This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Machine Learning and Cultural Heritage: An Italian Perspective

Giovanna Maria Dimitri

***Abstract*—Cultural heritage plays a crucial role for societal identity, allowing a deep connection between individuals and their historical roots. In this context the advent of deep learning technologies has significantly influenced the preservation and analysis of cultural heritage, offering innovative and technological solutions for heritage documentation, classification, and restoration. This article presents a short review of existing deep learning methodologies applied to cultural heritage preservation in Italy. By leveraging a systematic literature search, we investigate technological solutions deployed across various Italian regions and analyze their impact on the dissemination of artificial intelligence in this field. Furthermore, we perform a geolocation-based assessment of these approaches to uncover potential correlations between technological novel solutions and regional characteristics. Our findings indicate the critical role of deep learning in modernizing heritage conservation efforts, moreover, highlighting the necessity for continued research and interdisciplinary collaboration in this domain.**

***Index Terms*—Cultural heritage, deep learning, Italy.**

I. Introduction

Cultural heritage is more than a collection of traditions, languages, and artifacts; it is a foundation upon which individuals shape their identity, values, and sense of belonging. From the stories passed down through generations to the customs embedded in daily life, heritage provides a guiding thread that connects individuals to their past, while influencing their personal growth and world-view [1].

By embracing cultural heritage, individuals gain a deeper understanding of *who* they are, increasing their own self-awareness, identities and confidence. It shapes moral frameworks, ethical perspectives, and social behaviors, guiding decision-making and interactions with others [1], [2].

Furthermore, heritage serves as a source of resilience, offering a sense of continuity and stability in an ever-changing world. Through the preservation and appreciation of cultural roots, individuals not only enrich their own lives, but also contribute to a more diverse and inclusive [3] society.

In this context [4] we can place the use of deep learning methods and their role for the conservation of cultural heritage and societies.

In this article we reviewed existing solutions in the context of the Italian cultural heritage. Italy's cultural heritage is, in fact, an unparalleled treasure of history, art, and traditions, forged by millennia of human creativity and innovation. With the UNESCO[1] recognizing more World Heritage sites in Italy than in any other country in the world, Italy's cultural legacy remains a source of inspiration, an invitation for the world to experience its timeless beauty and profound historical significance [5].

But how much has artificial intelligence (AI) pervaded the landscape of Italian cultural heritage preservation? Which evidence do we have from the literature?

In order to answer these research questions, we performed a literature review search over technological solutions proposed for the Italian scenario.

We further performed a geolocalization of such solutions, investigating possible relationships between the solutions proposed and their feasibility.

The article is structured as follows. In Section II we describe the research methodology used to find the source articles for the review.

In Section III we report the statistical findings of the search conducted. In Section IV we offer an overview of the works included in the review. In Section V we further present an overview of the musical heritage works and, eventually, Section VI includes discussions and conclusions.

II. Research Methodology

To perform our systematic review we relied on two main databases: Scopus and Web of Science (WOS). WOS are two of the most widely used scientific journal databases for academic research. Both platforms serve as essential tools for researchers, institutions, and policymakers, providing access to a vast collection of peer-reviewed literature across multiple disciplines. Scopus,[2] managed by Elsevier, is known for its broad coverage of scientific journals, conference proceedings, and patents, including research fields such as life sciences, social sciences, physical sciences, and health sciences. It offers powerful analytical tools that help researchers track citations, measure research impact, and explore potential collaborations.

Received 28 February 2025; revised 31 August 2025 and 22 December 2025; accepted 22 April 2026. *(Corresponding author: Giovanna Maria Dimitri.)*

The author is with the Universitá degli Studi di Milano Statale, 20122 Milano, Italy (e-mail: giovanna.dimitri@unimi.it).

Digital Object Identifier 10.1109/TCSS.2026.3689422

[1]https://www.unesco.org/en
[2]https://www.elsevier.com/products/scopus

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Web of Science,[3] owned by Clarivate,[4] is a highly curated database with a strong emphasis on quality and selectivity. It includes a range of citation indexes, such as the science citation index (SCI), social sciences citation index (SSCI), and arts & humanities citation index (A&HCI). Web of Science is often considered more stringent in terms of journals inclusion, which can make it particularly valuable for identifying high-impact research. Its citation tracking and analysis tools provide insights into research trends, author influence, and institutional quality rankings.

Although the two databases share similarities, they differ in terms of journals coverage, indexing criteria, and analytical features. Scopus tends to have a broader scope, indexing more journals, while Web of Science is often regarded as more selective.

In our systematic literature review, to select the relevant articles, we performed the following search.

Query: cultural AND heritage AND deep learning AND Italy

As a preprocessing step we selected, among the results, those articles whose title, abstract and keywords contained the relevant queried keywords.

III. Investigating the Use of Deep Learning Methods in Cultural Heritage in the Italian Scenario

In the following section we will report statistics concerning the queries results, together with the data preprocessing steps.

A. Literature Search

In the list below we report the number of articles found in Scopus, in WOS and the *common* articles found in both databases.

1) Number of articles in Scopus: 17.
2) Number of articles in Web of Science: 7.
3) Number of articles common to both databases: 6.

As we can see the number of articles retrieved in Scopus was higher than the one retrieved through WOS, with 6 of them being in common between the search conducted in the two databases. A summary of the title distribution and most frequent title words is reported in Figs. 1 and 2.

