npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02403-z

Towards trustworthy AI in cultural heritage

Check for updates

Marina Paolanti[1] ✉, Emanuele Frontoni[1] & Roberto Pierdicca[2]

Artificial intelligence (AI) is increasingly being used in the cultural heritage (CH) sector to analyse, interpret and conserve artefacts and architectural features. While these technologies offer significant opportunities, concerns have been raised regarding transparency, fairness and interpretability. This paper proposes a methodology for fostering trustworthy AI in the CH sector that embeds explainability and bias-mitigation strategies directly into AI-driven analysis. The methodology integrates contextual insights with multidimensional explainability techniques to make AI decision-making processes more transparent and understandable. A case study based on an existing CH analysis framework shows that incorporating explainability can greatly increase user confidence, promote ethical alignment, and encourage responsible use. The findings emphasise the importance of clarifying AI outputs for heritage professionals, as well as ensuring that AI systems respect cultural specificity and interpretive accuracy.

Artificial Intelligence (AI) is increasingly becoming an indispensable tool in the cultural heritage (CH) domain, offering unprecedented opportunities for the study, documentation and interpretation of historic sites[1,2]. Techniques such as image analysis, semantic segmentation, and point cloud processing enable heritage professionals to manage and analyse vast amounts of data, preserve knowledge, and foster deeper engagement with cultural assets[3–6]. As AI methods become more accessible and widespread, they will greatly improve the efficiency and accuracy of heritage documentation, contributing to the sustainable management of cultural resources. However, the broad adoption of AI within the CH community also raises critical questions about trustworthiness. The unique characteristics of heritage data, which are often heterogeneous and context-dependent, expose AI models to biases that may compromise the representation of marginalized cultural representations or give prominence to dominant architectural styles[7–10]. In addition, the inherent non-transparency of some AI algorithms, especially those that use deep learning, poses challenges for heritage professionals who rely on these tools for decision-making.

The intersection of AI and CH is a growing area of research, primarily focused on automated analysis and documentation[11,12]. Techniques such as semantic point cloud segmentation and image-based classification have shown promising results in improving the digital preservation of architectural and archaeological sites[5,13,14]. Early attempts to address the ethical implications of AI in CH have highlighted the need for a structured framework for assessing the impact of these technologies[15]. A pioneering approach in this direction has been to propose key ethical principles specific to CH, including shared responsibility, meaningful participation, accountability, accessibility, sustainability, reliability and dignity[16]. Although general ethical principles for the responsible use of AI have been proposed, there remains a significant gap in the development of sector-specific guidelines tailored to the unique characteristics of CH data. Ethical concerns in CH go beyond general AI ethics and include issues such as social and cultural inclusion, subjectivism, and biases that may arise from algorithmic decisions. This shortcoming highlights the importance of moving from a general, horizontal ethical perspective to a sector-specific approach that takes into account the inherent values, cultural significance and social impact of CH data. Current studies tend to emphasise technical accuracy while neglecting how biases in training data or model interpretation can lead to biased results, potentially leading to misrepresentation of cultural sites. Thus, the development of robust and trustworthy AI practices in CH requires a dedicated framework that not only identifies potential risks, but also promotes sustainable and inclusive AI applications.

To address these gaps, the aim of this paper is to (i) provide an up-to-date, detailed analysis of the most advanced AI techniques for CH analysis and preservation, and (ii) highlight the key ethical issues that need to be considered in their design and use to ensure their development and deployment in accordance with AI ethical principles and requirements for trustworthy AI at the European (EU) level. In particular, we aim to address the following research questions: Can AI ethical principles support the trustworthy design and use of AI applications in CH? If so, how and what do they mean in the field of AI for CH analysis and preservation?

To this end, and given the high-level nature of many of the ethical frameworks proposed in AI ethics literature, we adopt a context-

[1]University of Macerata, Department of Political Sciences, Communication and International Relations, 62100 Macerata, Italy. [2]Dipartimento di Ingegneria Civile, Edile e dell'Architettura (DICEA), Universita' Politecnica delle Marche, Via Brecce Bianche 12, 60131 Ancona, Italy. ✉e-mail: marina.paolanti@unimc.it

sensitive approach that integrates high-level ethical principles with the specific challenges arising from the CH domain. In particular, we draw on commonly adopted AI ethics requirements (see HLEG-AI 2019[17] and ALTAI 2020 (https://altai.insight-centre.org) and adjust them to the CH context, focusing in particular on the ethical requirements of fairness and explainability. We combine a top-down approach, based on the latest AI ethics literature and policy documents, with the technical specificities and practical challenges discussed in recent AI and computer vision research in the CH sector. We present a tailored AI ethics framework that highlights the integration of bias detection and enhanced explainability, to improve both the concrete and perceived trustworthiness of AI systems applied to CH. Through a comprehensive analysis and evaluation of existing AI methodologies, we identify and address significant gaps implementing these systems within CH, focusing on issues of transparency, fairness and explainability. Our research highlights the need to integrate ethical considerations and bias mitigation strategies into AI development, ultimately leading to fairer and more understandable AI solutions tailored to the CH sector. In this way, we aim to fill a critical gap in the existing scholarly literature on AI and ethics by providing a structured and practical framework for managing ethical considerations and guiding various stakeholders in the development and use of trustworthy AI techniques for CH. In addition, our research outlines how ethical considerations and bias prevention strategies can be effectively incorporated into AI development processes, ultimately contributing to the responsible use of AI in the CH domain. From this perspective, our study serves as a practical guide for stakeholders in the CH sector to comply with evolving AI ethical standards and guidelines at the international level. We evaluate the work of Pierdicca et al.[18] as a case study, despite its dated approach, because it addresses a fundamental challenge in the CH field: semantic segmentation of 3D point clouds for accurate modeling of historic buildings (HBIM). The practical relevance of the study lies in the use of an enhanced Dynamic Graph Convolutional Neural Network (DGCNN)[19] for segmentation, which incorporates features such as normal and colour to improve processing. In addition, the ArCH dataset presented in the paper provides a diverse set of labelled point clouds, making it valuable for testing the robustness of the AI[20].

