algorithms

Article

Addressing Air Pollution Challenges: An Integrated Algorithmic Approach Towards Safeguarding Built Heritage

Dimitrios Mitsos [1] and Vassilis Poulopoulos [2,*]

[1] Laboratory of Archaeometry, University of the Peloponnese, Palaio Stratopedo, 24133 Kalamata, Greece; d.mitsos@go.uop.gr
[2] ΓAB LAB-Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, 22131 Tripoli, Greece
* Correspondence: vacilos@uop.gr; Tel.: +30-697-2700-533

Abstract

Air pollution poses significant risks to built heritage, yet traditional methods for diagnosing degradation patterns remain largely fragmented, often relying on isolated data streams and/or subjective comparative interpretations. This study proposes a novel modular workflow that integrates Raman spectroscopy and micro-XRF spectrometry data with user-defined contextual metadata to automate the characterisation of pollution-induced degradation layers on monuments. This method utilises algorithms for peak detection, dimensionality reduction, unsupervised machine learning clustering, variance analysis across centroids, and correlation analysis, as well as steps for data re-encoding and visualisation of the results, allowing for scalable and reproducible analyses on heterogeneous multidimensional datasets. Applied to case studies from Athens, Piraeus, and Eleusis, Greece, the workflow successfully identified pollution sources and degradation patterns, while also quantifying the contribution of features, including contextual variables such as surface orientation and sampling height. The results validate the method's capacity to combine molecular and elemental data streams, to enhance interpretive clarity, and to minimise manual effort and subjectivity. This work showcases the potential of algorithmic approaches in cultural heritage diagnostics to adapt dynamically and incorporate additional datasets and informs future applications of automated methods in the broader field of heritage science.

Keywords: air pollution; data fusion; statistical analysis; machine learning; heritage science

Academic Editor: Roberto Montemanni

Received: 15 July 2025
Revised: 20 September 2025
Accepted: 25 September 2025
Published: 1 October 2025

Citation: Mitsos, D.; Poulopoulos, V. Addressing Air Pollution Challenges: An Integrated Algorithmic Approach Towards Safeguarding Built Heritage. *Algorithms* **2025**, *18*, 619. https://doi.org/10.3390/a18100619

1. Introduction

The preservation of cultural heritage (CH), particularly historic monuments, has become increasingly urgent in the face of rapid urbanisation, industrial activity, and evolving air pollution dynamics [1–4]. Atmospheric pollutants such as sulphur oxides, nitrogen compounds, volatile organic compounds (VOCs), and particulate matter contribute to the formation of complex degradation layers on exposed stone surfaces, which compromise both the aesthetic and structural integrity of historic building materials [5–7].

There have been numerous efforts to address these issues from various angles, while facing both the threats from air pollution and the capabilities of new technologies, especially with the use of artificial intelligence (AI) and, subsequently, algorithms. The main focus of these efforts in the last decade has been on 3D modelling, virtual reality (VR) and augmented reality (AR), classification, and internet of things (IoT) platform configuration [8,9].

For example, ref. [10] showcased the use of VR tools for the preservation of CH through digitisation. Similar studies have implemented algorithms for 3D scanning [11] and modelling [12], as well as assessing the seismic vulnerability of monuments [13]. Regarding the investigation of degradation layers, the use of AI-assisted imaging has proven successful, both with supervised and unsupervised machine learning techniques [14–16]. Other studies have attempted to map the impact of air pollution on built heritage and its susceptibility, using fuzzy k-means [17], and Inverse Distance Weighting (IDW) algorithms [18,19]. Classification of historic buildings has been attempted through deep learning algorithms [20], and quantification of the contribution of atmospheric pollutants to degradation processes has been achieved through phenomenological models [21], although with a noted lack of measurements on physical samples from the monuments. Reversely, the estimation of the environmental impact on monuments has been conducted through physical samples without utilising algorithmic workflows [22].

The importance of using portable instrumentation, such as Raman spectroscopy and X-ray fluorescence (XRF) spectrometry, to perform non-destructive analyses on CH materials is documented [23–25]. Algorithms have been successfully applied to Raman datasets for spectral preprocessing [26,27], spectral searching and matching [28], and Principal Component Analysis (PCA) for dimensionality reduction [29]. These approaches have offered valuable solutions, with deep learning techniques reportedly offering the best results [30]. Similarly, machine learning algorithms have been used for quantitative elemental analyses on ceramics [31], and both qualitative and quantitative XRF measurements on archaeological materials [32,33]. In general, the use of machine learning algorithms in analytical methods for CH materials has focused mainly on case studies involving pigments, paper, metals and ceramics, with a significant lack of similar approaches regarding monuments. Furthermore, the use of highly complex models, such as neural networks, has been proven to make it impossible to determine which feature of the input data was affecting the outcomes of the models [34].

It is evident that traditional CH diagnostics have largely relied on isolated data streams from complementary analytical techniques and human-driven comparative interpretation of the results. This approach, while valuable, remains labour-intensive, subjective in various degrees, and often incapable of integrating heterogeneous multidimensional datasets. Notable efforts have been made in recent years for fusing data from complementary techniques, for example, for the characterisation of minerals [35] and soils [36]. To the best of our knowledge, there is a gap for such efforts focused on CH materials, thus showing the need for unified algorithmic frameworks that can handle data fusion, analysis, and interpretation in an automated process.

In response to these challenges, this study proposes a novel modular method that automates preprocessing and fusion of heterogeneous multidimensional datasets, advanced statistical analysis, dimensionality reduction, and unsupervised machine learning clustering, as well as visualisation of the results. It is designed to be fully reproducible and adaptable to new datasets with minimal adjustments.

Rather than focusing on analytical- and site-specific interpretations, which are under consideration for submission in an additional research article, the present study emphasises the algorithmic workflow design and innovative approach. Case studies from the Temple of Hephaestus and the Church of Ag. Theodoroi in Athens, marble sculptures in Piraeus and the archaeological site of Eleusis, Greece, are employed to validate the method. The hypothesis framing this work is that data fusion, when structured appropriately and coupled with algorithmic methods, can reveal pollution sources and pollution-induced degradation patterns on monuments, quantify the contribution of diverse variables in these patterns, and automate the analytical processes.

2. Materials and Methods

2.1. Overview of the Workflow

The proposed pipeline is developed in the Python programming language (version 3.9) and implemented entirely within Jupyter Notebook (version 7.3.2), a free web-based interactive computational environment. It utilises compartmentalised iterative code cells across separate kernels to ensure transparency, modularity, scalability, and repeatability. The Python libraries used within the context of this study are the following:

1. Pandas, for data manipulation, creation of DataFrames and correlation matrices;
2. Numpy, for numerical operations and array handling;
3. Matplotlib, for creating visualisations;
4. Scikit-learn, for machine learning algorithms;
5. Scipy, for spectral processing and statistical analysis;
6. Seaborn, for enhanced statistical data visualisation.

The workflow integrates Raman spectroscopy and micro-XRF spectrometry data with user-defined contextual metadata (i.e., surface orientation, sampling height, construction chronology, macroscopic evaluation of preservation state) into a unified DataFrame. Subsequent steps include spectral preprocessing, feature extraction, dimensionality reduction, advanced statistical analysis, unsupervised machine learning clustering and visualisation of the results in multiple stages (Figure 1).

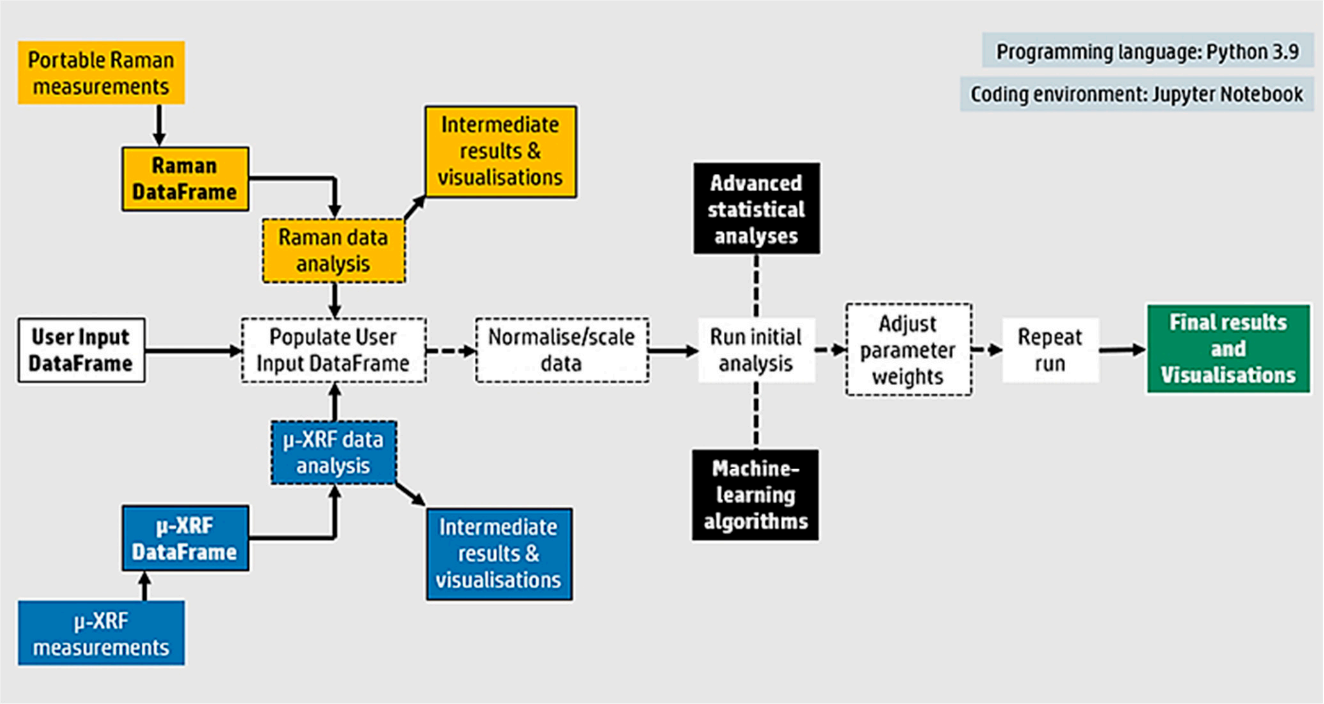


Figure 1. Schematic representation of the proposed workflow, including all the steps followed throughout the pipeline.