IV. Literature Background

In this section we will describe the articles which were selected from an initial preprocessing made upon the results of the WOS and Scopus literature search. More specifically, from the initial search results we removed articles whose topic was unrelated to deep learning applications to cultural heritage, and we also removed duplicates. The results allowed to have an overview of studies covering geographically a good part of the Italian country. The cities mentioned in the studies reported in our short review are pictured in the maps shown in Fig. 3. Moreover in Table II we report a thorough comparison of the several methods used. More specifically we tried to explain which technologies were considered more suitable trying to clarify the rationale behind their choices.

A. Screening the Stones of Venice: Mapping Social Perceptions of Cultural Significance Through Graph-Based Semisupervised Classification

In this work [6] the authors used a novel approach, based on graph neural networks (GNN) models, for mapping the perception of cultural significance in the city of Venice. The public perception has in fact become an extremely important aspect, as shown by the 2011 UNESCO Recommendation on the historic urban landscape (HUL).[5] More specifically, in the article, the authors used a semisupervised GNN approach to classify, summarize and map cultural significance, using as input content retrieved from social media. The input data used in this study, in fact, is a so called attributed multigraph. More specifically, social media posts were structured as a graph where nodes represent entities as for instance posts, users, or locations. Edges connect related entities, and in this way they can capture relationships, such as social connections between users or spatial proximity between posts and locations. This is the reason why the use of GNN architectures provide an important addition to this context. The use of graphs, allow in fact, to enhance the structure of social common behavior and proximities. In this way the information was been embedded in a spatio temporal fashion. Moreover, the semisupervised characteristic of the machine learning model implies that the GNN learns from a combination of a small set of labelled data (posts with known significance) and a larger set of unlabeled data. This is also an important aspect, which can help and be useful in several applications due to the frequent lack of labelled data. Another important characteristic of the model is the ensembling approach. Indeed, the overall architecture is defined as an ensemble of GNN models which are trained and combined to produce a final decision. This ensemble approach can improve robustness and incorporate different aspects of the data, leading to better overall classification performance. Additionally, message diffusion techniques on graphs were introduced to aggregate post labels onto nearby spatial nodes, aiding the mapping of cultural significance categories within their geographical contexts and performing spatial mapping. Once classified, in fact, a message diffusion method is applied to the graph. Labels are "diffused" or aggregated from the post nodes to their adjacent spatial nodes (e.g., locations). This process helps to map the perceived cultural significance categories to their geographical locations, creating visual maps. The proposed methodology is tested using data from Venice as a case study, illustrating the creation of social perception maps for this UNESCO World Heritage site. Such research framework and methodology can be successfully applied to other cities globally, enhancing socially inclusive heritage management practices. Furthermore, the proposed approach has the potential to diffuse any human-generated, location-based data

[3]https://www.webofscience.com/wos/
[4]https://clarivate.com/
[5]https://whc.unesco.org/en/hul/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

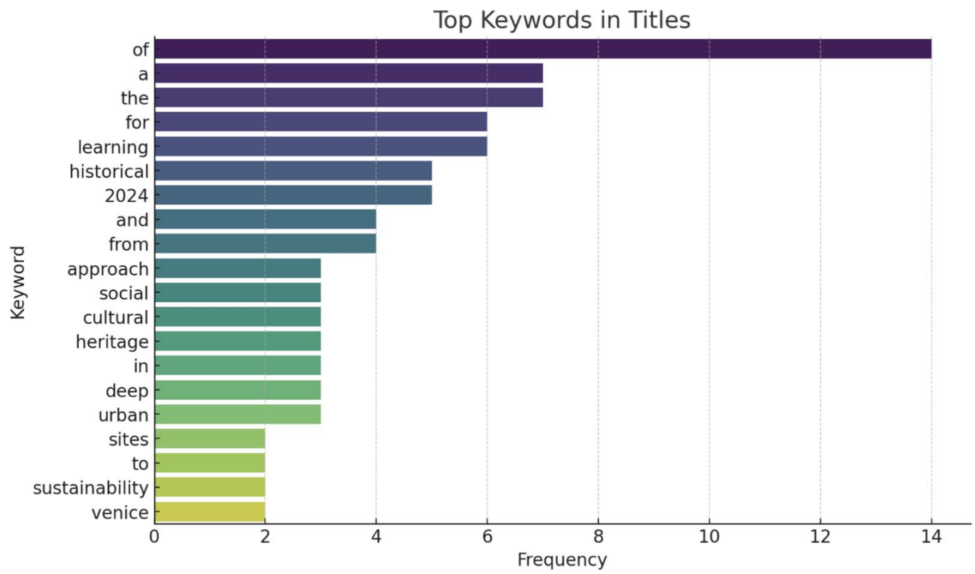

Fig. 1. Most frequent words in the titles of the articles selected from our query in both Scopus and WOS.