The main contributions of this paper are summarised as follows. First, we present an ethical framework specifically designed to evaluate AI applications in the CH domain, focusing on bias and explainability. Second, we define a set of quantifiable metrics to assess the fairness and transparency of AI models in the CH domain. Third, we demonstrate the application of the proposed framework through a case study on semantic point cloud segmentation in CH, highlighting potential biases and challenges related to explainability. Fourth, we provide insights into the broader implications of trustworthy AI for CH documentation and conservation practices, highlighting the need for ethical considerations to be embedded in the development and deployment of AI models.

The remainder of this paper is structured as follows: Section 2 outlines the proposed methodology, including key metrics for bias detection (SubSection "Bias detection in cultural heritage AI"), explainability (SubSection "Ensuring explainability in AI-driven cultural analysis"), and quantification (SubSection "Quantification and metrics in cultural heritage AI"). Section "Results" presents the results and discussions, highlighting the broader implications of responsible AI in CH analysis. Section "Enhancing trustworthiness in AI-driven cultural heritage analysis" details the proposed strategies for enhancing the trustworthiness of AI-driven cultural analysis. Finally, section 4 concludes the paper and discusses future research directions.

Methods

The promotion of trustworthy AI practices in CH is achieved through the introduction of a methodological framework specifically designed for the integration of ethical considerations into the development and application of AI models. To validate the proposed explainability metrics, we adopted a qualitative evaluation protocol inspired by Tiribelli et al.[21], who used a similar sample size to assess user experience and trust in AI-driven systems. Although the number of participants is limited, the selected experts represent a wide range of relevant stakeholders in the CH domain. This allows us to test the methodological reliability of our framework. This framework builds upon well-established principles of AI ethics and focuses on two key areas that are particularly critical in the CH context: Bias Detection and Explainability. The implementation of AI ethics in the CH domain requires context-specific assumptions to be formulated. In particular, our approach assumes the availability of domain-specific knowledge, such as architectural taxonomies, stylistic terminology and historical metadata, to support the generation of comprehensible explanations. Furthermore, adapting general AI ethics requirements (as defined in the ALTAI framework) to the CH context involves reframing core concepts such as fairness, which in this domain is closely tied to cultural representativeness, and explainability, which must consider visual and historical interpretability. The following sections detail the components of this approach:

- **Bias detection in cultural heritage AI:** This part of the framework aims to identify and mitigate biases that may arise during data collection, model training, or interpretation processes. By systematically analysing datasets, we ensure that they reflect diverse and representative cultural narratives, minimising the risk of marginalisation or bias. Advanced statistical methods and AI techniques are used to identify potential biases, enabling proactive adjustments and improvements.
- **Ensuring explainability in AI-driven cultural analysis:** Given the complexity of CH data, particularly in cases involving 3D point clouds or semantic segmentation, our framework emphasises the importance of making AI models transparent and interpretable. We focus on methods that enhance the clarity of AI-generated outputs, making them understandable to both technical experts and CH professionals. This includes developing strategies for visualising decision paths and clearly explaining model reasoning.
- **Metrics for evaluating AI ethics in CH:** To assess the practical effectiveness of our ethical framework, we propose a set of metrics tailored to the CH context. These include transparency ratings, explainability indices and user comprehension ratings. By measuring how well the explainability components are perceived by different stakeholders, we ensure that the integration of ethical principles into AI practices is both practical and impactful.

Figure 1 illustrates the proposed approach for trustworthy AI in CH. It outlines the key ethical principles, regulatory guidelines and evaluation metrics needed to ensure transparent, fair and responsible AI applications in this domain.

Bias detection in cultural heritage AI

Addressing the ethical challenges associated with AI applications in CH requires a focused examination of bias detection within the AI ethics framework. Bias detection is essential to ensure that AI systems do not reinforce discriminatory practices or produce harmful effects when applied to cultural data. In the context of AI, biases can arise at different stages, including data collection, model training, and decision-making processes. These biases can lead to distorted representations of marginalised or less documented cultural narratives. Our approach builds on the AI ethics principle of justice and fairness as outlined by the European Commission's High-Level Expert Group. We conceptualize this principle by proposing a structured framework that detects and addresses a range of biases, drawing on recent findings[22–26]. The framework emphasises methods for assessing the representativeness and fairness of AI outcomes, with the aim of ensuring that diverse cultural elements are appropriately and fairly represented. To support AI developers, CH professionals, and policymakers, we provide conceptual reference to guide the identification and correction of potential biases early in the AI development cycle (Table 1). This Table not only help to identify bias, but also recommend strategies