2.2. User Input DataFrame

The initial steps of the workflow aim to create and populate a User Input DataFrame, in the form of a .csv file. All steps are implemented through the use of custom-built code scripts (as described in Section 2.1), with the user required to insert contextual metadata and run the automated kernels for each subsequent step in the workflow. More specifically, the user is required to insert values for samples' identification numbers (which are used as vertical placeholders), the sample's surface orientation on the monument, the sampling height, the chronology of the monument, and a subjective condition rating after macroscopic visual inspection. Subsequent columns are populated gradually after separate analyses, which are described in detail in Sections 2.3–2.5. The structure of the User Input DataFrame is presented in Table 1.

Table 1. Example of the structure of the created User Input DataFrame with sample data.

Sample ID	Height	Orientation	Chronology	Macro Rating	Raman Source 1	Raman Source 2	Raman Rating	XRF Source 1	XRF Source 2	XRF Rating
IF1	1.8	N	−450	2	1	4	2	1	4	1
IF3	0.4	W	−450	2	1	2	1	1	2	1
AT1	9	SW	1050	1	3	1	1	1	2	2
AT4	7.1	SE	1050	1	1	2	2	1	3	1
EL1	2.1	NE	200	1	5	2	1	1	2	1
EL16	1.7	W	−450	2	1	2	1	2	1	0
PE1	0.4	NW	−450	2	1	3	2	1	2	1
PE5	0.6	W	−450	1	3	1	1	1	3	1

2.3. Analytical Techniques

2.3.1. Raman Spectroscopy

Raman measurements were performed in situ using a BRAVO handheld Raman spectrometer (Bruker Optics GmbH & Co. KG, Ettlingen, Germany), equipped with dual excitation lasers (DUO LASER™ technology, with wavelengths at 785 nm and 853 nm) and a CCD detector, spectral range of 170–3200 cm^{-1}, and spectral resolution of 10–12 cm^{-1}. Data acquisition and processing were carried out by using the free Spectragryph software (version 1.2.16.1). A total of 174 identified Raman bands were manually cross-referenced with spectral libraries from the free online RUFF database [37]. Baseline correction was avoided, as the algorithm used in the instrumentation's software already performs a similar step when merging the raw spectral data from both lasers, to exclude fluorescence signals (Sequentially Shifted Excitation patented technology by Bruker Optics GmbH & Co. KG (Ettlingen, Germany)). A total of 53 Raman measurements were performed across the four monuments, covering the four main orientations per site.

Raman spectral data were exported in raw format (.txt file) and processed using both the Spectragryph software for initial inspection and custom Python scripts for downstream analysis. Well-established Raman instrumentation and spectral processing methodological steps were used as a basis, as described in [38,39]. Smoothing of the spectra was conducted via a Savitzky–Golay filter algorithm (savgol_filter) from scipy.signal. The preprocessed data were then inserted through custom code scripts into a newly created unified DataFrame (.csv file). The range of wavenumber values (170–3200 cm^{-1}, 2 cm^{-1} step size) populated the first column, while the Raman intensity values measured for each wavenumber and for each sample populated the subsequent columns. Sample IDs were used as horizontal placeholders to represent each sample's Raman spectrum.

2.3.2. Micro-XRF Spectrometry

Elemental analysis was carried out on 46 physical samples, using a custom-built portable micro-XRF instrumentation from the X-ray fluorescence Laboratory of the Institute of Nuclear and Particle Physics at NCSR "Demokritos" (Athens, Greece). Each sample underwent three line-scan measurements across its cross-section's profile, using an unfiltered 50 kV, 600 μÅ X-ray beam and a 10 mm^2 silicon drift detector with a full width at half maximum of 140 eV at 5.89 keV. The line-scan's step size was 0.02 mm, and the acquisition time was set at 30 s per line-scan point measurement.

The presence of elements was determined after visually inspecting the spectra and the concentration of each element was calculated using the PyMCA software (version 5.8.0)through an established methodology [40], and their average concentrations for three line-scans per sample were computed through custom Python code scripts for integration into a unified dataset. Similarly to the Raman data, a unique DataFrame (.csv file) for each sample was created. The distance from the surface of the sample's cross-section (in mm) for each spot measurement populated the first column and the concentration value for each element (in weight %) populated the subsequent columns. The elements' chemical symbols were used as horizontal placeholders.

2.4. Data Fusion Architecture

This study is using a mid-level data fusion (MLF) (or feature level) approach, as it integrates a feature extraction step which can hold adequate original information from the analytical techniques, with the extracted features then combined to build further quantitative or qualitative outcomes [41,42].

After preprocessing, Raman spectra were transformed into binary matrices. More specifically, each Raman signal was encoded as the presence (=1) or absence (=0) of the 174 manually identified peaks, to reduce dimensionality. Similarly, elemental concentrations from the micro-XRF measurements were summarised as mean values for each sample.

User-defined variables (surface orientation, sampling height, construction chronology, macroscopic condition rating) were numerically encoded and normalised using the sklearn.preprocessing.MinMaxScaler algorithm. Orientation string values (N, NE, E, SE, S, SW, W, W or north, northeast, east, southeast, south, southwest, west, northwest) were circularly encoded via sine and cosine transforms in custom code shells. Chronology had a numerical user input value (in years) and was then normalised across all samples. Similarly, height was expressed in metres in user input values, and subsequently normalised to a 0–1 value range, to decrease domination of the variable in subsequent statistical analyses and machine learning techniques. Condition rating had a string user input value, based on macroscopic inspection, which was then ordinally encoded (0 = good, 1 = moderate, 2 = poor), to test the correlation between the visible decay on monuments and the degradation level observed through scientific methods.

A unified DataFrame was generated in the form of a .csv file, with all Raman, micro-XRF, and user-defined features for each sample. A unique Sample ID was used as an identifier for each sample and across all the other features. For testing purposes, an initial score was calculated for all the normalised values in the DataFrame. The feature weights were then automatically adjusted, based on the initial scores, and applied to the data utilised in all subsequent analytical steps.

2.5. Feature Extraction, Clustering and Visualisation

The find_peaks algorithm from scipy.signal was implemented for peak detection in each of the Raman spectra, using dynamic thresholding based on spectrum-specific min/max Raman intensity values. Negative Raman intensity value artefacts were treated as zero. The presence of combinations of specific Raman bands (e.g., 1008 cm^{-1} and 1036 cm^{-1} for gypsum) was used to automatically assign materials or molecular groups, to indicate the two primary pollution sources and provide an additional condition rating feature (titled "Raman state"), based on the molecular data that suggest degradation products. Pollution sources were numerically encoded to represent vehicular emissions (=1), industrial emissions (=2), domestic heating (=3), biological activity (=4), and soil dust (=5). The two main pollution sources according to the Raman measurements (Raman sources 1 and 2) were assigned according to the prevalence of specific pollutants' molecular compounds in the resulting Raman k-means clusters.

Selected elemental concentration ratios (e.g., Cu/Zn, Zn/Pb, Mn/Fe, S/Ca) were calculated for the micro-XRF measurements to identify the two primary pollution source indicators based on literature [43–45], and to assess condition ratings through selected elemental correlation scores, using the scipy.stats.pearsonr algorithm. Additionally, a Pearson's correlation matrix was created using the pandas.DataFrame.corr() algorithm for the elemental concentrations, to show the correlation between the distributions of elements across the samples' profiles.

Subsequently, PCA was applied via the sklearn.decomposition.PCA algorithm, to reduce dimensionality and eventually visualise principal axes of variance through scatter plots. A k-means clustering algorithm (sklearn.cluster.KMeans) was used to create groupings of samples explaining the variances of features across the dataset and to investigate the level of clarity using different types of datasets. These algorithms were applied in multiple stages of the workflow. More specifically, they were applied: (a) in the Raman dataset for each monument separately, (b) in the unified Raman dataset for all monuments with the exclusion of micro-XRF analyses, and (c) in a final step including both analytical techniques, all including the contextual metadata. The optimal number of clusters (k) for the k-means algorithm was determined after calculating silhouette scores (using the sklearn.metrics.silhouette_score algorithm) for a range of k values (1–10). The five most contributing features to each cluster centroid were calculated via the pandas.DataFrame.var() algorithm. Correlation scores between features were then calculated to identify potential underlying links between pollution source fingerprints, degradation patterns, and contextual metadata.