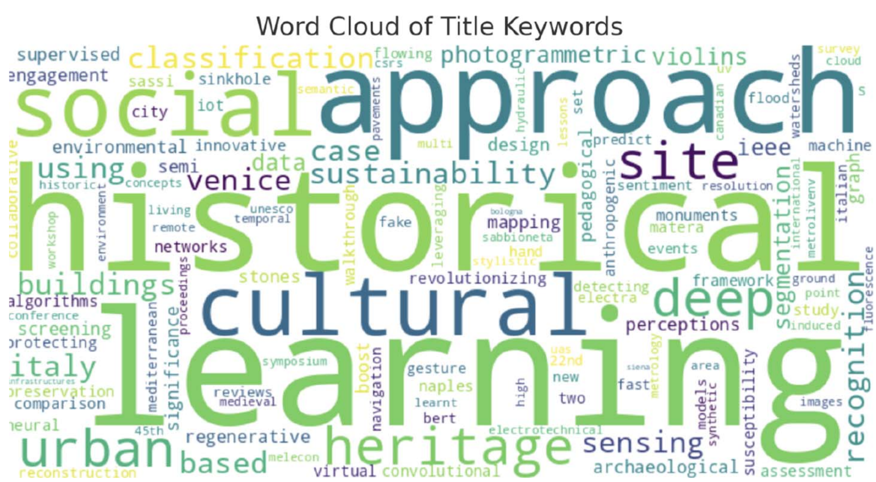

Fig. 2. Cloud of words included in the titles for of the articles selected in our review where the stop-words have been removed.

onto spatial networks and temporal timelines, offering insights into the safety, vitality, and general features of several urban spaces.

B. Revolutionizing Cultural Heritage Preservation: An Innovative IoT-Based Framework for Protecting Historical Buildings

In this work [7] the authors presented the possible use of machine learning and deep learning methodologies for the preservation of the Italian cultural heritage. Deterioration can in fact highly affect the heritage. This study presents an IoT-based system that integrates monitoring, predictive maintenance, and decision-making to support the protection of cultural heritage buildings. More specifically the IoT-based framework is technically implemented through a three-tier edge–cloud architecture and applied in a case study at the Royal Palace of Carditello (Caserta). Smart objects, equipped with micro-sensors, continuously collect environmental and structural data, with time-critical tasks (e.g., threshold detection of vibrations or anomalies) handled directly at the device or edge level to ensure rapid response. An edge server aggregates, normalizes, and fuses sensor streams, producing anomaly timelines and initial alerts, while the cloud tier manages long-term storage, complex analytics, and decision-support functionalities. The system design follows a situational awareness (SA) model enhanced by goal-directed task analysis (GDTA), which identifies operator goals and aligns them with system functionalities. Decision-making is distributed across tiers. More specifically devices manage immediate reactions, the edge coordinates local fusion and short-term reasoning, and the cloud provides broader contextual analysis. More technically the architecture is organized as follows. First of all a sensor layer is made of all of the IoT devices which collect all data coming from nature.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

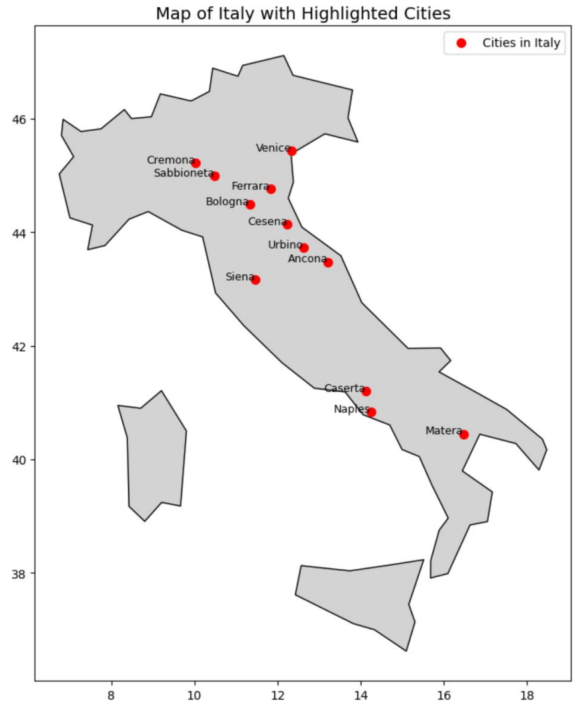

Fig. 3. Cities which were localized through our literature review.

Subsequently the information collected in the sensor layer are then passed to the Knowledge Base Layer, which acquires data from the sensor layer and storage them for later used. After this the architecture is composed by the Inference Engine Layer, which is composed by two sub-modules: the so called Predictive Maintenance Module and the Decision Support Module. The Predictive Maintenance Module is based on autoencoders (AEs) for anomaly detection, with a dedicated AE trained for each monitored damage type. These models learn to approximate the identity function on normal-condition data; anomalies are detected by evaluating reconstruction errors against predefined thresholds derived during training. When an anomalous behavior is detected, a k-nearest neighbors (kNN) algorithm is applied to classify the damage and associate it with the most appropriate intervention. The Decision Support Module integrates the outputs of the AEs and kNN classifiers. Detected anomalies and their classifications are mapped to corresponding maintenance or restoration actions, in accordance with relevant regulatory guidelines. The overall system operates in two phases: an offline phase, where AE and kNN models are trained using historical monitoring data, and an online phase. In the latter the trained models perform real-time inference to detect, classify, and support decision-making regarding structural damages.

