raman shift (wavenumber), typically reflecting molecular vibrational modes of the analysed materials. The acquisitions are assumed to be performed under partially controlled conditions, with variability in laser power, integration time, focus, fluorescence background, and environmental conditions. The following deterministic quality indices are considered.

- Signal-to-noise ratio is automatically estimated from Raman spectra by comparing the intensity of characteristic Raman peaks to the local noise level. The signal component is derived from peak intensities or integrated peak areas, while noise is estimated from high-frequency components, baseline residuals after smoothing, or spectral regions without significant peaks.
- Fluorescence background level and stability. Raman spectra often exhibit a fluorescence background that can obscure Raman peaks. This component is automatically estimated using baseline fitting techniques such as polynomial fitting, asymmetric least squares (ALS), or morphological filtering. The magnitude and variability of the background are quantified.
- Spectral smoothness and peak sharpness is evaluated using derivative-based measures and local fitting residuals, while peak sharpness is quantified through peak width (e.g., full width at half maximum, FWHM) and curvature around peak centres.
- Peak detectability and prominence. Peak detection is performed automatically using local maxima detection, derivative-based methods, or model fitting. Metrics such as peak height, prominence, and signal-to-background ratio are extracted.
- Baseline stability and residual noise. After baseline removal, residual noise is analysed to assess signal quality. Statistical descriptors of the residual signal (e.g., variance, entropy) are computed.
- Spectral alignment is evaluated automatically by analysing the consistency of peak positions across spectra within the dataset. Cross-correlation techniques or peak-matching algorithms are used to detect shifts.
- Saturation and dynamic range utilisation. This metric evaluates whether the detector operates within its optimal dynamic range. Saturation is detected by identifying intensity values near the maximum measurable level, while underexposure is detected through low signal levels.

Beyond deterministic metrics, statistical descriptors are computed over the spectral domain to capture the distributional and structural properties of the data. Each Raman acquisition is represented as a high-dimensional feature vector combining spectral statistics, peak-based features, and distributional descriptors. For each spectrum, the following are computed:

- classical statistics (mean, variance, skewness, kurtosis) of intensity values;
- quantile-based descriptors to capture distribution shape and robustness to outliers;
- histogram-based representations of intensity distributions;
- entropy measures to quantify signal complexity and variability;

In addition, Raman-specific feature engineering is applied:

- first- and second-order spectral derivatives to enhance peak detection;

- baseline-corrected and normalised spectra to isolate Raman features;
- peak fitting parameters (e.g., peak height, area, FWHM, symmetry, fitting residuals);
- peak ratios and relative intensities to capture compositional information.

Outlier detection features are derived by identifying spectra with anomalous peak structures, abnormal baseline behaviour, or irregular statistical properties, using distance-based or density-based methods. The resulting feature vector captures both global spectral behaviour and localised irregularities, enabling robust classification of acquisition quality.

7.1 Raman spectroscopy algorithmic calibration overview

Reference stage

At the beginning of each acquisition session, a set of reference materials with well-characterised Raman signatures is measured in order to establish a baseline response of the system. These include standards with known peak positions and intensities, enabling the assessment of spectral alignment and instrument performance. The acquired spectra are used to evaluate the quality indicators described in the previous section. The resulting statistical descriptors are compared with those obtained in previous sessions, providing a reference model of expected system behaviour under correct calibration conditions. Deviations observed in the reference measurements are analysed in terms of variability and confidence indicators. If the measured spectra fall within the expected statistical ranges, the system is considered correctly calibrated; otherwise, calibration issues are identified and require intervention before proceeding. The reference stage selected plots are on the top-left panel of Figure 7.1, showing the distributions of key Raman quality indicators derived from calibration measurements. Histograms and boxplots display relatively compact distributions, indicating stable acquisition conditions. These results establish the expected behaviour of the system under correct calibration.