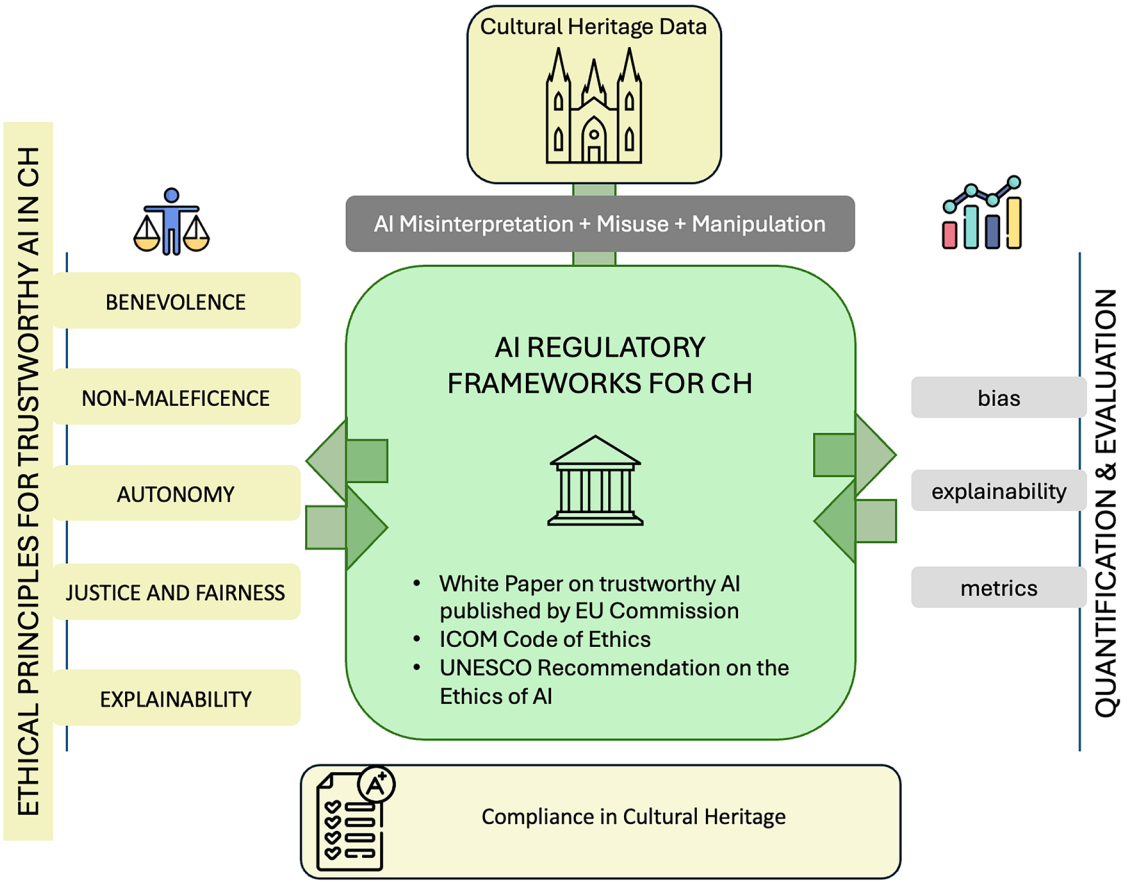

Fig. 1 | Ethical framework for trustworthy AI in CH.

Table 1 | Types of bias and recommended actions for trustworthy AI systems in cultural heritage

Type of Bias	Description	Recommended Action (RA)
Cultural target bias	Occurs when the characteristics of heritage data or cultural elements do not match the intended scope of the AI model, leading to skewed analyses or misinterpretations.	Clearly define the target cultural heritage elements and ensure data collection is inclusive of diverse cultural representations.
Missing data bias	Arises when historical data or cultural records are incomplete, impacting the model's ability to accurately analyze or represent cultural assets.	Implement robust data acquisition strategies, including digitizing undocumented heritage elements and using synthetic data where needed.
Minority bias	Occurs when cultural elements from minority groups are underrepresented, leading to biased heritage interpretations.	Increase data inclusivity by actively collecting and preserving minority cultural narratives and practices.
Informativeness bias	Results from features that inadequately capture the complexity of cultural heritage data, affecting interpretation accuracy.	Incorporate context-specific features that enhance the interpretability of heritage data, particularly for underrepresented cultural contexts.
Temporal bias	Occurs when AI models rely on outdated cultural data, neglecting contemporary heritage transformations or evolving cultural practices.	Continuously update heritage datasets to reflect current practices, traditions, and community perspectives
Contextual bias	Stems from variations in cultural practices and traditions across different regions or communities, leading to inaccurate generalizations.	Account for cultural variations by incorporating localized data and involving cultural experts during model training
Self-selection bias	Occurs when the data predominantly reflects self-documented heritage, overlooking broader community inputs.	Engage diverse community stakeholders to capture a wide range of heritage perspectives and practices.
Historical bias	Involves biases entrenched in historical records, which may reflect past discriminatory practices or dominant cultural perspectives.	Critically evaluate historical datasets for inherent biases and include contemporary reinterpretations of heritage.
Label bias	Arises from inconsistent or subjective labeling of heritage data, leading to varying interpretations.	Standardize labeling processes with input from multidisciplinary heritage experts to ensure consistency and accuracy.
Omitted variable bias	Missing critical contextual or cultural factors in the model can lead to incomplete heritage analyses.	Involve heritage professionals during variable selection to ensure comprehensive coverage of cultural elements.
Aggregation bias	Generalizing cultural patterns from aggregated data can overlook unique local traditions or practices.	Maintain data granularity to respect community-specific cultural practices and avoid overgeneralization.

Fig. 2 | Example of missing data bias: degraded tesserae mosaic from[34]. Loss of tiles leads to gaps in pattern recognition, risking misinterpretation.

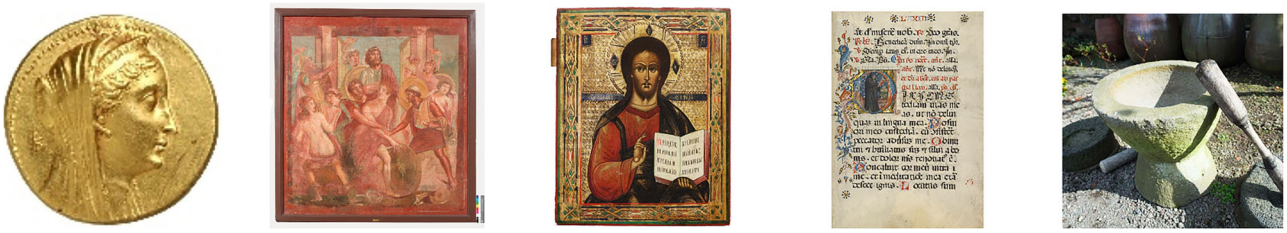

Fig. 3 | Example of population target bias: samples from the SIGNIFICANCE dataset[35] dominated by Eurocentric artifacts.