C. Comparison of Two Machine Learning Algorithms for Anthropogenic Sinkhole Susceptibility Assessment in the City of Naples (Italy)

In [8] the authors address an emerging issue of sinkhole susceptible areas, as the one around Naples. More specifically the authors highlight the common occurrence of sinkholes, which result from collapse processes due to both natural and human activities. Sinkholes are typically funnel-shaped depressions that can range from a few centimeters to several meters deep. In urban areas, particularly in Naples, sinkholes—often refer to as “anthropogenic” sinkholes, pose significant engineering challenges, threatening the stability of buildings, infrastructure, and potentially causing fatalities. The presence of a dense underground cavity network, formed by historical quarrying of volcanic tuff, increases sinkhole risk in Naples. In their study, the authors conducted a sinkhole susceptibility analysis using two well-known Machine Learning algorithms: Random Forest and Maximum Entropy MaxEnt, to generate and compare sinkhole susceptibility maps. Random Forest is a tree-based ensemble method that builds multiple decision trees and merges their outputs to improve prediction accuracy, while MaxEnt is a presence-only modelling approach that estimates a probability distribution of sinkhole occurrence by finding the maximum entropy (most uniform) solution conditioned on observed constraints. Twelve environmental variables were considered, including groundwater depth, bedrock depth, and proximity to factors like aqueducts, roads, sewers, anthropic cavities, and underground railroad networks. Results show that the Sinkhole Susceptibility Maps obtained with Maximum Entropy and Random Forest algorithms show good performances (AUC score = 0.93). The map resulting from the application of the Random Forest algorithm seems to be more inclined towards what is defined from the authors as a precautionary approach, than the one obtained with Maximum Entropy. This is due to the more neat distinction in the spatial distribution of the susceptibility classes. The good predictive performance, demonstrated by both algorithms, identify a high susceptibility to sinkholes in Naples’ city center, which correlates with the dense network of underground cavities. The road network, which also represents secondary aqueduct and sewer systems, was identified as a significant contributing factor. The limitations of these algorithms, highlighted by the authors, are the need for an inventory of the phenomena in the study area as well as the availability of predisposing factors data. Nonetheless this work can definitely be considered as a valuable tool for enhancing sinkhole susceptibility mapping and aiding local governments in protecting cultural heritage.

D. Social Sensing for Monuments Recognition Using Convolutional Neural Networks: A Case Study

In [9] the authors discuss how social media and AI are transforming various sectors, including cultural heritage. More specifically they highlighted the potential role of AI in promoting and sharing tourism activities, and points of interest in urban areas. Specifically, they explore the application of these techniques in automatically recognizing and classifying monuments and points of interest. Their study focuses on monument recognition using images captured by users’ mobile devices while visiting urban environments. The research uses images from citizens and tourists in the city of Cesena with the goal of enhancing the visibility of monuments and places. The authors evaluated different types of convolutional neural network (CNN) architectures, achieving an accuracy of over

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

90% in distinguishing among outdoor monuments. The comparison was done with the aim of trying to balance accuracy, robustness, and computational loads of the various experiments performed. The authors used well known CNN based architectures as ResNet-50, VGG16 and VGG-19. They also compared these larger networks to a lighter CNN, named Custom CNN, whose architecture is described in the article [9]. In brief the proposed Custom CNN is a lightweight, sequential architecture designed to balance classification accuracy and computational efficiency for monument recognition. The network takes an RGB image as input and processes it through a hierarchy of convolutional and pooling layers to progressively extract spatial and semantic features. The feature extraction stage consists of three convolutional blocks. Each block applies a 3 × 3 convolution with an increasing number of filters (16, 32, and 64 respectively), enabling the network to capture low-level visual features such as edges and textures in the early layers and more abstract spatial patterns in the deeper layers. Each convolutional operation is followed by a ReLU activation function to introduce nonlinearity and improve learning capability. After each convolution, a 2 × 2 max-pooling layer reduces the spatial resolution of the feature maps, limiting overfitting and decreasing computational cost while retaining the most salient features. After the convolutional stage, the extracted feature maps are flattened into a 1-D vector, which serves as input to the classification stage. This stage is composed of a sequence of fully connected (dense) layers with 256, 128, and 64 neurons, respectively, all using ReLU activation functions. These layers progressively refine the learned representation and enable high-level feature interactions relevant for monument discrimination. To further reduce overfitting, dropout layers with a rate of 0.5 are inserted between dense layers, randomly deactivating neurons during training. The network terminates with a dense output layer of 3 neurons, corresponding to the three monument classes considered in the case study. A soft-max activation function is applied to produce normalized class probabilities, enabling multiclass classification. The architecture follows a classic CNN design pattern—convolutional feature extraction followed by dense classification—while remaining intentionally shallow and parameter-efficient, making it suitable for scenarios with limited training data and potential deployment on resource-constrained platforms.

E. A Collaborative Virtual Walkthrough of Matera's Sassi Using Photogrammetric Reconstruction and Hand Gesture Navigation

