New technologies for the conservation and preservation of cultural heritage through a bibliometric analysis

Abstract:

Purpose

This study aims to analyze the impact of *Artificial Intelligence* (AI) and *Machine Learning* (ML) on heritage conservation and preservation, and to identify relevant future research trends, by applying scientometrics.

Design/methodology/approach

A total of 1646 articles, published between 1985 and 2021, concerning research on the application of ML and AI in cultural heritage was collected from the Scopus database and analyzed using bibliometric methodologies.

Findings

Our findings have shown that although there is a very important increase in academic literature in relation to AI and ML, publications that specifically deal with these issues in relation to cultural heritage and its conservation and preservation are significantly limited.

Originality/value

This study enriches the academic outline by highlighting the limited literature in this context and therefore the need to advance the study of AI and ML as key elements that support heritage researchers and practitioners in conservation and preservation work.

Keywords: machine learning, artificial intelligence, cultural heritage, conservation

1. Introduction

Cultural heritage is considered a strategic factor that contributes to the regional socioeconomic and cultural development (Carbone 2016; Di Pietro et al. 2015; Del Barrio-Garcia and Prados-Peña 2019). Cultural heritage, in its broadest sense, is both a product and a process that provides societies with a wealth of resources inherited from the past or created in the present and passed on to future generations for their benefit. It is important to recognize that it encompasses not only tangible heritage, but also natural and intangible heritage (UNESCO 2014). Cultural heritage assets offer the potential to act as a tourist attraction, thus contributing to the development of a region (Carbone

2016). Sustainable exploitation of these assets constitutes an excellent opportunity to improve the quality of life of a community (Timothy and Boyd 2006). Seen this way, investing in culture results in improving the quality of life in a specific region by attracting new economic, financial and human resources that empower sustainable growth (Sacco et al. 2013).

On the other hand, tourism is a phenomenon that dates to ancient times. In the early 1970s, tourism begun using the natural and cultural resources of a destination as a tourist attraction (Jayapalan 2001; Gyr 2010). The link between tourism and culture has always been strong. Cultural heritage offers great motivation to travel, and the trip itself generates culture (Del Barrio-Garcia and Prados-Peña 2019). In addition, the tourist traffic attracted by the presence of cultural heritage resources contributes to an increase in satellite activities that also produce a related economic impact (Di Pietro et al. 2015).

The maintenance of a heritage site, especially if large and complex, requires an in-depth knowledge on the conservation situation, which is not achievable through traditional means of on-site inspections, when different persons retrieve subjective data. This is particularly true when natural and cultural heritage are combined and are not easily accessible. A large variety of data and skills are necessary to carry out appropriate assessments, while the growing threat requires that data on conservation be often and scientifically retrieved.

WARMEST's[1] strategic goal is to optimize maintenance procedures in cultural and natural heritage sites through the introduction of new technologies and workflows towards a novel Decision Support System (DSS) that will carry out a Cultural Heritage Risk Analysis (CHRA) and suggest improvements in maintenance and disaster management procedures. The WARMEST consortium develops new technologies and workflows to collect data on the conservation status of the sites and to analyze them, as there is presently a lack of data that are specifically relevant to understanding the impact of climate change.

In this context, our research aims to analyze the degree of development of new technologies focused on *Artificial Intelligence* (AI) and *Machine Learning* (ML) as key elements that facilitate and support those responsible for heritage sites in their

[1] Project WARMEST - loW Altitude Remote sensing for the Monitoring of the state of cultural hEritage Sites: building an inTegrated model for maintenance, is an H2020 Marie Curie Research and Innovation Staff Mobility Project (H2020- Marie Skłodowska-Curie Actions-RISE-2017), online at https://warmestproject.eu, last accessed Apr. 2022.

conservation and preservation. It is also about identifying future research trends. For this, a bibliometric analysis is proposed. Bibliometric analysis is a well-known and rigorous method for exploring and analyzing large volumes of scientific data. It allows us to discover the evolution of a specific line of research, as well as sheds light on emerging areas in that field (Donthu et al. 2021). Therefore, the proposed analysis allows addressing the following research questions:

Q1. What is the trend of scientific publications in AI in the conservation of cultural heritage?

Q2. What are the main thematic areas and the most relevant publications in relation to AI?

Q3. Who are the most representative authors, journals, institutes, and countries in the line of research?

Q4. What are the main international cooperation networks of authors, institutes, and countries?

Q5. What are the main current and future research topics in this area of knowledge?

2. Methodology

2.1. Database and bibliometric analysis methodology

Scientometrics or bibliometric analysis is a technique whose main objective is to identify, organize and analyze metadata to examine the evolution of an area of knowledge in a specific period (Donthu et al. 2021; Kumar et al. 2021). Bibliometric methodologies apply statistical and graph-theoretic tools to bibliographic data analysis (Kumar et al., 2021) and include performance analysis and scientific mapping (Donthu et al. 2021; Kumar, et al. 2021).

2.2. Methodological procedure

The methodology followed to carry out this bibliometric study is organized in three sequential phases (see Figure 1). First, in phase 1, the selection of documents and the data retrieval is carried out, as well as the establishment of the search criteria. In the second phase, only the Journal articles, from which this bibliometric analysis will be carried out are selected. At the last phase, the bibliometric analysis is carried out using the VOSviewer[2]. Figure 2 presents the details of this methodological procedure.

[2] VOSviewer is a multi-platform free software tool for constructing and visualizing bibliometric networks, created by the Centre for Science and Technology Studies, Leiden University, The Netherlands, online at https://www.vosviewer.com, last accessed Apr. 2022. In this study v.1.6.18 of the tool was used.