Feature space stage

Each spectrum is transformed into a structured feature vector combining the deterministic spectral indicators and statistical descriptors described above. The representation captures both global spectral behaviour and localised irregularities, enabling a comprehensive characterisation of acquisition quality. Feature space stage selected plots are shown in the top-right panel of Figure 7.1. Cumulative distribution plots show differences between good and bad spectra, particularly in features related to noise, baseline behaviour, and peak characteristics. The tSNE projection reveals a clustering of good samples and a more dispersed distribution of bad samples. This indicates that the extracted features effectively capture spectral degradation patterns, although some overlap suggests non-linear or complex failure modes. As the dataset evolves, the statistical characterisation of spectral features is progressively updated, allowing the system to adapt to long-term variations in acquisition conditions and to refine its ability to distinguish between acceptable and problematic measurements.

Exploratory stage

The exploratory stage is dedicated to statistical investigation of the feature space to identify patterns, anomalies, and discriminative structures within the dataset. Both univariate and multivariate analyses are conducted to assess feature distributions, detect outliers, and analyse dependencies between variables. Exploratory stage charts are on the bottom-left panel of Figure 7.1. The exploratory stage includes a correlation matrix, highlighting relationships among features and showing strong correlations among groups

of variables, particularly those related to fluorescence background, noise and residual variability, peak sharpness and spectral smoothness. This confirms that Raman data quality is influenced by interdependent spectral factors, rather than isolated metrics. This stage supports the validation of the feature engineering process and provides guidance for the development and refinement of classification models.

Classification stage

In the classification stage, the extracted feature vectors are used to train and evaluate machine learning models for the automatic discrimination between reliable and problematic Raman spectra. The problem is formulated as a supervised or semi-supervised classification task, depending on the availability of labelled data, with models learning decision boundaries based on spectral quality indicators and statistical descriptors. The availability of labelled datasets allows the formulation of a supervised machine learning problem. Tree-based models such as Random Forest, XGBoost, and CatBoost are employed for the classification problem. Unsupervised learning techniques are used to explore the structure of the data. Clustering approaches such as Gaussian Mixture Models, DBSCAN/HDBSCAN, and spectral clustering are applied to group spectra with similar characteristics. Dimensionality reduction techniques such as PCA, t-SNE, or UMAP support visualisation and exploratory analysis. Anomaly detection methods, including Isolation Forest and One-Class SVM, are used to identify outliers and rare failure modes. At the current stage, Raman data have been acquired under heterogeneous conditions, with variability in acquisition setup and environmental factors. This variability is beneficial for constructing a representative dataset but may not reflect future automated acquisition scenarios. As with other modalities, future integration with robotic systems is expected to produce more consistent acquisition conditions, potentially altering the statistical properties of the data. Therefore, the current statistical models and classifiers are considered as an initial baseline. The pipeline is designed to be adaptive, allowing continuous integration of new data, updating of statistical descriptors, and retraining of machine learning models. Figure 7.1 shows the classification stage (bottom-centre panel) of Raman analyses. A ROC curve with AUC ≈ 0.83 demonstrates discriminative performance between acceptable and problematic spectra, with moderate values for Precision, Recall and F1-score (see Table 7.1). These results reflect the inherent complexity of Raman data, where factors such as fluorescence and noise introduce variability that is more difficult to model than in other modalities.

XAI stage

The XAI stage introduces interpretability into the classification framework by analysing the contribution of individual spectral features to model predictions. While the classification stage provides an automatic decision, this stage aims to explain the underlying factors driving the classification outcome. XAI stage (bottom-right panel of Figure 7.1) provides interpretability through feature importance rankings, identifying the most influential variables, and local explanation plots, showing feature contributions for individual predictions. The most relevant features, at the current stage, include: low-signal fraction and noise-related indicators; peak curvature and sharpness metrics; fluorescence-related features. These results demonstrate that classification decisions are driven by a combination of signal quality, background effects, and spectral structure, enabling diagnostic interpretation of degradation causes. From a calibration perspective, these insights enable the identification of the root causes of spectral degradation and support the definition of targeted corrective actions, by linking model outputs to physically interpretable spectral properties.