In [10] the authors highlight how the COVID-19 pandemic has emphasized the need for real-time, collaborative virtual tools to facilitate remote activities, particularly in fields like education and cultural heritage. Virtual walk-through, which allow users to explore and interact with historical sites, are presented as an effective means of engagement. However, creating realistic and user-friendly applications remains a significant challenge. This study explores the use of collaborative virtual walk-throughs as an educational tool for cultural heritage, specifically focusing on the Sassi of Matera, a UNESCO World Heritage Site in Italy. The virtual walk-through application was developed using RealityCapture and Unreal Engine, incorporating photogrammetric reconstruction and deep learning-based hand gesture recognition to create an immersive and accessible experience. The application allowed users to interact with the environment using intuitive gestures. A dataset of nearly 2000 photographs was processed with RealityCapture to produce a dense 3-D reconstruction. The resulting mesh was simplified to reduce computational load while retaining visual realism via high-resolution textures. The optimized model was then imported into Unreal Engine, where ambient illumination and interactive info-stands were integrated to create a realistic, educational, virtual environments accessible to multiple users. To enable natural navigation, the authors designed a hand gesture recognition system using a monocular RGB camera with the MediaPipe framework. In this case the machine learning algorithm used consisted in a lightweight feed-forward neural network which was trained on hand landmark coordinates to classify 13 gestures (six static and seven dynamic). These gestures were mapped to actions such as tobject interaction, and interface control. By combining accurate photogrammetric modelling with real-time gesture recognition, the system allowed users to collaboratively explore the reconstructed heritage site in an intuitive and immersive way. A test with 36 participants provided positive feedbacks on the application's effectiveness, ease of use, and user-friendliness. The findings indicate that virtual walk-throughs can offer accurate representations of complex historical locations and promote both tangible and intangible heritage. The authors suggest future work should expand the reconstructed site, improve performance, and assess the impact on learning outcomes. Overall, the study underscores the potential of virtual walk-through applications as valuable tools for architecture, cultural heritage, and environmental education. More specifically the model works as follows. As mentioned above, the system combines photogrammetric reconstruction with hand-gesture–based interaction, where neural networks are used exclusively for gesture recognition and not for 3-D reconstruction. The physical environment is first captured through a large set of RGB images and reconstructed into a detailed 3-D model using standard photogrammetry software. This stage relies on geometric and computer vision techniques rather than learning-based models, and the resulting mesh is optimized and imported into Unreal Engine to create an interactive virtual scene. User interaction is enabled through a hand gesture recognition pipeline driven by neural networks: video frames from a monocular RGB camera are processed by MediaPipe Hands, which employs deep neural models to detect the hand and estimate the 2-D coordinates of 21 anatomical landmarks. These landmark coordinates are then fed into a lightweight feed-forward neural network with a single hidden layer, trained to classify static and dynamic hand gestures. The output gesture classes are mapped to navigation and interaction commands within the virtual environment, allowing users to move and interact naturally without specialized hardware. In summary, neural networks are used specifically for hand detection, landmark estimation, and gesture classification, while the virtual environment reconstruction and rendering

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

are handled by nonneural photogrammetric and graphics pipelines.

F. A New Italian Cultural Heritage Data Set: Detecting Fake Reviews With BERT and ELECTRA Leveraging the Sentiment

In [4] the authors discuss the growing phenomenon of online reviews, in the context of cultural heritage, which has become an important source of tourism and economic growth. However, this growth has also been accompanied by the rise of fake reviews. In response, the scientific community has developed language models capable of distinguishing fake reviews from authentic ones. These models, often based on deep neural networks with transformer-type architectures, face limitations due to the lack of specific local language datasets for certain domains, which are crucial for both training and verification. This article has two main objectives. First, the authors created a new dataset in Italian, a low-resource language, focused on the domain of cultural heritage in Italy. They collected online reviews and reorganized them into a format suitable for use by language models. Second, they established a baseline for detecting misleading reviews by applying two widely-used and well known language models, BERT and ELECTRA. Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language model trained with a masked language modelling objective, in which some input tokens are hidden and the model learns to predict them using both left and right context, enabling deep bidirectional representations. It is pretrained on large text corpora and then fine-tuned for downstream tasks such as classification, question answering, or sequence labeling. ELECTRA is also a transformer-based model but replaces masked language modeling with a more sample-efficient pretraining scheme: a small generator model corrupts the input by replacing tokens, and a discriminator model is trained to detect which tokens have been replaced. The models were trained and tested using standard splits (80/20 and 90/10), achieving approximately 95% accuracy and F1 score in distinguishing between genuine and deceptive reviews. Importantly, the study also employs SHapley Additive exPlanations (SHAP) to interpret model decisions, demonstrating that sentiment-related features notably help in detecting deceptiveness. This integration of explainability underscores the role of sentiment analysis in the automated identification of fake reviews in the dataset [4].

G. Remote Sensing Analysis

We include here also the proceedings of two remote sensing conferences in which machine learning techniques for the preservation of the Italian cultural heritage were taken into account.

1) 45th Canadian Symposium on Remote Sensing (CSRS 2024):[6] The CSRS 2024 proceedings feature 10 articles covering a variety of topics related to remote sensing applications. Key subjects related to the cultural heritage application included comparison of classifiers for urban change detection in Kingston, Ontario, and the fusion of UAS and TLS 3-D data for cultural heritage assessments, particularly at St. Catherine Monastery in Ferrara, Italy. Other topics part of the proceedings and related to cultural heritage preservation, includes, for instance, studies related to the automation of methods for evaluating pavement surface distress.

2) 2024 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv 2024):[7] The MetroLivEnv 2024 proceedings include 104 articles, addressing various aspects of digital technology and its role in enhancing living environments. Topics include structural investigations on the San Domenico complex in Cosenza, the effects of lockdown on electricity demand patterns in institutional buildings, and innovative applications of deep learning for cultural heritage preservation. Additional discussions cover the use of digital twins for structural health monitoring, managing large-scale data from terrestrial laser scanners, and enhancing smart city management with geomatic techniques and CIM. The proceedings also explore strategies for reducing urban heat islands, mapping carbon dynamics in Calabria (Italy), and using data-driven digital services to build environmental innovation.