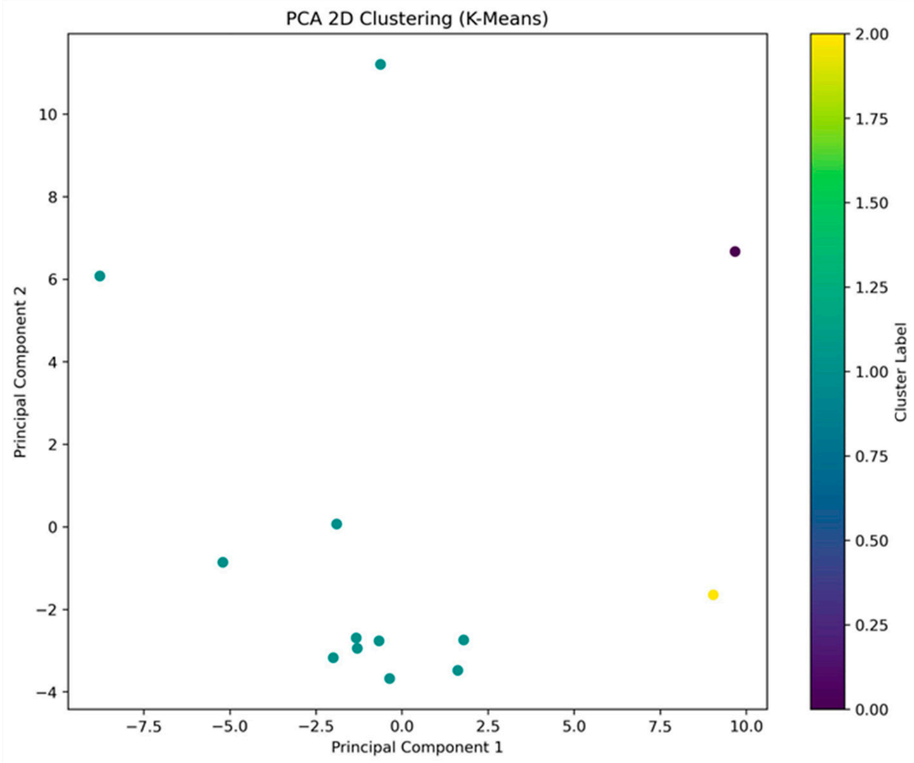
The final visualisations include scatter plots, accompanied by interpretive tables as per the steps described above, heatmaps, and Pearson's correlation matrices. All steps described in this section were automated through custom-built code scripts.

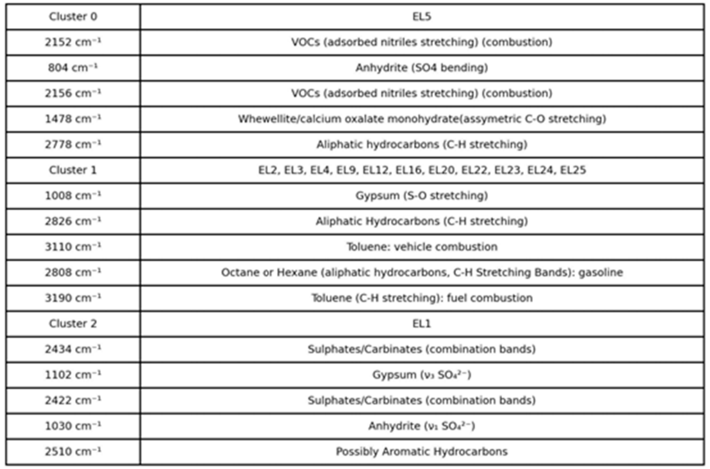
3. Results and Discussion

3.1. Raman Data Analysis

Figure 2 shows the PCA scatter plot generated for the archaeological site of Eleusis, after k-means clustering, with the samples colour-coded by cluster assignment. Additionally, an adjacent table shows the Sample IDs that are included in each cluster, along with the five most contributing Raman bands for each cluster centroid, re-encoded again as string values and presented here as text. For example, a combination of specific Raman bands is here automatically typed as "nitriles".

The results show distinct groupings based on gypsum content and specific pollutants, such as VOCs, polycyclic aromatic hydrocarbons (PAHs), and soot. The resulting scatter plot and table for each monument gives a clear view of sample groupings, the molecular phases that vary across the samples, and the pollutants that are associated with their degradation mechanisms. This is invaluable for both a site-specific conservation strategy, and for further implementation on additional sites or CH materials groups of samples in general.



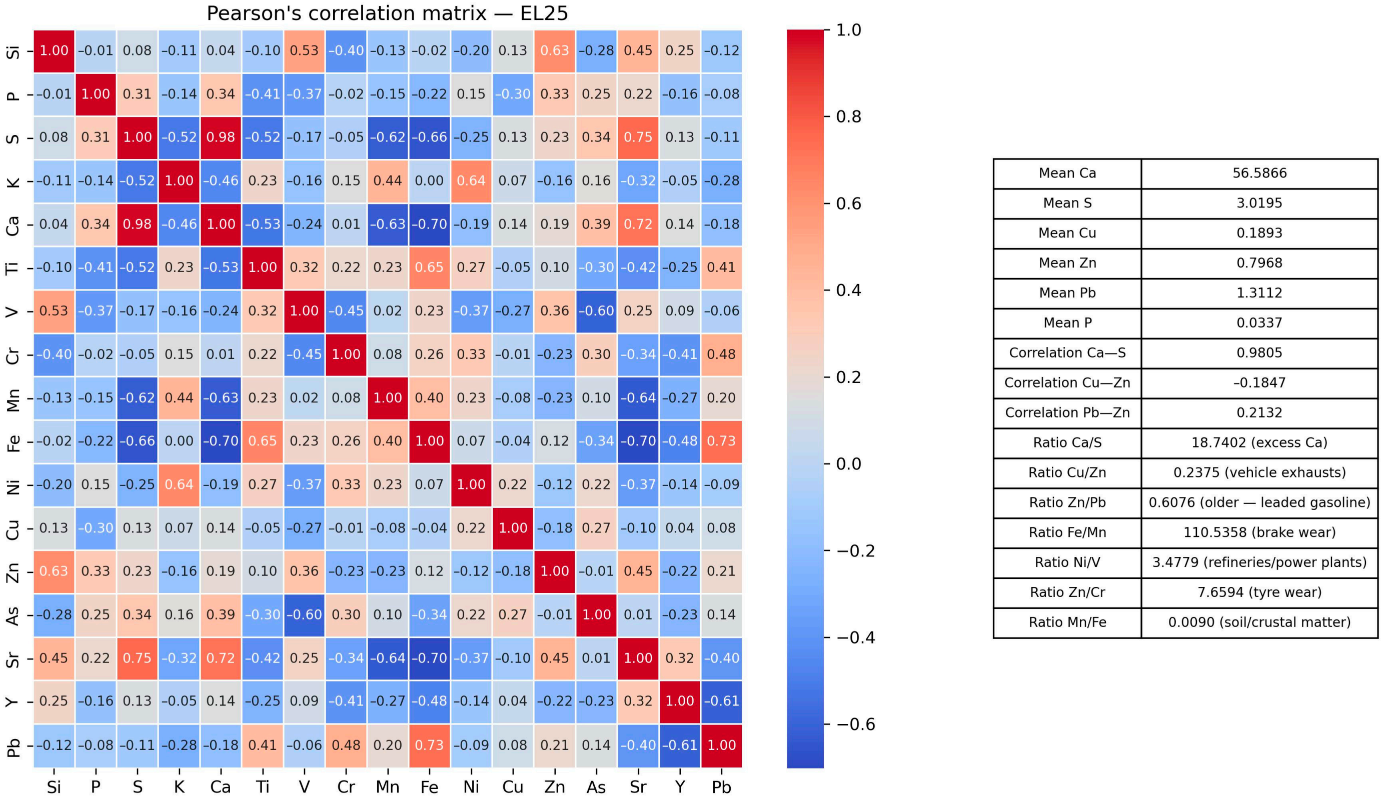


Cluster 0	EL5
2152 cm^{-1}	VOCs (adsorbed nitriles stretching) (combustion)
804 cm^{-1}	Anhydrite (SO4 bending)
2156 cm^{-1}	VOCs (adsorbed nitriles stretching) (combustion)
1478 cm^{-1}	Whewellite/calcium oxalate monohydrate(assymetric C-O stretching)
2778 cm^{-1}	Aliphatic hydrocarbons (C-H stretching)
Cluster 1	EL2, EL3, EL4, EL9, EL12, EL16, EL20, EL22, EL23, EL24, EL25
1008 cm^{-1}	Gypsum (S-O stretching)
2826 cm^{-1}	Aliphatic Hydrocarbons (C-H stretching)
3110 cm^{-1}	Toluene: vehicle combustion
2808 cm^{-1}	Octane or Hexane (aliphatic hydrocarbons, C-H Stretching Bands): gasoline
3190 cm^{-1}	Toluene (C-H stretching): fuel combustion
Cluster 2	EL1
2434 cm^{-1}	Sulphates/Carbinates (combination bands)
1102 cm^{-1}	Gypsum (ν_3 SO_4^{2-})
2422 cm^{-1}	Sulphates/Carbinates (combination bands)
1030 cm^{-1}	Anhydrite (ν_1 SO_4^{2-})
2510 cm^{-1}	Possibly Aromatic Hydrocarbons

Figure 2. Generated PCA scatter plot (after k-means clustering) for the archaeological site of Eleusis, with colour-coded cluster labels, and an adjacent table with Sample IDs groupings in clusters and most contributing molecular phases for the cluster centroids.