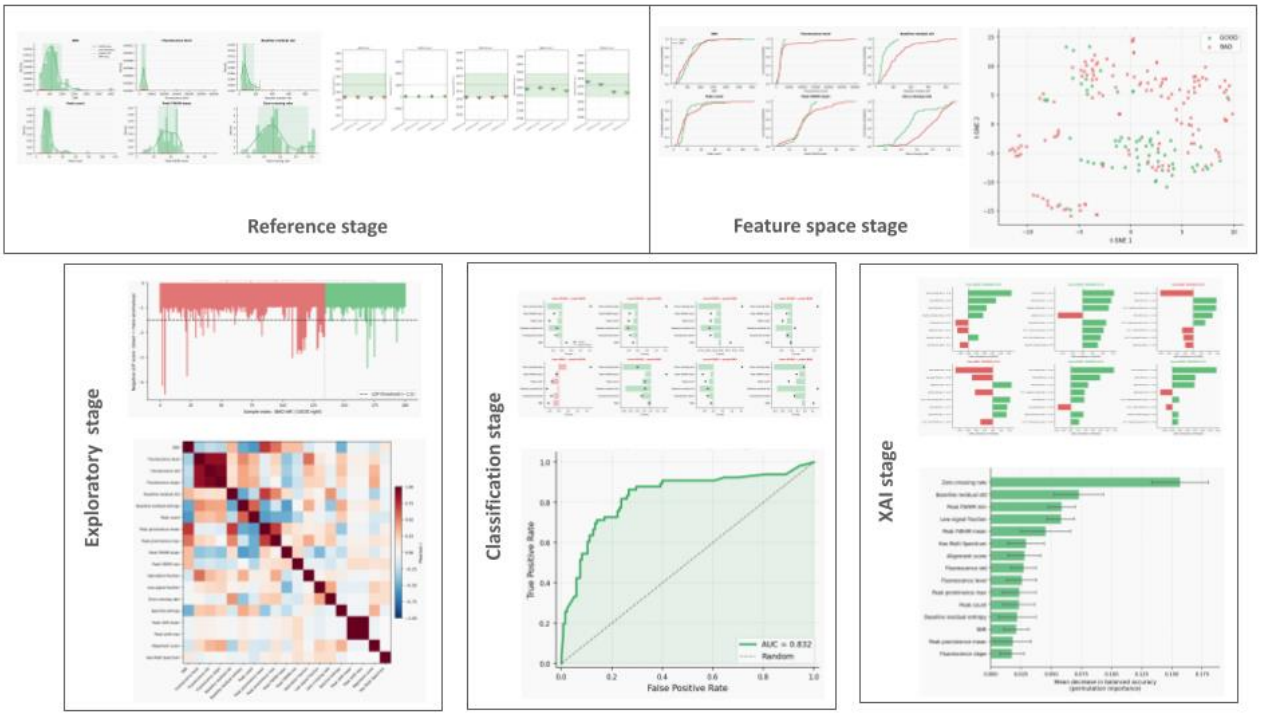

***Figure 7.1**: Overview of the Raman calibration pipeline across its main stages. Representative plots illustrate the progression from reference validation to feature space construction, exploratory analysis, classification, and explainability. The pipeline captures the statistical and spectral complexity of Raman data, enabling separation between acceptable and degraded spectra, as evidenced by classification performance (AUC ≈ 0.83). The integration of XAI methods identifies the most influential spectral features, linking classification outcomes to physically meaningful causes.*

Metric	Accuracy	Precision	Recall	F1	AUC
Value	0.8	0.73	0.61	0.66	0.83

***Table 7.1**: Performance evaluation of the Raman classification model (XGBoost). The results show solid discriminative ability (AUC = 0.83) and satisfactory accuracy, while the moderate Precision, lower Recall, and F1-score suggest limitations in capturing all positive instances and potential class imbalance effects in currently available data.*

Feedback stage

The feedback stage translates the outputs of feature extraction, classification, and XAI analysis into concrete parameter adjustments for the Raman spectroscopy acquisition system. The decision process is driven by modality-specific spectral quality indicators. Based on these indicators, deviations are mapped to corrective actions, following rule-based logic or threshold-driven policies. Typical mappings include:

- low signal-to-noise ratio → increase integration time, increase laser power (within safe limits), optimise focusing conditions;
- high fluorescence background → adjust acquisition parameters (e.g., reduce laser power, modify integration time), apply baseline correction, or consider alternative acquisition settings;
- peak broadening or loss of sharpness → improve focus, stabilise acquisition conditions, reduce thermal effects by adjusting laser power;
- spectral noise or irregular fluctuations → apply smoothing or filtering, verify detector stability and environmental conditions;
- peak shifts or spectral misalignment → trigger recalibration using reference materials, verify instrument stability;
- saturation or limited dynamic range utilisation → reduce laser power or integration time to avoid detector saturation;
- absence or weak presence of expected peaks → verify measurement location, adjust focus and acquisition parameters, or repeat acquisition.

Corrective actions are prioritised according to severity. Moderate deviations are addressed through parameter tuning within predefined operational bounds, while severe deviations trigger acquisition rejection. For Raman spectroscopy, this distinction is critical because not all degraded spectra are equally actionable: low signal or moderate baseline issues may be corrected through parameter adjustment or reacquisition, whereas severe fluorescence, missing diagnostic peaks, or unstable peak positions usually require expert review. The decision process also relies on the XAI stage, identifying the main features contributing to deviations, which are mapped to their underlying causes and associated corrective actions. Automatic correction is therefore limited to deviations whose likely cause is identifiable and linked to a controllable acquisition parameter. The feedback stage is designed to operate iteratively: over successive iterations, the system progressively stabilises measurement conditions and improves the reliability of the acquired spectra. In addition, the Raman feedback stage supports data efficiency and resource-aware operation by enforcing quality-driven data selection and reducing the propagation of noise and artefacts into downstream analysis, limiting unnecessary data storage and processing. Figure 7.2 illustrates selected plots of the feedback stage of the Raman calibration pipeline, showing how statistical indicators, anomaly detection, and model uncertainty are integrated to support decision-making and diagnostic interpretation. The upper-left section presents the distributions of the three components (D(x), A(x) and U(x)) contributing to the global severity score. The aggregated score shows progressive separation across decision classes, though with some overlap, reflecting the more challenging nature of Raman data. The violin plots show the distribution of actionability scores across decision classes. The bar chart identifies the most frequent causes of non-acceptable decisions, linking feature-space deviations to physically meaningful spectral artefacts. The lower panel presents a representative feedback output: the bar chart shows the individual contributions of D(x), A(x), and U(x), with relatively high contributions from anomaly and uncertainty; the resulting decision is “Reacquire”, flagged as rule-based; the dominant cause is identified as baseline residual standard deviation, indicating issues with background correction. The actionability score (~60%) indicates moderate feasibility of correction, supporting the decision to reacquire rather than accept or escalate to human intervention.