H. Learning From Synthetic Point Cloud Data for Historical Buildings Semantic Segmentation

Historical heritage [11] requires efficient pipelines for creating fully interoperable and informative heritage building information modeling (HBIM). Developing effective Scan-to-BIM workflows is crucial for managing historical real estate, as generating 3-D models from point clouds is complex and time-consuming. In this context, semantic segmentation of 3-D point clouds, particularly with the use of Deep Learning approaches, is becoming increasingly important, facing challenges in historical architecture due to the complexity of shapes and the lack of large annotated datasets. To address this, in [11] the authors explore the use of synthetic point cloud data to train Deep Learning models for semantic segmentation of historical buildings, introducing an improved version of the Dynamic Graph CNN (RadDGCNN) and testing it on datasets from the Ducal Palace in Urbino and Palazzo Ferretti in Ancona, Italy, showing promising results in segmenting real (Terrestrial Laser Scanning) TLS datasets. RadDGCNN is an enhanced variant of the original dynamic graph convolutional neural network (DGCNN), adapted for semantic segmentation of historical building point clouds. While standard DGCNN dynamically constructs local neighborhood graphs using kNN, RadDGCNN replaces this with a radius-based neighborhood definition—points within a specified radial distance—to better capture the local geometry of complex architectural features. More specifically the Dynamic Graph CNN is designed for point-wise semantic segmentation of 3-D point clouds. The network takes as input a block of N points, each represented by its spatial coordinates (augmented in the experiments with normalized local coordinates), and builds a graph by associating each point with K neighboring points. Local feature learning

[6] https://crss-sct.ca/publications/cjrs/

[7] https://www.metrolivenv.org/metrolivenv2024/

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

is performed through a sequence of three EdgeConv layers. In each EdgeConv, edge features are computed by combining the feature vector of a point with the relative differences between that point and its neighbors; these edge features are processed by shared multilayer perceptrons and aggregated with a max-pooling operation over the neighborhood, ensuring permutation invariance. By recomputing the neighborhood graph at each layer, the model progressively captures higher-level local geometric structures. The outputs of the three EdgeConv layers are concatenated to preserve multiscale local information and then passed through a further multilayer perceptron (MLP) to learn a global representation of the entire point cloud block. This global feature is subsequently concatenated back with the per-point local features, allowing each point's prediction to be informed by both its immediate geometric context and the overall scene structure. A final shared MLP maps these fused features to P class scores, producing a semantic label for each point in the cloud.

I. A Deep Learning Approach for the Recognition of Urban Ground Pavements in Historical Sites

Urban management [12], particularly in relation to smart cities, is a key area of focus for local administrators due to its significant impact on enhancing sustainability. One aspect of urban management, especially in historical city centers and sites, is the physical accessibility of urban areas, where the variety of pavement types can pose challenges. To improve the management of these areas, comprehensive mapping of different pavements is essential. This article presents a case study of Sabbioneta, a historical city in northern Italy, where a mobile mapping system (MMS) was used for data collection. The study employs DL, more specifically well-known CNN based architectures, ResNet and U-Net to classify different pavement types. ResNet (Residual Network) is a deep CNN architecture designed to address the degradation problem that arises when networks become very deep. Its core idea is the use of residual connections, or skip connections, which allow layers to learn residual functions with respect to their inputs rather than direct mappings. By adding the input of a block to its output, ResNet facilitates gradient flow during backpropagation, enabling the effective training of networks with tens or even hundreds of layers. This design has made ResNet a foundational architecture for image classification and a common backbone for many computer vision tasks. U-Net is a CNN architecture specifically developed for image segmentation, particularly in biomedical imaging. It follows an encoder–decoder structure, where the encoder progressively extracts high-level features through downsampling, and the decoder reconstructs spatial resolution through upsampling. Crucially, U-Net introduces skip connections between corresponding encoder and decoder layers, allowing fine-grained spatial information to be combined with semantic features. This structure enables precise, pixel-level predictions even with limited training data, making U-Net widely adopted for semantic segmentation tasks. [13]. The process described in the article included selecting ground surface points from the point cloud, rasterizing the data, and then applying a material classification using a U-Net architecture coupled with ResNet–18. The study successfully identified five pavement classes: sampietrini, bricks, cobblestone, stone, and asphalt, with an average accuracy of 94%.

TABLE I
SUMMARY OF DEEP LEARNING MODELS USED IN CULTURAL HERITAGE APPLICATIONS

Paper	Deep Learning Model
[16]	GNN
[7]	CNN
[9]	CNN
[10]	Deep Learning for Gesture Recognition
[4]	BERT, ELECTRA
[11]	Dynamic Graph CNN (RadDGCNN)
[12]	U-Net with ResNet18
[15]	VGG16, ResNet50, InceptionV3

V. MUSICAL HERITAGE

If it is true that mostly we would think of architectural operas, when thinking about preservation, it is also well known how also musical and artistic objects have to be inevitably preserved. In the following we will describe two interesting studies in which preservation of musical instruments in Italy are proposed, using machine learning technologies, and which were returned by our literature search.

A. Segmentation of Multitemporal UV-Induced Fluorescence Images of Historical Violins [14]

Monitoring the conservation state of a historical violin is challenging due to the complex, layered surface created by multiple restorations over time. The reflectance of varnishes and the violin's rounded shape can produce noise, which may be mistaken for genuine alterations. To address this, the authors proposed an improvement of their previous segmentation method based on hue saturation and value (HSV) histogram quantization using a genetic algorithm, and tested it on images of two important violins from the "Museo del Violino" in Cremona, Italy, taken periodically over six months, as well as images of a sample violin altered in a lab to simulate long-term changes.