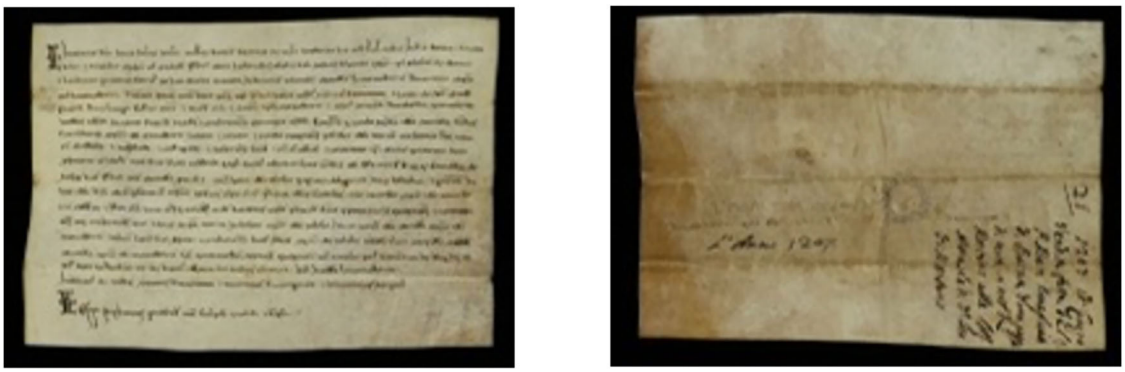

Fig. 4 | Example of label bias: scanned parchments from[36], which may be inconsistently annotated across heritage experts.

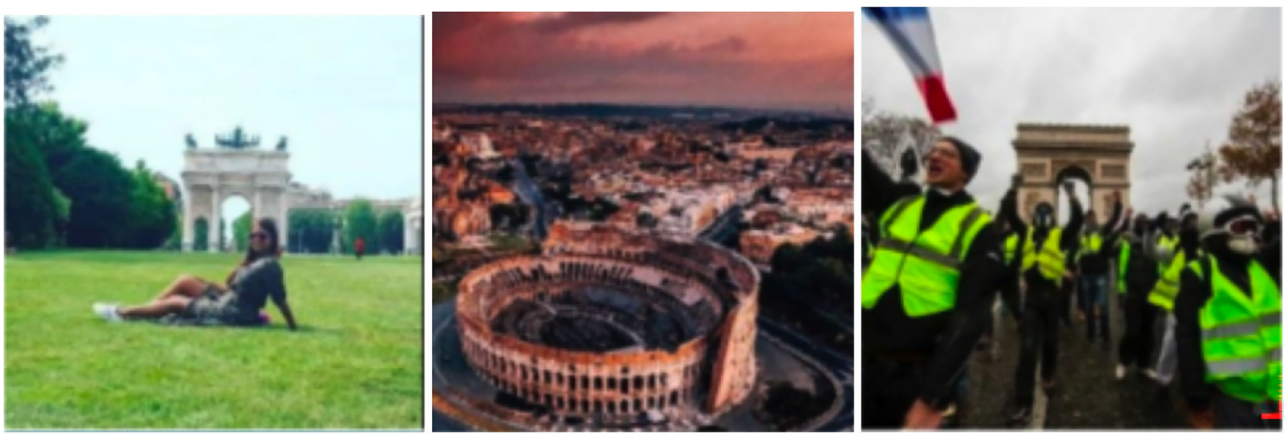

Fig. 5 | Example of contextual bias: social media images from the Sentiment CH dataset[37], shaped by tourists and current events.

to mitigate its impact, ensuring that AI applications in CH are both equitable and reliable.

To illustrate the various types of bias that can influence AI-driven analysis of CH, we have selected five representative examples from publicly accessible CH datasets. Figure 2 illustrates missing data bias, where the incomplete digitisation of
tesserae mosaics can lead to erroneous historical reconstructions if damaged or missing tiles are not properly accounted for. Figure 3 highlights population target bias, as the SIGNIFICANCE dataset predominantly comprises canonical Western artefacts, such as coins, frescoes and icons, which may not adequately represent non-Western or underdocumented cultures. Figure 4 shows an example of label bias, where scanned parchments exhibit transcription and annotation inconsistencies due to subjective interpretation by different experts. Figure 5 represents contextual bias, as images from social media associated with cultural sites are shaped by tourists' perspectives, seasonal effects or social movements, potentially

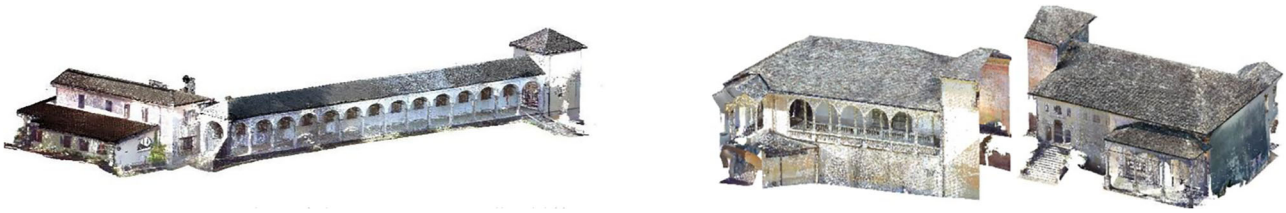

Fig. 6 | An example of a temporal bias is that the static 3D scans from the ArCH dataset[20] do not capture recent transformations in architectures.

Table 2 | Dimensions and levels of explainability in ai-driven cultural heritage analysis

Dimension	Definition	Example in cultural heritage
Computational	How the algorithm produces any output related to cultural data analysis.	Analyzing architectural features from 3D point clouds to identify structural elements.
Mechanistic	Why the algorithm produced a specific output related to cultural interpretation.	Determining structural stability based on detected geometric anomalies in heritage structures.
Justificatory	Why the output is correct or valid.	Verifying the classification of a fresco restoration technique based on color analysis.
Informative	What the output means in practical terms for heritage preservation.	Assessing conservation needs based on the detection of deterioration patterns.
Cautionary	The degree of uncertainty behind the output in heritage analysis.	Providing accuracy metrics for the classification of architectural features.
Expalinability by Level		
Global	Provides a global understanding of the AI model's logic in analyzing heritage data.	Analyzing overall patterns in point cloud segmentation for heritage sites.
Local	Segmenting the solution space to provide explanations for less complex parts of heritage data.	Explaining why a specific architectural feature was classified as Gothic or Romanesque.
Semi-Local	Combining global and local explanations to provide insights on individual predictions and overall model characteristics.	Analyzing a specific segment of a heritage site to understand why it is categorized differently from the rest.

distorting local cultural narratives. Finally, Fig. 6 exemplifies temporal bias, with static 3D scans from the ArCH dataset capturing the built environment at a fixed point in time and neglecting recent restorations or evolving community practices. These examples highlight the importance of bias-aware methodologies in CH to ensure more accurate and inclusive AI-driven interpretations.