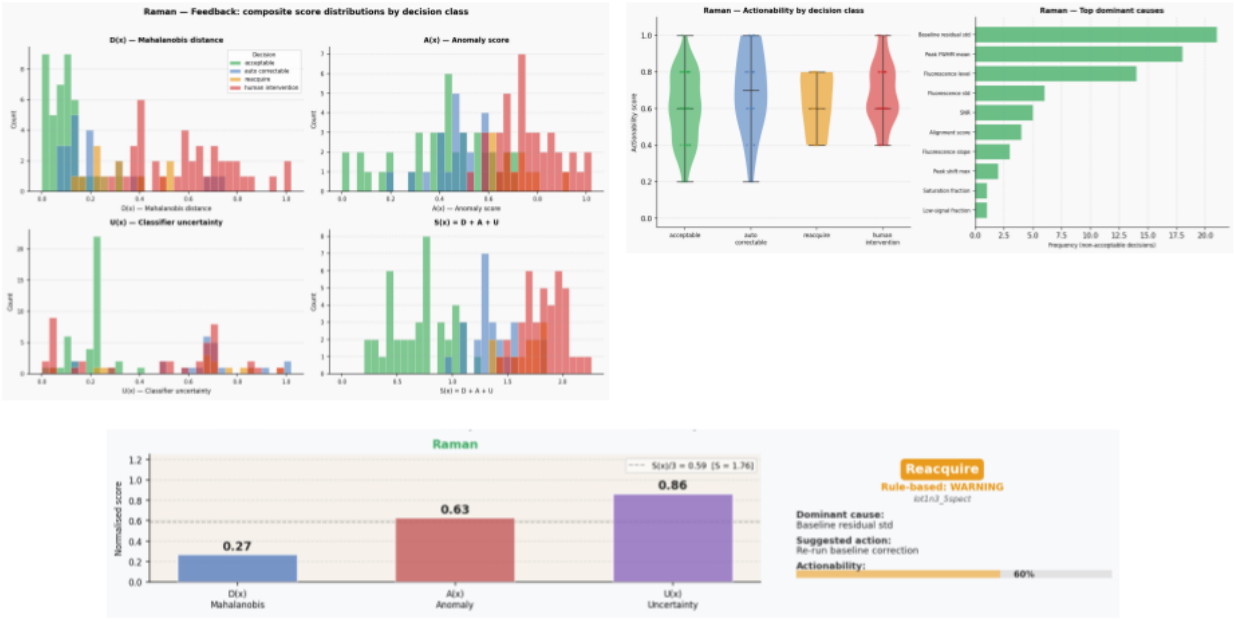

***Figure 7.2**: Feedback stage of the Raman calibration pipeline. The figure illustrates the computation of the composite severity score from statistical deviation, anomaly detection, and classification uncertainty, enabling separation between operational decision classes despite the higher variability of Raman data. The bottom panel presents a representative decision case, demonstrating how quantitative indicators can be translated into interpretable feedback and targeted corrective actions within the calibration loop.*

8 Automatic calibration loop and final comments

The calibration pipeline described in this deliverable is designed to operate as an iterative, closed-loop system in which acquisition, analysis, and control are continuously integrated. Rather than representing a linear sequence of independent processing steps, the pipeline can be understood as a dynamic cycle in which each acquisition is evaluated, interpreted, and used to inform subsequent system behaviour.

At the beginning of each acquisition session, a predefined set of reference objects is scanned under nominal operating conditions. These reference acquisitions provide a baseline representation of the expected system response. From these data and those collected in previous sessions, quality indicators are computed and aggregated into empirical distributions, which define the statistical envelope of correct system behaviour. In practice, this can be visualised as a set of reference curves or distributions against which all subsequent measurements are compared. Variability and confidence intervals derived from these distributions act as acceptance regions: measurements falling within these regions are considered consistent with a properly calibrated system, while deviations indicate potential calibration drift or acquisition instability. Each new measurement enters the pipeline and undergoes feature extraction, producing a structured representation of its statistical and physical properties. These features are then projected against the reference distributions and processed through the classification models. Deviations from expected behaviour are detected through a combination of statistical thresholding and machine learning-based classification. These deviations may manifest as shifts in feature distributions, dispersion outside confidence bounds, or anomalous clustering behaviour.

A central component of the loop is the mapping from detected deviations to corrective actions (see figure 8.1). This mapping can be conceptualised as a transformation from feature space to control space: specific regions or patterns in the feature space correspond to adjustments of acquisition parameters. Typical control variables include illumination intensity, exposure time, sensor gain, acquisition geometry, and measurement duration. Following the identification of corrective actions, the system proceeds to re-acquisition, if needed. The updated parameters are applied, and a new measurement is generated under modified conditions. This new acquisition is reintroduced into the pipeline and evaluated using the same statistical and machine learning framework. If detected deviations exceed acceptable tolerance bounds, the system calls for human intervention.

An important aspect of this loop is its adaptability over time. As new data are acquired across different sessions, objects, and environmental conditions, the statistical descriptors, reference distributions, and classification models can be progressively updated. This allows the system to refine its understanding of normal and abnormal behaviour, improve the accuracy of its decisions, and adapt to changes in instrumentation or acquisition context. In this sense, the calibration loop is not static but evolves with the data, increasing its robustness and generalisation capability.