B. Stylistic Classification of Historical Violins: A Deep Learning Approach [15]

Stylistic analysis of artworks is a common challenge in the Cultural Heritage field, with experts typically using standard methods to analyze statues and paintings. However, historical violins present a unique challenge, as even though the key stylistic features are known, only a few experts can confidently attribute a violin to its maker. In this article the authors explore solutions using deep learning for violin style discrimination, testing three CNN models (VGG16, ResNet50, InceptionV3) on a binary classification task (Stradivari versus NotStradivari), achieving 77.27% accuracy and a 0.72 F1 score, and comparing the model's identified regions of interest with those identified in a previous study of expert luthiers.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
COMPARISON OF AI METHODS USED IN THE REVIEWED STUDIES: METHODOLOGICAL RATIONALE AND STRENGTHS

Study	AI Method	Why this Method was Chosen	Main Strengths
Bai et al. [6]	Semi-supervised GNN ensemble	Graph Neural Networks were selected over CNNs or standard classifiers because the data are inherently relational (users, posts, locations). GNNs explicitly model spatial, social, and semantic dependencies, while semi-supervised learning addresses the scarcity of labelled data.	Captures complex spatial–social relationships; effective with limited labels; supports message diffusion for spatial mapping; high interpretability of collective social perception patterns.
Casillo et al. [7]	ML/DL in IoT edge–cloud architecture	Learning-based models were preferred to rule-based systems to enable adaptive anomaly detection and predictive maintenance under heterogeneous sensor streams. Edge intelligence reduces latency compared to cloud-only solutions.	Scalable real-time monitoring; resilience to noise; supports predictive maintenance; suitable for safety-critical heritage protection scenarios.
Bausilio et al. [8]	Random Forest and MaxEnt	RF and MaxEnt were chosen over deep neural networks due to limited labeled sinkhole events and the need for interpretable susceptibility maps. MaxEnt handles presence-only data, while RF captures nonlinear interactions.	High predictive performance (AUC = 0.93); robustness to small datasets; interpretability for decision-makers; suitability for precautionary risk assessment.
Delnevo et al. [9]	CNNs (VGG16, VGG19, ResNet50, Custom CNN)	CNNs outperform traditional feature-based methods for visual recognition tasks. Comparing heavy and lightweight architectures allows balancing accuracy and computational cost for mobile and social sensing scenarios.	High recognition accuracy (>90%); robustness to visual variability; adaptability to resource-constrained environments; transferable to other urban contexts.
Notarangelo et al. [10]	Lightweight NN for gesture recognition	A shallow neural network was preferred over complex temporal models (e.g., LSTM) to ensure real-time performance with monocular RGB input and low computational overhead in VR applications.	Low latency; real-time interaction; intuitive human–computer interaction; effective integration with immersive environments.
Catelli et al. [4]	BERT and ELECTRA	Transformer-based language models were selected over classical ML due to their superior contextual representation of language. ELECTRA provides competitive performance with reduced training cost, suitable for low-resource Italian datasets.	State-of-the-art accuracy (∼95%); strong contextual understanding; explainability via SHAP; effective for domain-specific NLP tasks.
Morbidoni et al. [11]	RadDGCNN	Graph-based point cloud networks were preferred over voxel- or image-based methods to preserve geometric fidelity. Radius-based neighborhoods improve robustness to irregular historical geometries.	Accurate semantic segmentation; adaptability to complex architectural shapes; reduced dependency on real annotated data via synthetic training.
Treccani et al. [12]	U-Net + ResNet18	U-Net was chosen for pixel-wise segmentation, while ResNet18 provides a good trade-off between depth and efficiency compared to deeper backbones. Suitable for rasterized point cloud data.	High segmentation accuracy (94%); efficient training; strong generalization; effective for urban-scale mapping tasks.
Dondi et al. [14]	Genetic algorithm + HSV segmentation	Evolutionary optimization was preferred over fixed-threshold methods to adapt segmentation to noise, varnish variability, and illumination changes in UV imagery.	Robust to noise; adaptive to temporal changes; suitable for conservation monitoring; low data requirements.
Dondi et al. [15]	CNNs (VGG16, ResNet50, InceptionV3)	Deep CNNs were chosen over handcrafted stylistic features to automatically learn discriminative visual cues aligned with expert knowledge. Architecture comparison highlights robustness vs. complexity.	Captures subtle stylistic traits; supports expert validation; transferable to other attribution tasks; bridges human and AI analysis.

VI. DISCUSSIONS AND CONCLUSION

The analysis of machine learning applications in Italian cultural heritage preservation revealed the presence of interesting possible studies and applications in the field. The systematic literature review we conducted highlights a growing trend toward integrating AI with heritage conservation efforts, showcasing the transformative potential of these technologies. In Table I we summarize the methodologies which we obtained through the literature search.

From Table I we can see that deep learning models have been successfully applied to tasks such as object recognition, damage classification, and semantic segmentation, significantly enhancing the efficiency and accuracy of heritage preservation.

Moreover we further summarize in Table II a comparison between the models used, and also some comments on which other possible models could have been used and added to the various studies.