3.2. Micro-XRF Data Analysis

Pollution source indicators derived from elemental ratios allowed classification of samples into vehicular emissions, industrial emissions, domestic heating, and soil dust. A Pearson's correlation matrix (Figure 3) for sample EL25 from the site of Eleusis, for example, shows a similar distribution between Ca and other marble constituents (e.g., Sr), as well as between metal-bearing particulate pollutants (e.g., Pb, Ti, Fe), within the sample. The adjacent table provides the mean concentration values, correlation scores and ratios of selected elements, which provide a full profile of the degradation state of the sample and the pollution sources indicated by the particulate pollutants' elemental composition.



	Si	P	S	K	Ca	Ti	V	Cr	Mn	Fe	Ni	Cu	Zn	As	Sr	Y	Pb
Si	1.00	−0.01	0.08	−0.11	0.04	−0.10	0.53	−0.40	−0.13	−0.02	−0.20	0.13	0.63	−0.28	0.45	0.25	−0.12
P	−0.01	1.00	0.31	−0.14	0.34	−0.41	−0.37	−0.02	−0.15	−0.22	0.15	−0.30	0.33	0.25	0.22	−0.16	−0.08
S	0.08	0.31	1.00	−0.52	0.98	−0.52	−0.17	−0.05	−0.62	−0.66	−0.25	0.13	0.23	0.34	0.75	0.13	−0.11
K	−0.11	−0.14	−0.52	1.00	−0.46	0.23	−0.16	0.15	0.44	0.00	0.64	0.07	−0.16	0.16	−0.32	−0.05	−0.28
Ca	0.04	0.34	0.98	−0.46	1.00	−0.53	−0.24	0.01	−0.63	−0.70	−0.19	0.14	0.19	0.39	0.72	0.14	−0.18
Ti	−0.10	−0.41	−0.52	0.23	−0.53	1.00	0.32	0.22	0.23	0.65	0.27	−0.05	0.10	−0.30	−0.42	−0.25	0.41
V	0.53	−0.37	−0.17	−0.16	−0.24	0.32	1.00	−0.45	0.02	0.23	−0.37	−0.27	0.36	−0.60	0.25	0.09	−0.06
Cr	−0.40	−0.02	−0.05	0.15	0.01	0.22	−0.45	1.00	0.08	0.26	0.33	−0.01	−0.23	0.30	−0.34	−0.41	0.48
Mn	−0.13	−0.15	−0.62	0.44	−0.63	0.23	0.02	0.08	1.00	0.40	0.23	−0.08	−0.23	0.10	−0.64	−0.27	0.20
Fe	−0.02	−0.22	−0.66	0.00	−0.70	0.65	0.23	0.26	0.40	1.00	0.07	−0.04	0.12	−0.34	−0.70	−0.48	0.73
Ni	−0.20	0.15	−0.25	0.64	−0.19	0.27	−0.37	0.33	0.23	0.07	1.00	0.22	−0.12	0.22	−0.37	−0.14	−0.09
Cu	0.13	−0.30	0.13	0.07	0.14	−0.05	−0.27	−0.01	−0.08	−0.04	0.22	1.00	−0.18	0.27	−0.10	0.04	0.08
Zn	0.63	0.33	0.23	−0.16	0.19	0.10	0.36	−0.23	−0.23	0.12	−0.12	−0.18	1.00	−0.01	0.45	−0.22	0.21
As	−0.28	0.25	0.34	0.16	0.39	−0.30	−0.60	0.30	0.10	−0.34	0.22	0.27	−0.01	1.00	0.01	−0.23	0.14
Sr	0.45	0.22	0.75	−0.32	0.72	−0.42	0.25	−0.34	−0.64	−0.70	−0.37	−0.10	0.45	0.01	1.00	0.32	−0.40
Y	0.25	−0.16	0.13	−0.05	0.14	−0.25	0.09	−0.41	−0.27	−0.48	−0.14	0.04	−0.22	−0.23	0.32	1.00	−0.61
Pb	−0.12	−0.08	−0.11	−0.28	−0.18	0.41	−0.06	0.48	0.20	0.73	−0.09	0.08	0.21	0.14	−0.40	−0.61	1.00

Mean Ca	56.5866
Mean S	3.0195
Mean Cu	0.1893
Mean Zn	0.7968
Mean Pb	1.3112
Mean P	0.0337
Correlation Ca—S	0.9805
Correlation Cu—Zn	−0.1847
Correlation Pb—Zn	0.2132
Ratio Ca/S	18.7402 (excess Ca)
Ratio Cu/Zn	0.2375 (vehicle exhausts)
Ratio Zn/Pb	0.6076 (older — leaded gasoline)
Ratio Fe/Mn	110.5358 (brake wear)
Ratio Ni/V	3.4779 (refineries/power plants)
Ratio Zn/Cr	7.6594 (tyre wear)
Ratio Mn/Fe	0.0090 (soil/crustal matter)

Figure 3. Generated Pearson's correlation matrix heatmap for the elemental concentrations measured via micro-XRF for sample EL25, and the adjacent table with mean concentration values, correlation scores, and ratios of selected elements.

In the given example, the Ca/S ratio indicates that the degradation level is not high, as there is superfluous Ca content from fresh marble (other than that existing in gypsum). In addition, the majority of the elemental ratios indicate vehicular emissions (exhausts, brake wear, tyre wear, and gasoline combustion) as the main pollution source that participated in the formation of the degradation layer on the monument, although it is situated in a heavily industrialized area.

This showcases the proposed method's capability to reveal patterns that would otherwise be extremely time-consuming to manually investigate for a high number of samples and/or measurements or even misinterpreted or ignored through a mere manual viewing of XRF spectra. It also enhances the interpretation process via visualisation of the results, allowing the user to perform a rapid analysis of complex multidimensional data. Furthermore, the generated outcome can easily be adjusted through minor tweaks in the modular code.

3.3. Fused Data Analysis

Figure 4 illustrates the output for the fused dataset for Raman measurements from all monuments and excluding micro-XRF measurements, with sample labels coloured according to the surface orientation and sample dots coloured according to the cluster they were assigned to.

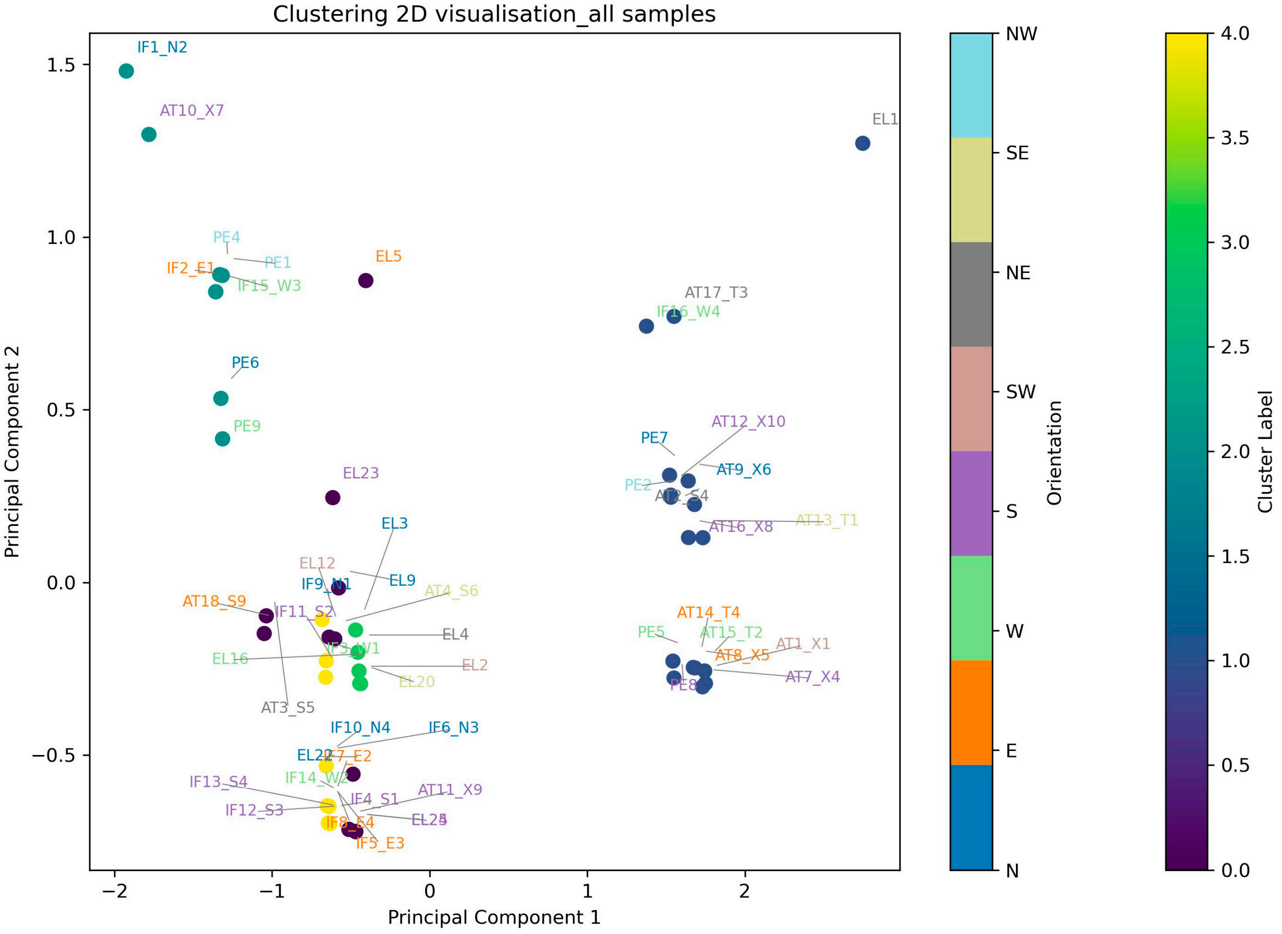


Figure 4. Generated PCA scatter plot (after k-means clustering) for the fused dataset, including all Raman measurements and user-defined contextual metadata, and excluding the micro-XRF measurements. Dots are coloured according to the different cluster assignments, and Sample ID labels are coloured according to the surface orientation of each sample.