Ensuring explainability in AI-driven cultural analysis

In order to promote transparency and the responsible use of AI in the CH domain, it is essential to establish a well-defined approach to accountability. Such an approach is crucial not only for improving the clarity of AI models, but also for fostering trust among diverse stakeholders, including CH professionals, AI developers, and policy makers. By making AI processes understandable, we ensure that the insights derived from AI systems are accessible to both technical and non-technical audiences.

Our approach takes a multidimensional perspective on explainability, building on recent advances in the field[27,28]. It incorporates several dimensions, each of which addresses a different aspect of how AI models handle data and produce interpretable results:

- **Operational explainability:** Focuses on the technical processes within the AI system, detailing how algorithms work and how they process cultural data.
- **Contextual explainability:** Examines how specific cultural or historical contexts influence the AI's decision-making processes, ensuring that heritage-specific nuances are taken into account.
- **Reasoning explainability:** Articulates the reasoning behind the AI's results, highlighting the cultural and contextual factors that influence predictions.
- **Practical explainability:** Aims to clearly communicate the practical implications of the AI's outputs, particularly in the context of heritage conservation and interpretation.
- **Confidence explainability:** Communicates the levels of certainty associated with AI results, helping stakeholders to assess the reliability and robustness of the results.

In addition, the approach distinguishes between global and local levels of explainability, as well as intermediate (semi-local) levels. Global explainability provides a holistic understanding of the model's logic and behavior, while local explainability focuses on explaining specific predictions or decisions. Semi-local explainability addresses explanations within subsets of data, which is particularly relevant when dealing with diverse cultural datasets. By adopting this comprehensive approach, AI developers and CH professionals can systematically address the explainability challenges inherent in CH applications. This approach not only facilitates greater transparency and ethical compliance, but also ensures that AI-driven interpretations are both contextually grounded and trustworthy. Adopting this comprehensive approach enables AI developers and CH professionals to systematically address the explainability challenges inherent in CH applications. This approach facilitates greater transparency and ethical compliance, ensuring that AI-driven interpretations are contextually grounded and trustworthy. Table 2 provides a summary of these explainability dimensions and levels, along with illustrative examples from the CH domain.

Quantification and metrics in cultural heritage AI

Inspired by Tiribelli et al.'s approach to analysing customer behaviour in the retail sector[21], we adapt their methodology to assess explainability in AI-driven CH analysis. Leveraging an established AI explainability literature[29], we define a set of metrics that provide a stakeholder-oriented perspective on the explainability of AI systems. These metrics are essential for assessing the transparency of AI decision-making processes and how they are perceived by CH professionals and other stakeholders. The proposed metrics are organized into four main categories:

- **Goodness and satisfaction**: Assesses the overall satisfaction of the user and the quality of the explanations provided by the AI system. This metric examines whether the explanations improve understanding of the system's operations and whether the information presented is appropriate and actionable.
- **Curiosity**: Measures the system's ability to address user queries comprehensively, including hypothetical scenarios and alternative decision paths, thereby encouraging deeper engagement with the AI's functionality.
- **Trust**: Assesses user confidence in the AI system, focusing on predictability, reliability and the perception of safety when relying on AI-based decisions in cultural analysis.
- **Performance**: Analyses the impact of explainability on user outcomes, in particular how understanding the AI system influences decision making and improves interaction quality.

Although the evaluation was conducted within the CH domain, the metrics were designed to be generalisable to other fields where ethics are a key consideration. Following the approach of Tiribelli et al.[21], we adopted a sample of twelve multidisciplinary participants to ensure a diverse yet consistent methodological assessment of AI explainability and trustworthiness. To ensure diverse perspectives, we formed two focus groups consisting of different stakeholders, including ethicists, engineers, CH experts and public policy makers (see Table 3 for more details). Each participant rated the AI system on a scale of 0 to 10 for each metric, with consensus discussions held to resolve significant discrepancies in scores. The final scores reflect a balanced and collaborative assessment, following a structured consensus-building process as outlined by Pokholkova et al.[30]. In order to ensure consistency and reduce potential bias due to differences in background, each member's input was given equal weight in the calculation of the average rating. A comprehensive overview of the metrics and the specific aspects they address is provided in Table 4. This table provides a structured presentation of each metric, along with a quantification scale, to facilitate a systematic assessment of the explainability of our AI systems.

Table 3 | Background distribution of participants in the study group

Stakeholder group	# Participants
Ethicists	2
Engineers	2
Cultural Heritage Experts	2
Privacy Experts	2
Members of Society at Large	2
Public Decision-Makers (City Level, EU and US)	2

Results

In this study, we extend the methodology of Tiribelli et al.[21] on retail environments to the domain of CH, evaluating the deep learning system proposed by Pierdicca et al.[18]. This work introduced a deep learning framework for analysing 3D point cloud data, with the aim of semantically segmenting architectural elements from historic buildings. The model used a modified version of DGCNN enriched with additional features such as normals and colours to improve classification accuracy. This approach was applied to the ArCH dataset, which consists of labelled point clouds of various architectural elements from heritage sites. Although this study focuses on evaluating a model that was not originally designed with fairness constraints in mind, the proposed framework can guide the integration of technical interventions, such as data augmentation to enhance stylistic diversity and balanced sampling across architectural periods, as well as fairness-aware loss functions, to reduce bias in future AI systems for CH analysis. However, despite the technical achievements of this framework, challenges remain in terms of fairness, transparency and explainability. The framework was evaluated using our ethical approach for AI in CH, focusing on bias detection and explainability. Specifically, the evaluation addressed the following biases:

- **Population targeting bias**: The dataset used may not adequately cover the diversity of architectural styles and historical contexts, potentially leading to biased representations.
- **Contextual bias**: Failure to account for regional variations in architectural features can result in misclassification or generalization errors.
- **Historical bias**: Outdated interpretations may be perpetuated by relying on older, static datasets that may not reflect recent restoration or conservation efforts.