One of the key findings is the correlation between the geographic distribution of deep learning applications and the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

concentration of cultural heritage sites in Italy. The prevalence of AI-driven projects in historically rich regions suggests that institutions and researchers are prioritizing the digital transformation of heritage conservation in these areas. However, disparities remain in the deployment of such technologies across different regions, indicating the need for broader access to resources and funding, as also our Italian map, shown in Fig. 3.

Future research should focus on enhancing the robustness of deep learning models by incorporating more diverse datasets and developing hybrid approaches that integrate traditional conservation methods with AI-driven techniques. Furthermore, policy frameworks should be established to promote the responsible use of AI in cultural heritage, ensuring that technological progress supports the long-term sustainability of historical assets. By facilitating continued innovation and collaboration, deep learning can play a pivotal role in preserving cultural heritage for the future.

REFERENCES

[1] C. Dümcke and M. Gnedovsky, "The social and economic value of cultural heritage: Literature review," *EENC Paper*, vol. 1, pp. 101–114, Jul. 2013.

[2] M. Mishra and P. B. Lourenço, "Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review," *J. Cultural Heritage*, vol. 66, pp. 536–550, Mar.-Apr. 2024.

[3] A. Klamer, "The values of cultural heritage," in *Handbook on the Economics of Cultural Heritage.* U.K.: Edward Elgar Publishing, 2013, pp. 421–437.

[4] R. Catelli et al., "A new Italian cultural heritage data set: Detecting fake reviews with Bert and Electra leveraging the sentiment," *IEEE Access*, vol. 11, pp. 52214–52225, 2023.

[5] E. Bertacchini, F. Revelli, and R. Zotti, "The economic impact of UNESCO world heritage: Evidence from Italy," *Regional Sci. Urban Econ.*, vol. 105, 2024, Art. no. 103996.

[6] N. Bai, P. Nourian, R. Luo, T. Cheng, and A. P. Roders, "Screening the stones of Venice: Mapping social perceptions of cultural significance through graph-based semi-supervised classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 203, pp. 135–164, Sep. 2023.

[7] M. Casillo, F. Colace, R. Gaeta, A. Lorusso, D. Santaniello, and C. Valentino, "Revolutionizing cultural heritage preservation: An innovative IoT-based framework for protecting historical buildings," *Evol. Intell.*, vol. 17, no. 5, pp. 3815–3831, 2024.

[8] G. Bausilio et al., "Comparison of two machine learning algorithms for anthropogenic sinkhole susceptibility assessment in the city of Naples (Italy)," in *Geotechnical Engineering for the Preservation of Monuments and Historic Sites III.* Boca Raton, FL, USA: CRC Press, 2022, pp. 1112–1123.

[9] G. Delnevo, A. Kazazi, E. Amadori, and S. Mirri, "Social sensing for monuments recognition using convolutional neural networks: A case study," in *Proc. IEEE Symp. Comput. Commun. (ISCC).* Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–6.

[10] N. M. Notarangelo, G. Manfredi, and G. Gilio, "A collaborative virtual walkthrough of Matera's Sassi using photogrammetric reconstruction and hand gesture navigation," *J. Imag.*, vol. 9, no. 4, p. 88, 2023.

[11] C. Morbidoni, R. Pierdicca, M. Paolanti, R. Quattrini, and R. Mammoli, "Learning from synthetic point cloud data for historical buildings semantic segmentation," *J. Comput. Cultural Heritage (JOCCH)*, vol. 13, no. 4, pp. 1–16, 2020.

[12] D. Treccani, J. Balado, A. Fernández, A. Adami, L. Diaz-Vilarino et al., "A deep learning approach for the recognition of urban ground pavements in historical sites," *Int. Arch. Photo. Remote Sens. Spatial Inf. Sci.*, vol. 43, no. B4, pp. 321–326, 2022.

[13] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, Jul. 2022.

[14] P. Dondi, L. Lombardi, M. Malagodi, and M. Licchelli, "Segmentation of multi-temporal UV-induced fluorescence images of historical violins," in *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy.* Berlin, Germany: Springer, 2019, pp. 81–91.

[15] P. Dondi, L. Lombardi, M. Malagodi, and M. Licchelli, "Stylistic classification of historical violins: a deep learning approach," in *Proc. Int. Conf. Pattern Recognit.* Berlin, Germany: Springer, 2021, pp. 112–125.

[16] C. López-Bravo and J. P. López, "Learning from medieval Italy: Urban sustainability concepts from the heritage of hydraulic infrastructures. The cases of Venice, Siena and Bologna," *J. Cultural Heritage Manage. Sustain. Develop.*, vol. 16, no. 7, pp. 111–126, Dec. 2024.

Giovanna Maria Dimitri received the Ph.D. degree in artificial intelligence from the Department of Computer Science, University of Cambridge, Cambridge, U.K., 2020.

She is a Tenure Track Assistant Professor in Artificial Intelligence with the University of Milan, Milan, Italy, and a Guest Lecturer with the University of Cambridge. Her research interests include applications of AI to a wide variety of fields of applications, with a focus on biomedical data and forensics applications, with publications in top peer reviewed journals and conferences. She is Senior AE of IEEE TRANSACTIONS IN TECHNOLOGY AND SOCIETY and *Neurocomputing Journal (Elsevier)*. She was the General Chair of CCS25 and she will be the General Chair of the Artificial Neural Networks and Pattern Recognition Workshop ANNPR (Milan, 2026). Website: https://sites.google.com/view/gmdimitri/.

Open Access funding provided by 'Universita degli Studi di Milano' within the CRUI CARE Agreement