Figure 5 shows the score for all the features, based on variance across cluster centroids, through a bar chart, and Figure 6 shows the correlation scores between all the features through a heatmap.

The results show that pollution source assignments dominated the clustering behaviour, while orientation and height had a surprisingly minimal effect, although orientation is traditionally considered a major factor for different types of material degradation [46]. The low contribution score of the "sampling height" parameter can be attributed to the relatively narrow range of heights available in the dataset. High correlation scores can be seen between indicated pollution sources and the state of degradation according to Raman data analysis, as well as between pollution sources and sampling height. Feature-level fusion produced improved cluster coherence and reduced redundancy.

In a final step, Figure 7 shows the outcome for the fused dataset with both Raman and micro-XRF data, along with user-defined contextual metadata.

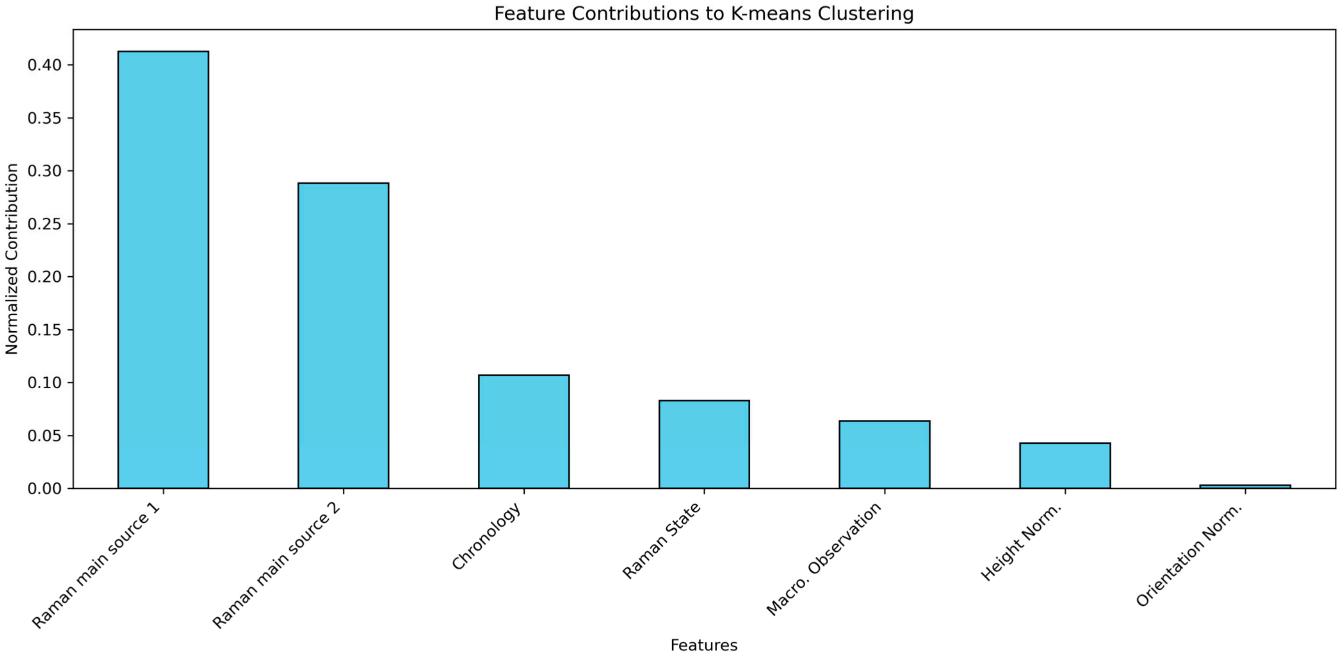
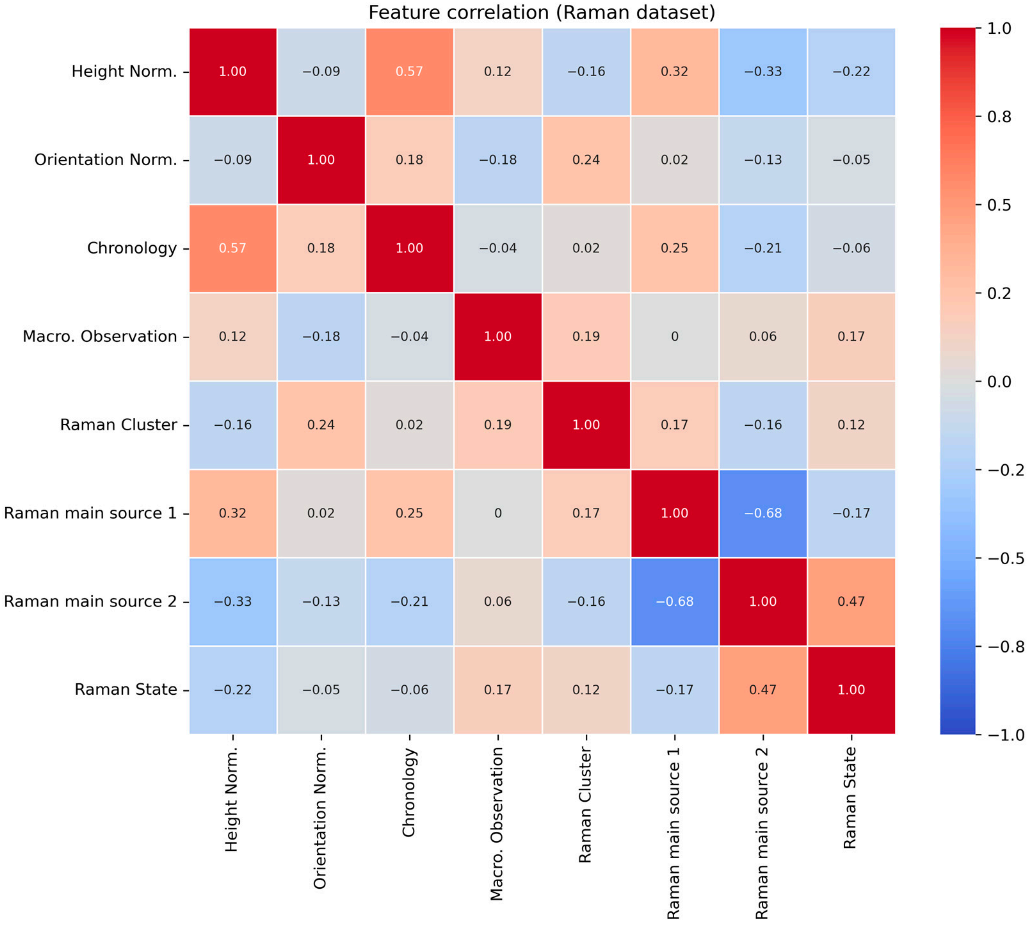


Figure 5. Generated bar chart with the most contributing features to the formation of clusters, excluding the micro-XRF data.



	Height Norm.	Orientation Norm.	Chronology	Macro. Observation	Raman Cluster	Raman main source 1	Raman main source 2	Raman State
Height Norm.	1.00	−0.09	0.57	0.12	−0.16	0.32	−0.33	−0.22
Orientation Norm.	−0.09	1.00	0.18	−0.18	0.24	0.02	−0.13	−0.05
Chronology	0.57	0.18	1.00	−0.04	0.02	0.25	−0.21	−0.06
Macro. Observation	0.12	−0.18	−0.04	1.00	0.19	0	0.06	0.17
Raman Cluster	−0.16	0.24	0.02	0.19	1.00	0.17	−0.16	0.12
Raman main source 1	0.32	0.02	0.25	0	0.17	1.00	−0.68	−0.17
Raman main source 2	−0.33	−0.13	−0.21	0.06	−0.16	−0.68	1.00	0.47
Raman State	−0.22	−0.05	−0.06	0.17	0.12	−0.17	0.47	1.00

Figure 6. Generated heatmap with correlation scores for all the features in the fused Raman dataset, excluding the micro-XRF data.