Future validation will focus on three complementary aspects. First, the datasets will be expanded through repeated acquisitions under controlled robotic conditions. Second, the current binary labels will be refined into a more diagnostic classification of failure modes, enabling the system to distinguish between correctable acquisition problems, non-correctable analytical limitations, and cases requiring human review. Third, the diagnostic outputs of the calibration pipeline will be connected to the robotic and sensor-control layer, so that decisions such as adjust, reacquire, reject, or continue with a post-processing flag can be translated into operational actions. In parallel, the representation of validated outputs and their integration into enriched 3D models will be further developed in connection with D5.3. These steps will transform the present decision-support framework into an increasingly automated calibration loop for enriched digitisation.

Finally, the proposed approach supports efficient and sustainable operation by focusing computational and storage resources on informative data. Measurements identified as non-informative or severely degraded can be discarded, while marginal cases can be retained in processed form (e.g., denoised or reduced representations). This selective retention strategy reduces unnecessary data handling, limits computational overhead, and contributes to lowering the energy consumption associated with large-scale data processing, without compromising the quality and reliability of the calibration process.

***Figure 8.1**: Local feature attribution analysis using local XAI methods, illustrating the contribution of individual features to classification outcomes, for one ceramic and one lithic artefact. The plots highlight how different features positively or negatively influence the decision process, revealing the underlying causes of quality degradation and supporting the interpretability of the calibration pipeline.*

Bibliography

Al-Tameemi, A. A., Li, F., Xiao, Z., and Raza, G. 2026. Advancements in machine learning, deep learning, and data fusion techniques for XRF spectrometry in heavy metal detection: a critical review. *Journal of Analytical Atomic Spectrometry* https://doi.org/10.1039/D5JA00458F

Andric, V., Kvascev, G., Cvetanovic, M., Stojanovic, S., Bacanin, N., and Gajic-Kvascev, M. 2024. Deep learning assisted XRF spectra classification. *Scientific Reports* 14: 3666. https://doi.org/10.1038/s41598-024-53988-z

Bryndza, S., Marmol, U., and Borowiec, N. 2024. 3D Modelling of the Svetovid Statue from Zbrucz Based on Integrated Photogrammetric and Laser Data. *International Journal of Conservation Science*, *15*(4): 1713–1730. https://doi.org/10.36868/ijcs.2024.04.09

Charton, J., Baek, S., and Kim, Y. 2021. Mesh repairing using topology graphs. *Journal of Computational Design and Engineering* 8(1): 251-267. https://doi.org/10.1093/jcde/qwaa076

Costopoulos, S. And Papaioannou, G. 2026. The impact of normal mapping on the appearance of geometrically simplified archaeological 3D models. *Peer Community Journal*, 6: e17. https://doi.org/10.24072/pcjournal.683

Daneshmand, M., Helmi, A., Avots, E., Noroozi, F., Alisinanoglu, F., Arslan, H.S., Gorbova, J., Haamer, R.E., Ozcinar, C., and Anbarjafari, G. 2018. 3D Scanning: A Comprehensive Survey. ArXiv preprint: https://doi.org/10.48550/arXiv.1801.08863

di Filippo, A., Antinozzi, S., Cappetti, N., and Villecco, F. 2024. Methodologies for assessing the quality of 3D models obtained using close-range photogrammetry. *International Journal on Interactive Design and Manufacturing* 18:5917–5924. https://doi.org/10.1007/s12008-023-01428-z

Frahm, E. (2024). Protocols, pitfalls, and publishing for pXRF analyses: From “know how” to “best practices”. Journal of Archaeological Science: Reports, 60, 104831.

Grahn, H., & Geladi, P. 2007. Techniques and applications of hyperspectral image analysis. John Wiley & Sons.

Hernández-Muñoz, Ó. 2023. Analysis of Digitized 3D Models Published by Archaeological Museums. *Heritage*, *6*(5): 3885-3902. https://doi.org/10.3390/heritage6050206

Hubbard, M., Waugh, D., & Ortiz, J. 2004. Provenance determination of chert by VIS/NIR diffuse reflectance spectrometry. Journal of Earth Sciences Sigma Gamma Epsilon, 78(3), 119–129.