Another critical limitation of [18] is its black box nature. The system lacks transparency in decision-making, making it challenging for heritage professionals to understand how specific classifications are arrived at. This opacity raises concerns about accountability, particularly when the results are used to inform conservation strategies or heritage interpretation.

Evaluation against our explainability metrics reveals significant gaps in this work's ability to make AI decisions under-standable and interpretable. The results, summarised in Table 5, highlight that while the framework demonstrates moderate performance in terms of trust, it shows limited capability in providing comprehensive explanations of its outputs, particularly regarding user understanding and satisfaction. In particular, the system performs poorly in helping heritage professionals understand how the model works (scores: 3-4). Similarly, the explanations provided lack depth and do not adequately address users' curiosity about alternative

Table 4 | Explainability metrics for AI-driven cultural heritage analyses

Category	Indicator	Scale (1–10)
Goodness and Satisfaction	Does the explanation help the user to understand how the system works?	1–10
	Is the explanation satisfying?	1–10
	Is the explanation sufficiently detailed and complete?	1–10
	Is the explanation actionable?	1–10
	Does the explanation indicate the system's accuracy and reliability?	1–10
Curiosity	Does the system address user questions and hypothetical scenarios?	1–10
	Can the system explain why it did not choose an alternative action?	1–10
Trust	Does the user feel confident in the AI system's performance?	1–10
	Are the AI system's outputs predictable and reliable?	1–10
	Does the user feel safe relying on the system?	1–10
Performance	Does understanding the AI system improve user performance?	1–10
	Does trust in the system positively impact outcomes?	1–10

interpretations (scores: 2-3). While the model is moderately efficient and predictable (scores 4-5), these attributes alone do not compensate for the lack of transparency and accountability. These results highlight the urgent need to develop more transparent AI methods in the CH field. Integrating accountability into AI-driven heritage analysis would improve stakeholder trust and ensure that the application of technology is consistent with ethical standards, particularly when influencing conservation practices and cultural interpretation.

Enhancing trustworthiness in AI-driven cultural heritage analysis

To address the challenges identified in ref. 18, particularly those related to explainability and bias, we propose several strategic enhancements to increase the trustworthiness of AI applications in CH. These improvements not only address current limitations, but also provide a foundation for future AI-driven heritage analysis systems. The work in ref. 18, originally designed for the semantic segmentation of architectural heritage elements, faces challenges similar to those in other domains when it comes to transparency and fairness. Given the importance of preserving historical accuracy and cultural authenticity, it is essential to develop explicable models that go beyond traditional black-box approaches. Inspired by the approach in ref. 21 to consumer behaviour analysis in retail, we adapt similar explainability strategies tailored to CH contexts. One of the main tasks in the work of Pierdicca et al.[18] is the semantic segmentation of architectural elements, where the challenge is to explain why certain parts are labelled as belonging to a particular architectural style or period. The use of transparent and interpretable models, such as BubblEX[31,32], can significantly improve explainability by associating segmentation decisions with prototypical parts. For example, when identifying elements such as Gothic arches or Romanesque columns, BubblEX can highlight features that match learned prototypes, making the decision process more intuitive.

Unlike post-hoc methods like Seg-XRes-CAM[33], which are unsuitable for real-time applications, BubblEX integrates explainability directly into the segmentation process, aligning well with the need for transparent and privacy-preserving AI applications in heritage analysis. By embedding interpretability within the model architecture, BubblEX provides localized explanations that are essential for verifying the historical and stylistic classification of architectural elements.

To further enhance trustworthiness, we integrate a multi-dimensional explainability approach that includes:

- **Contextual explainability**: Ensuring that identified architectural features are contextually accurate by considering the historical and cultural background of the site.
- **Mechanistic explainability**: Clarifying the algorithm's reasoning behind labeling specific elements as belonging to a particular architectural style.
- **Interactive explainability**: Allowing experts to query the model and receive detailed insights into why certain features were classified as they were.

By incorporating these explainability dimensions, the work in ref. 18 becomes more aligned with ethical AI practices, particularly in environments where preserving cultural integrity is paramount. In addition, we propose a hybrid approach that combines real-time interpretability with

Table 5 | Evaluation of the work introduced by Pierdicca et al.[13] based on explainability metrics.

Category	Mean Score (1–10)
Goodness and satisfaction	3.0
Curiosity	3.0
Trust	4.5
Performance	3.2