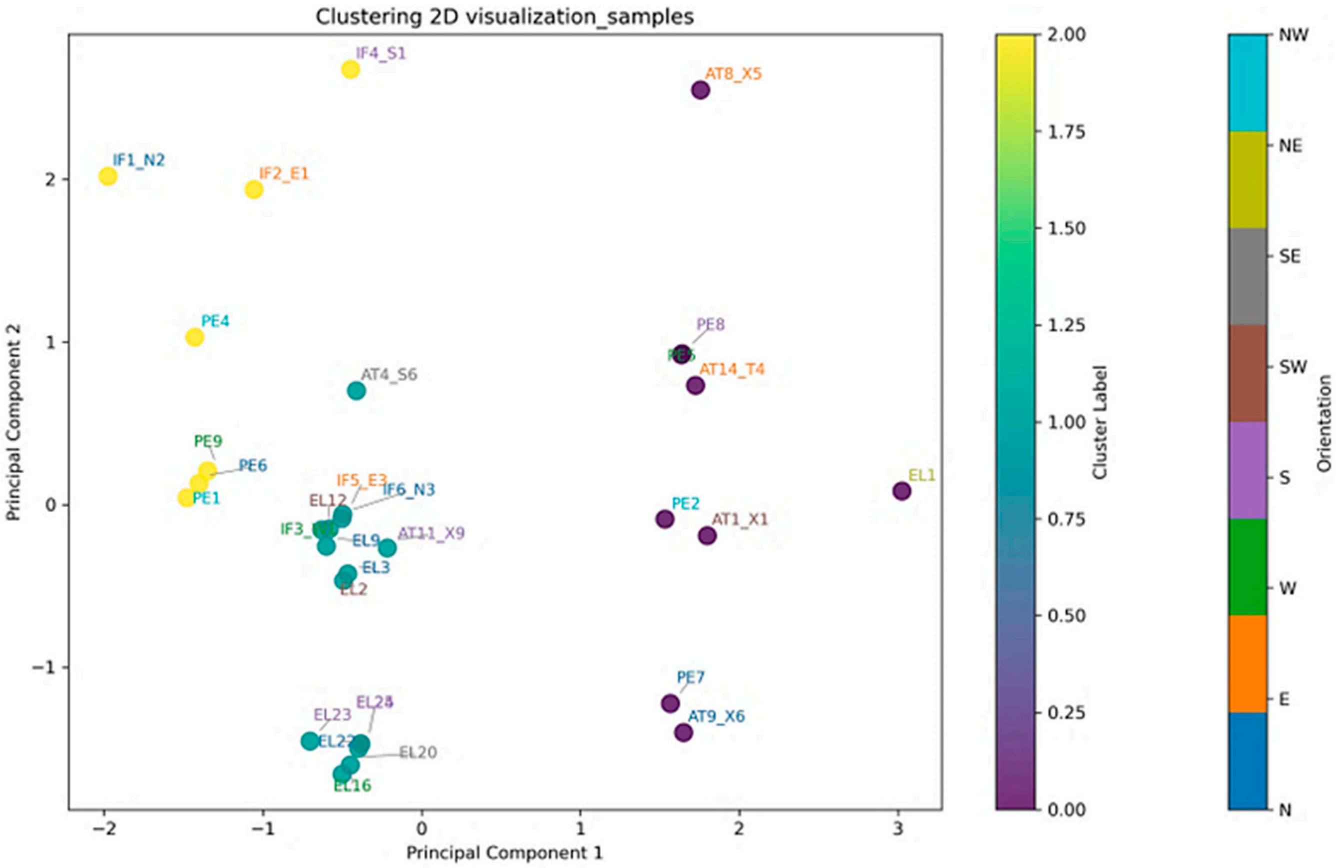


Figure 7. Generated PCA scatter plot (after k-means clustering) for the fused dataset, including both Raman and micro-XRF measurements, as well as user-defined contextual metadata. Dots are coloured according to the different cluster assignments, and Sample ID labels are coloured according to the surface orientation of each sample.

Figure 8 shows the contribution score for all the features, based on variance across cluster centroids, through a bar chart, and Figure 9 shows the correlation scores between all the features through a heatmap.

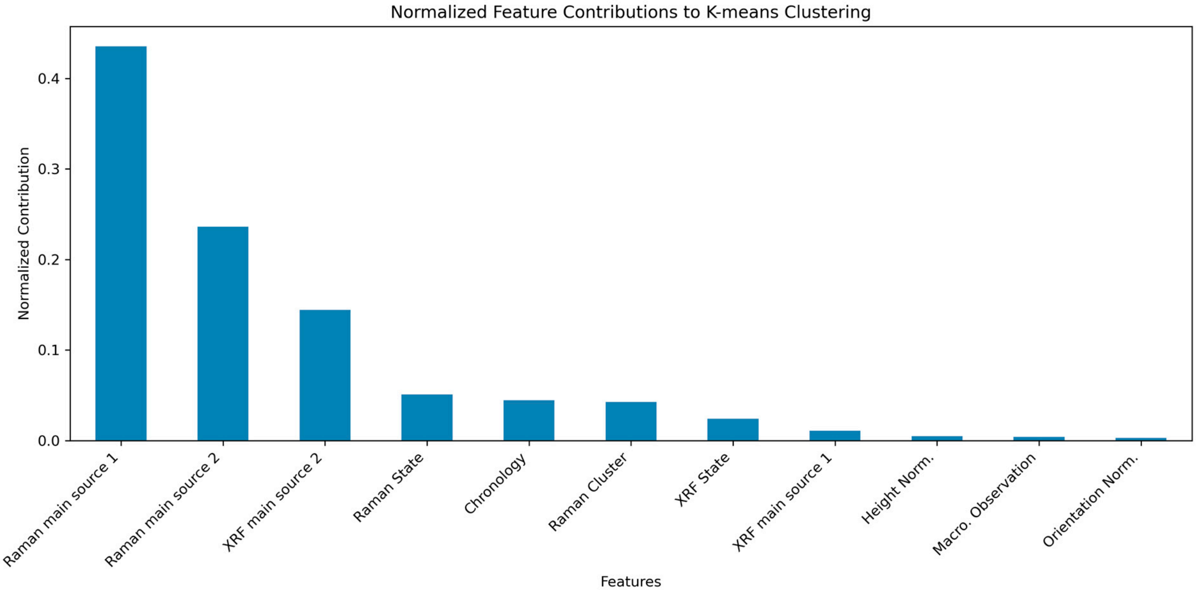
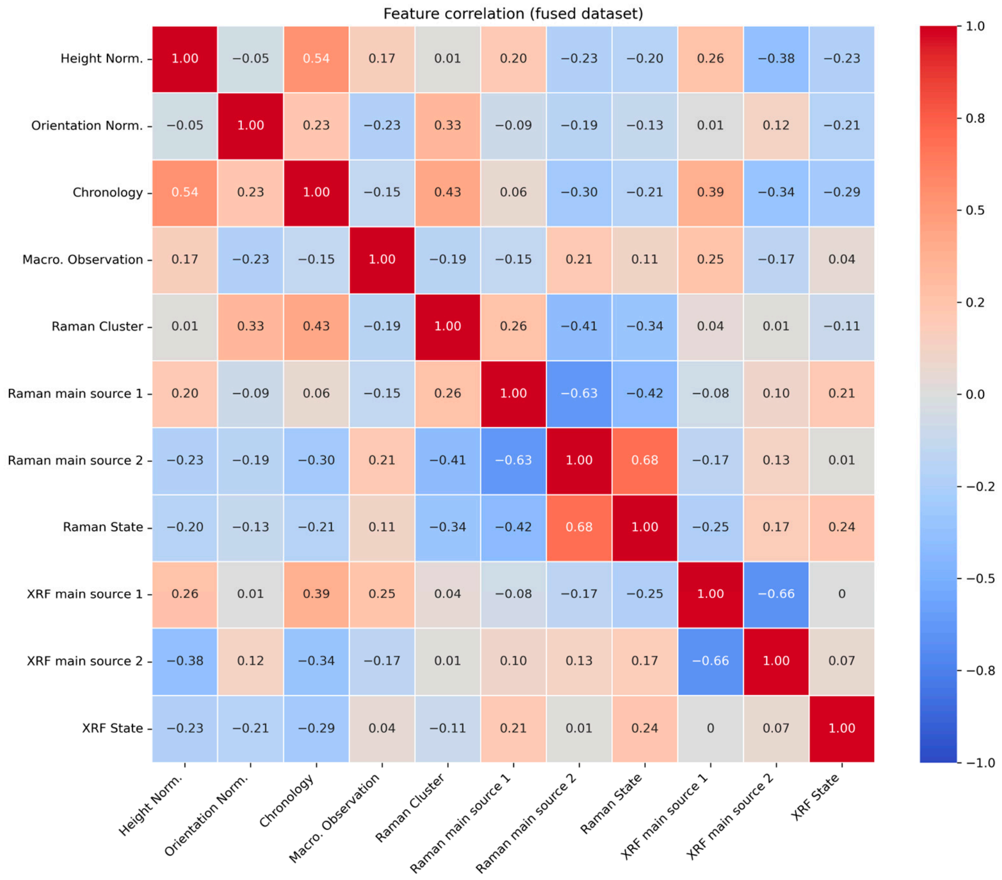


Figure 8. Generated bar chart with the most contributing features to the formation of clusters for the fused dataset.



	Height Norm.	Orientation Norm.	Chronology	Macro. Observation	Raman Cluster	Raman main source 1	Raman main source 2	Raman State	XRF main source 1	XRF main source 2	XRF State
Height Norm.	1.00	−0.05	0.54	0.17	0.01	0.20	−0.23	−0.20	0.26	−0.38	−0.23
Orientation Norm.	−0.05	1.00	0.23	−0.23	0.33	−0.09	−0.19	−0.13	0.01	0.12	−0.21
Chronology	0.54	0.23	1.00	−0.15	0.43	0.06	−0.30	−0.21	0.39	−0.34	−0.29
Macro. Observation	0.17	−0.23	−0.15	1.00	−0.19	−0.15	0.21	0.11	0.25	−0.17	0.04
Raman Cluster	0.01	0.33	0.43	−0.19	1.00	0.26	−0.41	−0.34	0.04	0.01	−0.11
Raman main source 1	0.20	−0.09	0.06	−0.15	0.26	1.00	−0.63	−0.42	−0.08	0.10	0.21
Raman main source 2	−0.23	−0.19	−0.30	0.21	−0.41	−0.63	1.00	0.68	−0.17	0.13	0.01
Raman State	−0.20	−0.13	−0.21	0.11	−0.34	−0.42	0.68	1.00	−0.25	0.17	0.24
XRF main source 1	0.26	0.01	0.39	0.25	0.04	−0.08	−0.17	−0.25	1.00	−0.66	0
XRF main source 2	−0.38	0.12	−0.34	−0.17	0.01	0.10	0.13	0.17	−0.66	1.00	0.07
XRF State	−0.23	−0.21	−0.29	0.04	−0.11	0.21	0.01	0.24	0	0.07	1.00

Figure 9. Generated heatmap with correlation scores for all the features in the fused dataset.