Itoh, N., and Hanari, N. 2021. Development of a Polystyrene Reference Material for Raman Spectrometer (NMIJ RM 8158-a). *Analytical Sciences* 37: 1533–1539. https://doi.org/10.2116/analsci.21P054

Lellinger, D., Thomson, J., Coca-Lopez, N., Ntziouni, A., Nikoloudakis, N., Fernández-Álvarez, M., Jeliazkova, N., Bañares, M.A., Portela, R., and Lozano Diz, E. 2025. Interlaboratory study to minimize wavelength calibration uncertainty due to peak fitting of reference material spectra in Raman spectroscopy. *Applied Spectroscopy* 79(12): 1669–1679. https://doi.org/10.1177/00037028251330654

Linderholm, J., Geladi, P., Gorretta, N., Bendoula, R., & Gobrecht, A. 2019. Near infrared and hyperspectral studies of archaeological stratigraphy and statistical considerations. Geoarchaeology, 34(3), 311–321.

Lou, L., Li, W., Gan, W., Yu, Y., Wang, T., Wang, X., and Zhan, Z. 2025. On-the-fly Feedback SfM: Online Explore-and-Exploit UAV Photogrammetry with Incremental Mesh Quality-Aware Indicator and Predictive Path Planning. ArXiv preprint: https://doi.org/10.48550/arXiv.2512.02375

Madden, O., Chan, D.M.W., Dundon, M., and France, C.A.M. 2018. Quantifying collagen quality in archaeological bone: Improving data accuracy with benchtop and handheld Raman spectrometers. *Journal of Archaeological Science: Reports* 18: 596–605. https://doi.org/10.1016/j.jasrep.2017.11.034

Malik, U.S., Tissen, L.N.M., Vermeeren, A.P.O.S. 2021. 3D Reproductions of Cultural Heritage Artefacts: Evaluation of significance and experience. *Studies in Digital Heritage*, 5(1): 1-29. https://doi.org/10.14434/sdh.v5i1.32323

Menna, F., Nocerino, E., Remondino, F., Dellepiane, M., Callieri, M., Scopigno, R. 2016. 3D Digitization of a Heritage Masterpiece - A Critical Analysis on Quality Assessment. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B5: 675–683, https://doi.org/10.5194/isprs-archives-XLI-B5-675-2016.

Montusiewicz, J., Milosz, M., Sarnowski, W., and Kayumov, R. 2026. Quality Assessment of Digital 3D Models of Museum Artefacts from the Mobile LiDAR iPhone and Structured Light Scanners. *Applied Sciences*, *16*(4): 2100. https://doi.org/10.3390/app16042100

Polo, M-.E., Felicísimo, A.M., and Durán-Domínguez, G. 2022. Accurate 3D models in both geometry and texture: An archaeological application. *Digital Applications in Archaeology and Cultural Heritage* 27: e00248. https://doi.org/10.1016/j.daach.2022.e00248.

Sciuto, C., Cantini, F., Chapoulie, R., Cou, C., De la Codre, H., Gattiglia, G., Granier, X., Mounier, A., Palleschi, V., Sorrentino, G., & Raneri, S. 2022. What Lies Beyond Sight? Applications of Ultraportable Hyperspectral Imaging (VIS-NIR) for Archaeological Fieldwork. Journal of Field Archaeology, 1–14. https://doi.org/10.1080/00934690.2022.2135066

Sfikas, K., Perakis, P., and Theoharis, T. 2022. FoR^{2}M: Recognition and Repair of Foldings in Mesh Surfaces. Application to 3D Object Degradation. ArXiv preprint: https://doi.org/10.48550/arXiv.2206.09699

Sorgente, T., Biasotti, S., Manzini, G., and Spagnuolo, M. 2023. A Survey of Indicators for Mesh Quality Assessment. *Computer Graphics forum* 42(2): 461-483. https://doi.org/10.1111/cgf.14779

Yan, C. 2025. A review on spectral data preprocessing techniques for machine learning and quantitative analysis. *iScience* 28: 112759. https://doi.org/10.1016/j.isci.2025.112759

Zhang, C., Zhou, H., and Duan, J. 2023. A method for identifying and repairing holes on the surface of unorganized point cloud. *Measurement* 230. https://doi.org/10.1016/j.measurement.2023.112575