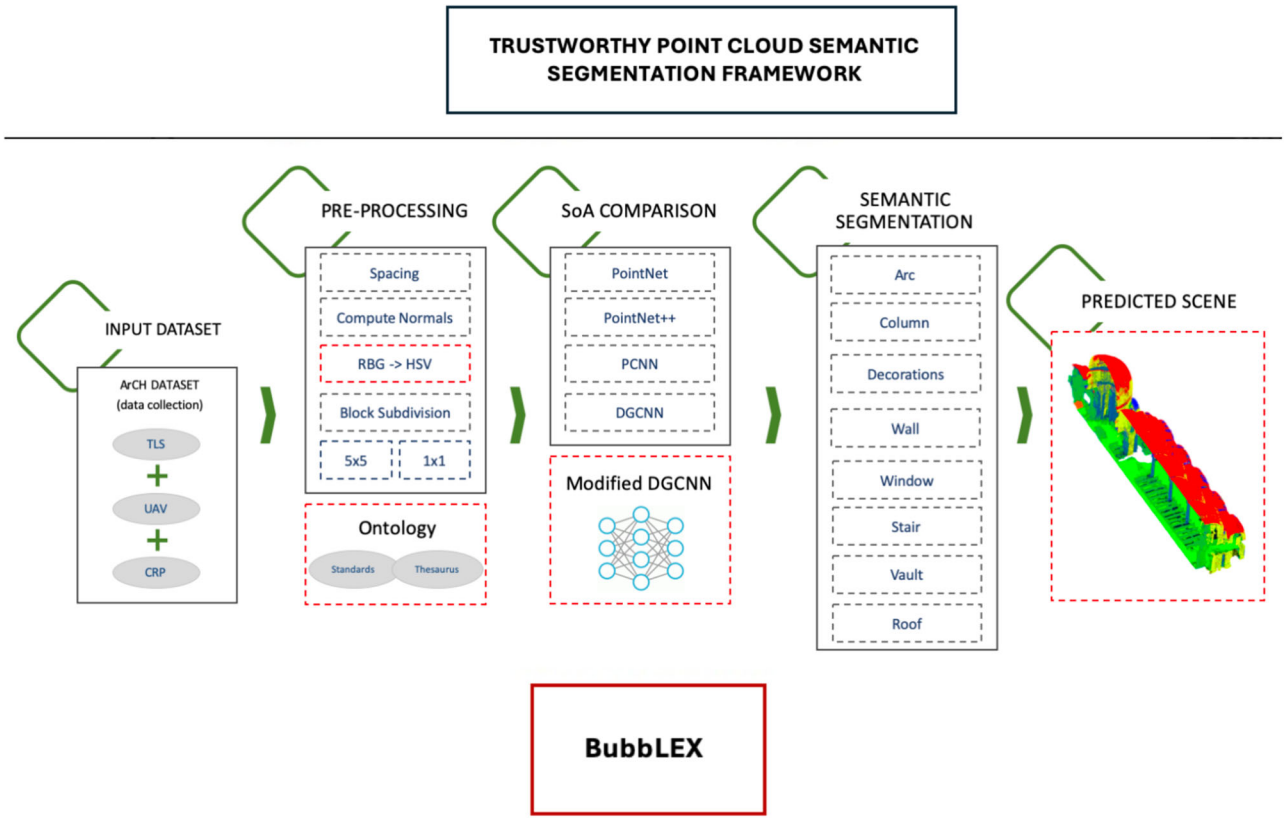

Fig. 7 | Different steps of a possible explainable framework for CH point cloud sematic segmentation which leverages on BubbLEX to integrate explainability directly into the segmentation process.

model transparency. This involves using BubblEX for segmentation tasks, while also applying contextual annotations derived from architectural databases. Such integration ensures that both the AI-driven segmentation and the contextual justification are transparent and verifiable. The proposed approach ensures that automated analyses respect the nuanced complexities inherent to heritage documentation and preservation. Figure 7 schematically illustrates the architecture of a trustworthy point cloud semantic segmentation framework for CH. It demonstrates how explainability can be embedded directly into the processing pipeline, from raw data input and pre-processing through to semantic segmentation, using BubblEX to associate architectural elements with interpretable prototypical parts.

Discussion

In this paper, we addressed the challenge of enhancing trustworthiness in AI-driven CH analysis, focusing on improving explainability and reducing bias in the context of semantic segmentation. Building on the pioneering work of Pierdicca et al.[18], we proposed strategic improvements inspired by approaches successfully implemented in other domains, such as consumer behavior analysis[21]. By integrating transparent models such as BubblEX[31,32], we demonstrated how embedding explainability directly into the AI pipeline can significantly increase user confidence and interpretability, particularly in sensitive heritage applications. The main contributions of this work include the proposal of a hybrid approach that combines real-time interpretability with contextual explainability for architectural segmentation. We addressed the challenges of transparency and fairness by using models that provide localised explanations directly linked to learned prototypes. Furthermore, we demonstrated the practical implementation of enhanced explainability metrics to assess model performance in CH contexts. Although the proposed methodology is tailored to the specific context of CH, its structure, which is based on bias detection and multidimensional explainability, can be extended to other critical areas such as healthcare, education and environmental monitoring. Ethical concerns around transparency, fairness, and stakeholder trust also arise in these areas. While adapting the methodology would require the contextualisation of ethical concerns, datasets and norms specific to each sector, the general principles would remain applicable across sectors. While the proposed methodology has been validated through a representative case study, its robustness when applied to different datasets is still to be determined. As such, further research will involve applying the framework to multiple CH datasets with different properties, including variations in scale, modality and cultural context, to assess its generalisability. Additionally, extending the evaluation to a larger, more diverse group of stakeholders will strengthen the empirical validity of the proposed metrics and enhance the reliability of the insights obtained. Future work will focus on further refining the hybrid approach to include a broader range of explainability metrics, particularly those that address temporal changes in heritage sites. In addition, we plan to extend the model evaluation to different architectural styles and time periods to ensure robustness and fairness across different cultural contexts. Another avenue for future research will be to explore the integration of multimodal data sources, such as textual descriptions and historical records, to complement visual analysis and enhance contextual understanding. By continuously advancing the integration of explainable AI techniques into CH applications, we aim to establish robust, ethically robust frameworks that support sustainable and transparent heritage conservation practices.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Received: 31 October 2025; Accepted: 21 February 2026;
Published online: 03 March 2026