The inclusion of both analytical techniques' datasets contributed towards a clearer clustering of the samples, while also validating the necessary use of complementary techniques for a more complete chemical profile. It is evident that both techniques converge on the same degradation patterns, even when combined with the user-defined contextual metadata. The main pollution source indicated by Raman measurements again was the most contributing feature for the formation of clusters, indicating the variance of molecular phases as significant across the assemblage of samples. Interestingly, the main pollution source indicated by both Raman and micro-XRF measurements had good correlation with sampling height and chronology, indicating that the time of exposure of the samples and the sampling height are connected to different pollutants' deposition patterns. Additionally, the degradation state indicated by both Raman and micro-XRF showed good correlation with each other and the pollution sources indicated by their separate analyses kernels, indicating a good alignment of molecular and elemental information across all samples and of separate workflow kernels.

The results of this study affirm the central hypothesis: that a modular, algorithmic approach to multi-analytical data fusion can provide interpretable, reproducible, and chemically coherent insights into monument degradation patterns. From an algorithmic standpoint, the method succeeded on two critical fronts: first, in addressing data heterogeneity without compromising analytical depth; and second, in maintaining transparency and interpretability across all stages of processing. It also incorporated the integrated contextual metadata and variables that are often omitted or handled post hoc in traditional methodologies.

Furthermore, unlike prior workflows that rely on isolated outputs or static clustering, the modular design employed here enables researchers to iteratively adjust feature weights and re-run analyses without restructuring the code. This flexibility supports site-specific research needs, while at the same time preserving methodological consistency. It also resolves key criticisms in heritage science about the subjectivity of interpretation and the difficulty of scaling traditional approaches to larger or more complex and multidimensional datasets [47].

Importantly, this study suggests a shift from the traditional research question of “what material is present?” to that of “what data architecture best reveals underlying degradation patterns?”. The current workflow’s potential scalability and inclusion of larger datasets or additional features (such as measurements from additional analytical techniques or environmental measurements) could support predictive modelling, comparative site studies, or real-time monitoring when combined with environmental sensors.

Specific limitations regarding this study, include the lack of quantitative validation against ground-truth data, such as direct measurements from known pollution sources or lab-controlled simulated degradation samples. The addition of targeted environmental data, such as temperature, relative humidity, and concentration of various types of pollutants to the proposed pipeline is included in the immediate future steps and continuation of this study. The simulation of degradation on sample materials would provide a deeper understanding of the degradation mechanisms, although it would simultaneously introduce greatly increased equipment and time costs to the current methodology.

Finally, a Graphical User Interface (GUI) front-end would democratise access to the pipeline, allowing non-programmers in conservation practice to apply it to their own datasets. Public release of the codebase with sample datasets and documentation would support community validation, reproducibility, and future collaboration.

4. Conclusions

This study presented a novel modular algorithmic workflow that integrates Raman spectroscopy, micro-XRF spectrometry, and user-defined contextual variables to characterise pollution-induced degradation on historic monuments. The pipeline automates preprocessing, feature encoding and extraction, dimensionality reduction, unsupervised clustering, and visualisation. It offers an alternative to traditional, manually driven methods, that reduce subjectivity and manual effort.

Applied to case studies from Athens, Piraeus, and Eleusis, the workflow demonstrated its capacity to identify pollutant fingerprints and degradation patterns across samples. The analysis revealed that chemical features, particularly those related to pollution sources, were the most significant contributors to cluster formation, while contextual spatial parameters such as surface orientation and sampling height played a comparatively minor role, challenging assumptions about their significance.

The results highlight the value of combining molecular, elemental, and contextual data within a unified algorithmic framework. The pipeline is carefully designed to allow researchers to adjust feature weights, expand datasets, and integrate additional analytical techniques with minimal restructuring or coding skills. Furthermore, the computational resources required for the implementation of the pipeline are limited, as the software and the computational environments used in this study can be downloaded and used free of charge, and the methodology was implemented by using a low-end laptop computer. Future work could build on this approach to support predictive modelling, cross-site comparisons, or the development of a user-friendly GUI for broader adoption in conservation practice. Ultimately, this study underscores the transformative potential of algorithmic pipelines in CH science, not only in terms of efficiency and accuracy, but also in promoting more targeted, data-driven conservation strategies.

Author Contributions: Conceptualisation, D.M. and V.P.; methodology, D.M.; software, D.M.; validation, D.M. and V.P.; formal analysis, D.M.; investigation, D.M.; resources, D.M.; data curation, D.M.; writing—original draft preparation, D.M.; writing—review and editing, D.M. and V.P.; visualisation, D.M.; supervision, V.P.; project administration, V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Access to the full raw Raman and micro-XRF datasets is subject to approval from the Greek Ministry of Culture. Requests for data access can be directed to the corresponding author.

Acknowledgments: Acknowledgements to the Ephorate of Antiquities of Athens, the Ephorate of Antiquities of West Attica, and the Ephorate of Antiquities of Piraeus for the permission to receive material for the analysis, as well as the Laboratory of Archaeometry of the University of the Peloponnese and the X-ray fluorescence Laboratory of NCSR "Demokritos" for providing access to the analytical instrumentation used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vidal, F.; Vicente, R.; Mendes Silva, J. Review of environmental and air pollution impacts on built heritage: 10 questions on corrosion and soiling effects for urban intervention. *J. Cult. Herit.* **2019**, *37*, 273–295. [CrossRef]
2. Vidović, K.; Hočevar, S.; Menart, E.; Drventić, I.; Grgić, I.; Kroflič, A. Impact of air pollution on outdoor cultural heritage objects and decoding the role of particulate matter: A critical review. *Environ. Sci. Pollut. Res.* **2022**, *29*, 46405–46437. [CrossRef]
3. Ruffolo, S.A.; La Russa, M.F.; Rovella, N.; Ricca, M. The impact of air pollution on stone materials. *Environments* **2023**, *10*, 119. [CrossRef]
4. Silva, F.M.; Arreiol, M.; Fragata, A. The Impact of Pollution on Cultural Heritage in the Historic Centre of Porto, Portugal. *Urban Sci.* **2024**, *8*, 31. [CrossRef]
5. Moropoulou, A.; Bisbikou, K.; Torfs, K.; Van Grieken, R.; Zezza, F.; Macri, F. Origin and growth of weathering crusts on ancient marbles in industrial atmosphere. *Atmos. Environ.* **1998**, *32*, 967–982. [CrossRef]
6. Sabbioni, C. Mechanism of Air Pollution Damage to Stone. In *The Effects of Air Pollution on the Built Environment*; Brimblecombe, P., Ed.; Imperial College Press: London, UK, 2003; pp. 63–106. [CrossRef]
7. Mitsos, D.; Kantarelou, V.; Palamara, E.; Karydas, A.G.; Zacharias, N.; Gerasopoulos, E. Characterization of black crust on archaeological marble from the Library of Hadrian in Athens and inferences about contributing pollution sources. *J. Cult. Herit.* **2022**, *53*, 236–243. [CrossRef]
8. Mendoza, M.A.D.; De La Hoz Franco, E.; Gómez, J.E.G. Technologies for the Preservation of Cultural Heritage—A Systematic Review of the Literature. *Sustainability* **2023**, *15*, 1059. [CrossRef]
9. Gîrbacia, F. An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage. *Electronics* **2024**, *13*, 3738. [CrossRef]
10. Li, J. Application of Artificial Intelligence in Cultural Heritage Protection. *J. Phys. Conf. Ser.* **2021**, *1881*, 032007. [CrossRef]
11. Wojcicki, P.; Korga, S.; Milosz, M. Preliminary Application of the Algorithm Highlighting Petroglyph Patterns. *Appl. Sci.* **2022**, *12*, 1660. [CrossRef]
12. Dzwierzynska, J.; Prokop, A. Reconstruction of Historic Monuments—A Dual Approach. *Sustainability* **2022**, *14*, 14651. [CrossRef]
13. Shabani, A.; Kioumarsi, M.; Zucconi, M. State of the art of simplified analytical methods for seismic vulnerability assessment of unreinforced masonry buildings. *Eng. Struct.* **2021**, *239*, 112280. [CrossRef]
14. Scatigno, C.; Festa, G. Neutron Imaging and Learning Algorithms: New Perspectives in Cultural Heritage Applications. *J. Imaging* **2022**, *8*, 284. [CrossRef] [PubMed]
15. Chrysogonos, N.; Lampropoulos, K.; Tzortzis, I.N.; Zafeiropoulos, C.; Doulamis, A.; Doulamis, N. Evaluating the Effectiveness of Unsupervised and Supervised Techniques for Identifying Deteriorations on Cultural Heritage Monuments Using Hyper-Spectral Imagery. In *Transdisciplinary Multispectral Modeling and Cooperation for the Preservation of Cultural Heritage. TMM_CH 2023*; Moropoulou, A., Georgopoulos, A., Ioannides, M., Doulamis, A., Lampropoulos, K., Ronchi, A., Eds.; Communications in Computer and Information Science; Springer: Cham, Switzerland, 2023; Volume 1889, pp. 114–124. [CrossRef]
16. Sazanova, K.V.; Zelenskaya, M.S.; Vlasov, A.D.; Bobir, S.Y.; Yakkonen, K.L.; Vlasov, D.Y. Microorganisms in Superficial Deposits on the Stone Monuments in Saint Petersburg. *Microorganisms* **2022**, *10*, 316. [CrossRef]

17. Sahin, H.; Sözen, A. The tangible impacts of air pollution on built heritage during COVID-19 period on the Historical Peninsula of Istanbul, Turkey. *Environ. Sci. Pollut. Res.* **2025**, *32*, 14202–14219. [CrossRef]
18. Spezzano, P. Mapping the susceptibility of UNESCO World Cultural Heritage sites in Europe to ambient (outdoor) air pollution. *Sci. Total Environ.* **2021**, *754*, 142345. [CrossRef]
19. Xiao, B.; Ning, L.; Lin, Z.; Wang, S.; Zang, H. The Impact of Air Pollution on the Protection of World Cultural Heritage in China. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10226. [CrossRef]
20. Siountri, K.; Anagnostopoulos, C.N. The Classification of Cultural Heritage Buildings in Athens Using Deep Learning Techniques. *Heritage* **2023**, *6*, 3673–3705. [CrossRef]
21. Bonetti, E.; Cavaterra, C.; Freddi, F.; Grasselli, M.; Natalini, R. Chemomechanical Degradation of Monumental Stones: Preliminary Results. In *Mathematical Modeling in Cultural Heritage*; Bonetti, E., Cavaterra, C., Natalini, R., Solci, M., Eds.; Springer INdAM Series; Springer: Cham, Switzerland, 2021; Volume 41, pp. 59–72. [CrossRef]
22. Yi, Y.; Wang, C.; Li, K.; Jia, X.; Wang, C.; Wang, Y. Revealing Black Stains on the Surface of Stone Artifacts from Material Properties to Environmental Sustainability: The Case of Xianling Tomb, China. *Sustainability* **2025**, *17*, 3422. [CrossRef]
23. Casadio, F.; Daher, C.; Bellot-Gurlet, L. Raman Spectroscopy of cultural heritage Materials: Overview of Applications and New Frontiers in Instrumentation, Sampling Modalities, and Data Processing. *Top. Curr. Chem.* **2016**, *374*, 62. [CrossRef]
24. Odelli, E.; Rousaki, A.; Raneri, S.; Vandenabeele, P. Advantages and pitfalls of the use of mobile Raman and XRF systems applied on cultural heritage objects in Tuscany (Italy). *Eur. Phys. J. Plus* **2021**, *136*, 449. [CrossRef]
25. Pozzi, F.; Rizzo, A.; Basso, E.; Angelin, E.M.; de Sá, S.F.; Cucci, C.; Picollo, M. Portable Spectroscopy for Cultural Heritage: Applications and Practical Challenges. In *Portable Spectroscopy and Spectrometry*; Crocombe, R., Leary, P., Kammrath, B., Eds.; Wiley Online Books; John Wiley & Sons: Hoboken, NJ, USA, 2021; Volume 2, pp. 499–522. [CrossRef]
26. Sheehy, G.; Picot, F.; Dallaire, F.; Ember, K.J.; Nguyen, T.; Petrecca, K.; Trudel, D.; Leblond, F. Open-sourced Raman spectroscopy data processing package implementing a baseline removal algorithm validated from multiple datasets acquired in human tissue and biofluids. *J. Biomed. Opt.* **2023**, *28*, 025002. [CrossRef]
27. Offroy, M.; Marchetti, M.; Kauffmann, T.H.; Bourson, P.; Duponchel, L.; Savarese, L.; Mechling, J.M. Using clustering as pre-processing in the framework of signal unmixing for exhaustive exploration of archaeological artefacts in Raman imaging. *Talanta* **2024**, *274*, 125955. [CrossRef]
28. Rousaki, A.; Paolin, E.; Sciutto, G.; Vandenabeele, P. Development and evaluation of a simple Raman spectral searching algorithm. *Eur. Phys. J. Plus* **2021**, *136*, 620. [CrossRef]
29. Coccato, A.; Caggiani, M.C. An overview of Principal Components Analysis approaches in Raman studies of cultural heritage materials. *J. Raman Spectrosc.* **2023**, *55*, 125–147. [CrossRef]
30. Fang, S.; Wu, S.; Chen, Z.; He, C.; Lin, L.L.; Ye, J. Recent progress and applications of Raman spectrum denoising algorithms in chemical and biological analyses: A review. *Trends Analyt. Chem.* **2024**, *172*, 117578. [CrossRef]
31. Li, M.; Ruan, F.; Li, R.; Zhou, J.; Zhang, T.; Tang, H.; Li, H. In situ simultaneous quantitative analysis multi-elements of archaeological ceramics via laser-induced breakdown spectroscopy combined with machine learning strategy. *Microchem. J.* **2022**, *182*, 107928. [CrossRef]
32. Andrić, V.; Gajić-Kvaščev, M.; Crkvenjakov, D.K.; Marić-Stojanović, M.; Gadžurić, S. Evaluation of pattern recognition techniques for the attribution of cultural heritage objects based on the qualitative XRF data. *Microchem. J.* **2021**, *167*, 106267. [CrossRef]
33. Ruschioni, G.; Malchiodi, D.; Zanaboni, A.M.; Bonizzoni, L. Supervised learning algorithms as a tool for archaeology: Classification of ceramic samples described by chemical element concentrations. *J. Archaeol. Sci. Rep.* **2023**, *49*, 103995. [CrossRef]
34. Towarek, A.; Halicz, L.; Matwin, S.; Wagner, B. Machine learning in analytical chemistry for cultural heritage: A comprehensive review. *J. Cult. Herit.* **2024**, *70*, 64–70. [CrossRef]
35. Gibbons, E.; Léveillé, R.; Berlo, K. Data fusion of laser-induced breakdown and Raman spectroscopies: Enhancing clay mineral identification. *Spectrochim. Acta B At. Spectrosc.* **2020**, *170*, 105905. [CrossRef]
36. Zhang, Z.; Wang, Z.; Luo, Y.; Zhang, J.; Feng, X.; Zeng, Q.; Tian, D.; Li, C.; Zhang, Y.; Wang, Y.; et al. 2Quantitative Analysis of Soil Cd Content Based on the Fusion of Vis-NIR and XRF Spectral Data in the Impacted Area of a Metallurgical Slag Site in Gejiu, Yunnan. *Processes* **2023**, *11*, 2714. [CrossRef]
37. RRUFF Project Database. Available online: https://rruff.info/ (accessed on 7 June 2025).
38. Chi, M.; Han, X.; Xu, Y.; Wang, Y.; Shu, F.; Zhou, W.; Wu, Y. An Improved Background-Correction Algorithm for Raman Spectroscopy Based on the Wavelet Transform. *Appl. Spectrosc.* **2018**, *73*, 78–87. [CrossRef]
39. Xu, Y.; Du, P.; Senger, R.; Robertson, J.; Pirkle, J.L. ISREA: An Efficient Peak-Preserving Baseline Correction Algorithm for Raman Spectra. *Appl. Spectrosc.* **2020**, *75*, 34–45. [CrossRef] [PubMed]
40. Solé, V.A.; Papillon, E.; Cotte, M.; Walter, P.; Susini, J. A multiplatform code for the analysis of energy-dispersive X-ray fluorescence spectra. *Spectrochim. Acta B At. Spectrosc.* **2007**, *62*, 63–68. [CrossRef]

41. Smolinska, A.; Engel, J.; Szymanska, E.; Buydens, L.; Blanchet, L. Chapter 3—General Framing of Low-, Mid-, and High-Level Data Fusion with Examples in the Life Sciences. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 51–79. [CrossRef]
42. Robert, C.; Jessep, W.; Sutton, J.J.; Hicks, T.M.; Loeffen, M.; Farouk, M.; Ward, J.F.; Bain, W.E.; Craigie, C.R.; Fraser-Miller, S.J.; et al. Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat. *Food Chem.* **2021**, *361*, 130154. [CrossRef] [PubMed]
43. Valavanidis, A.; Fiotakis, K.; Vlahogianni, T.; Bakeas, E.B.; Triantafillaki, S.; Paraskevopoulou, V.; Dassenakis, M. Characterization of atmospheric particulates, particle-bound transition metals and polycyclic aromatic hydrocarbons of urban air in the centre of Athens (Greece). *Chemosphere* **2006**, *65*, 760–768. [CrossRef]
44. Thorpe, A.; Harrison, R.M. Sources and properties of non-exhaust particulate matter from road traffic: A review. *Sci. Total Environ.* **2008**, *400*, 270–282. [CrossRef]
45. Koukoulakis, K.G.; Chrysohou, E.; Kanellopoulos, P.G.; Karavoltsos, S.; Katsouras, G.; Dassenakis, M.; Nikolelis, D.; Bakeas, E. Trace elements bound to airborne PM10 in a heavily industrialized site nearby Athens. *Atmos. Pollut. Res.* **2019**, *10*, 1347–1356. [CrossRef]
46. Maurenbrecher, P. *Water-Shedding Details Improve Masonry Performance*; Construction Technology Update 23. 1998; Institute for Research in Construction, National Research Council of Canada: Ottawa, ON, Canada, 1998. [CrossRef]
47. Laohaviraphap, N.; Waroonkun, T. Integrating Artificial Intelligence and the Internet of Things in Cultural Heritage Preservation: A Systematic Review of Risk Management and Environmental Monitoring Strategies. *Buildings* **2024**, *14*, 3979. [CrossRef]