References

1. Fiorucci, M. et al. Machine learning for cultural heritage: A survey. *Pattern Recognit. Lett.* **133**, 102–108 (2020).
2. Xing, Y., Gan, W. & Chen, Q. Artificial intelligence in landscape architecture: a survey. *Int. J. Mach. Learn. Cybern*. 1–26 (2025).
3. Korro, J., Valle-Melón, J. M., Zornoza-Indart, A. & Rodríguez Miranda, Á. New perspectives and usual challenges: present technologies for document management in architectural heritage conservation-restoration works. *ACM J. on Comput. Cult. Herit*. (2025).
4. Jiménez-Díaz, G. et al. Interpretable clusters for representing citizens' sense of belonging through interaction with cultural heritage. *ACM J. on Comput. Cult. Herit.* **17**, 1–22 (2025).
5. Pang, B. et al. Automated heritage building component recognition and modelling based on local features. *J. Cult. Herit.* **71**, 252–264 (2025).
6. Zhao, J. et al. Semantic segmentation of point clouds of ancient buildings based on weak supervision. *Herit. Sci.* **12**, 232 (2024).
7. Zhou, Z., Xi, Y., Xing, S. & Chen, Y. Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques. *Artif. Intell. Mach. Learn. Rev.* **5**, 28–40 (2024).
8. Wang, X., Zhang, J., Cenci, J. & Becue, V. Spatial distribution characteristics and influencing factors of the world architectural heritage. *Heritage* **4**, 2942–2959 (2021).
9. Karadag, I. Machine learning for conservation of architectural heritage. *Open House Int* **48**, 23–37 (2023).
10. Foka, A. & Griffin, G. Ai, cultural heritage, and bias: Some key queries that arise from the use of genai. *Heritage* **7**, 6125–6136 (2024).
11. Jain, V., Mohanan, P. & Naira, M. Role of artificial intelligence in management and preservation of old text through new tech. *Artif. Intell. Businesses: How to Dev. Strateg. for Innov*. 25–38 (2025).
12. Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W. & Choudhury, T. Exploring ai-driven approaches for unstructured document analysis and future horizons. *J. Big Data* **11**, 92 (2024).
13. Galantucci, R. A., Musicco, A., Verdoscia, C. & Fatiguso, F. Machine learning for the semi-automatic 3d decay segmentation and mapping of heritage assets. *Int. J. Archit. Herit.* **19**, 389–407 (2025).
14. Seo, H., Sihag, P., Fu, L. & Kim, D. Novel rcc fusion machine learning method for automatic damage detection of heritage buildings using 3d point cloud data. *J. Civ. Struct. Heal. Monit*. 1–18 (2025).
15. Pansoni, S., Tiribelli, S., Paolanti, M., Frontoni, E. & Giovanola, B. Design of an ethical framework for artificial intelligence in cultural heritage. In *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 1–5 (IEEE, 2023).
16. Pansoni, S. et al. Artificial intelligence and cultural heritage: Design and assessment of an ethical framework. *Int. Arch. Of The Photogramm. Remote. Sens. And Spatial Inf. Sci.* **48**, 1149–1155 (2023).
17. Ai, H. High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI* **6**, 1005 (2019).
18. Pierdicca, R. et al. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote. Sens.* **12**, 1005 (2020).
19. Wang, Y. et al. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graph. (tog)* **38**, 1–12 (2019).
20. Matrone, F. et al. A benchmark for large-scale heritage point cloud semantic segmentation. *The Int. Arch. Photogramm. Remote. Sens. Spatial Inf. Sci.* **43**, 1419–1426 (2020).
21. Tiribelli, S., Giovanola, B., Pietrini, R., Frontoni, E. & Paolanti, M. Embedding ai ethics into the design and use of computer vision technology for consumer's behaviour understanding. *Comput. Vis. Image Underst.* **248**, 104142 (2024).
22. Giovanola, B. & Tiribelli, S. Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms. *AI & society* **38**, 549–563 (2023).
23. Migliorelli, L. et al. Accountable deep-learning-based vision systems for preterm infant monitoring. *Computer* **56**, 84–93 (2023).
24. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54**, 1–35 (2021).

25. Suresh, H. & Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9 (2021).
26. Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. big data* **2**, 13 (2019).
27. Cabitza, F. et al. Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai. *Expert. systems with Appl* **213**, 118888 (2023).
28. Ding, W., Abdel-Basset, M., Hawash, H. & Ali, A. M. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Inf. Sci.* **615**, 238–292 (2022).
29. Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
30. Pokholkova, M., Boch, A., Hohma, E. & Lütge, C. Measuring adherence to ai ethics: a methodology for assessing adherence to ethical principles in the use case of ai-enabled credit scoring application. *AI Ethics* 1–23 (2024).
31. Matrone, F., Paolanti, M., Felicetti, A., Martini, M. & Pierdicca, R. Bubblex: An explainable deep learning framework for point-cloud classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **15**, 6571–6587 (2022).
32. Matrone, F., Felicetti, A., Paolanti, M. & Pierdicca, R. Explaining ai: understanding deep learning models for heritage point clouds. *ISPRS Annals Photogramm. Remote. Sens. Spatial Inf. Sci*. 207–214 (2023).
33. Hasany, S. N., Petitjean, C. & Mériaudeau, F. Seg-xres-cam: Explaining spatially local regions in image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3733–3738 (2023).
34. Felicetti, A., Paolanti, M., Zingaretti, P., Pierdicca, R. & Malinverni, E. S. Mo. se.: Mosaic image segmentation based on deep cascading learning. *Virtual Archaeol. Rev.* **12**, 25–38 (2021).
35. Malinverni, E. S. et al. Significance deep learning based platform to fight illicit trafficking of cultural heritage goods. *Sci. Reports* **14**, 15081 (2024).
36. Paolanti, M. et al. Perganet: A deep learning framework for automatic appearance-based analysis of ancient parchment collections. In *International Conference on Image Analysis and Processing*, 290–301 (Springer, 2022).
37. Paolanti, M. et al. Deep convolutional neural networks for sentiment analysis of cultural heritage. *The Int. Arch. Photogramm. Remote. Sens. Spatial Inf. Sci.* **42**, 871–878 (2019).

Acknowledgements

This work was supported by the project titled ”Open, Reusable, and Multimodal Large Language Model Technology for Point Cloud Understanding” (acronym ”3D.LLM”), funded under the cascade call of Spoke 7 Politecnico di Torino, as part of the FAIR research programme financed through the PNRR Mission 4 – Component 2 – Investment 1.3 Extended Partnerships.

Author contributions

Conceptualization, M.P., E.F., and R.P.; methodology, M.P., E.F., and R.P.; validation, M.P., E.F., and R.P.; formal analysis, M.P., E.F., and R.P.; investigation, M.P., E.F., and R.P.; writing—original draft preparation, M.P., E.F., and R.P.; writing—review and editing, M.P., E.F., and R.P.; visualization, M.P., E.F., and R.P.; supervision, M.P., E.F., and R.P. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Marina Paolanti.