INSERT FIGURE 1

Phase 1 included the establishment of the search criteria and the execution of the data retrieval. Phase 2 included the filtering of the retrieved data by selecting only the journal articles and by excluding the irrelevant subject areas. Phase 3 was the main bibliometric analysis using the VOSviewer[2]; this phase included a publication analysis to identify publication trends, outlets and performance, along with leading authors, institutes and countries, using also a citation analysis; in addition, this phase included a co-authorship and a keyword co-occurrence analysis. Phase 4 of this process included the further analysis on the key-findings for the identification of potential significant patterns.

INSERT FIGURE 2

2.2.1. Retrieval of documents and data collection

The Scopus database was used for the retrieval of documents, for several reasons: . Scopus is the repository that provides the greatest volume of information in terms of authors, countries, and institutes (Zhang and Eichmann-Kalwara, 2019); it contains the largest volume of articles and journals that meet peer review scientific quality requirements (Baas et al. 2020); the coverage of WOS is less (Paul et al., 2021); it shows additional details of the publications (Nascimento and Rodrigues 2015). In conclusion, Scopus is the most suitable repository for bibliometric reviews (Donthu et al. 2021).

The search for the selected terms has been carried out in the fields “article title, abstract and keywords”. The search query was as follows:

```
( TITLE-ABS-KEY ( ( "advanced conservation" OR "cultural heritage" OR "preservation" OR "extended digitization" OR "advanced digitization" OR "preventive preservation" OR "heritage interpretation" OR "heritage restoration" OR "advanced reconstruction" OR "predictive modelling" OR "heritage analytics" OR "personalized cultural content" OR "personalized heritage" ) ) AND TITLE-ABS-KEY ( ( "machine learning" OR "artificial intelligence" ) ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( EXCLUDE ( SUBJAREA , "MEDI" ) OR EXCLUDE ( SUBJAREA , "AGRI" ) OR EXCLUDE ( SUBJAREA , "BIOC" ) OR EXCLUDE ( SUBJAREA , "HEAL" ) OR EXCLUDE ( SUBJAREA , "IMMU" ) OR EXCLUDE ( SUBJAREA , "NURS" ) OR EXCLUDE ( SUBJAREA , "PHAR" ) OR EXCLUDE ( SUBJAREA , "PSYC" ) OR EXCLUDE ( SUBJAREA , "DENT" ) OR EXCLUDE ( SUBJAREA , "VETE" ) ) AND ( EXCLUDE ( PUBYEAR , 2022 ) )
```

Data retrieval was carried out in April 2022. Initially, a total of 4489 documents were obtained that met the established search requirements. Setting the filter to only Journal articles yielded a total of 2478 documents, considered to be of high quality. This high quality is a consequence of the rigorous blind peer review process, according to Paul et al. (2021). Also, the search query excluded subject areas not in relation with this study: Medicine, Biochemistry, Genetics and Molecular Biology, Agricultural and Biological Sciences, Health Professions, Psychology, Immunology and Microbiology Pharmacology, Toxicology and Pharmaceutics, Nursing, Dentistry and Veterinary. The time frame for this research included all published research until 2021. The final search query yielded a total of 1646 documents.

2.2.2. Bibliometric analysis

The analysis included the main authors, thematic areas, countries and authors' affiliations, as well as international cooperation networks and indexing keywords. The analysis of the authors, institutes, countries and international cooperation networks was carried out based on the analysis of co-authorship. In this sense, the greater the frequency of co-authorship, the greater the interrelation among them, increasing their conceptual relationship. For the analysis of the keywords, the co-occurrence among them was taken into account. This methodology allowed to identify a conceptual and thematic structure, showing an overview of the most explored research topics in the relationship between the conservation of cultural heritage and technologies, focused on AI.

3. Results

The bibliographic analysis shows that the first article on the topics of interest was published in 1985 and that the total number of articles indexed in Scopus between then and 2021 was 1646. The following sections provide detailed information on the bibliometric analysis of the basic attributes and research topics related to academic research related to the development of "*artificial intelligence"* (AI) and "*machine learning*"(ML) in the conservation of cultural heritage to answer to the main review questions (Q1-Q5) presented in the Introduction.

3.1 Trend of publications related to AI in the conservation of cultural heritage (Q1).

This section shows the results related to the main characteristics of the scientific production on artificial intelligence (AI) and machine learning (ML) in the conservation

of cultural heritage in the period 1985-2021. First, the annual evolution since 1985 of the number of articles is shown in Figure 3.

INSERT FIGURE 3

As shown in Figure 3, 2008 marks the beginning of a significant increase in the scientific production in the domain of interest. This increase can potentially be linked to the consequences of the economic crisis of that period, which led to a drop in budget allocations for conversation projects. It can also be linked with the rapid advancement in technology, and particularly machine learning, in that period. Taking into account this evolution of scientific production, two periods are distinguished in our study, clearly differentiated: one period ranging between 1985 and 2007 that includes 92 articles corresponding to a 5.5% of all relevant scientific production; another period ranging between 2008 and 2021, which is analyzed on per year basis. Table 1 shows the main characteristics in these periods, including the number of published articles, the number of authors, countries, institutes, citations, journals, as well as the average number of citations and average number of authors in the domain of interest.

The first relevant publication was by *Herrod, Richard A.*, *Papas, Barbara* in 1985, who, discussed the issues, applications, and tools available related to Artificial Intelligence as a key factor in future applications of computers in different fields.

Since then, a total of 1646 articles have been published in this line of research, at least, available in the Scopus database.

The data in Table 1 show the significant annual increase of all the scientometric indicators analyzed in the period 2008-2021, which clearly indicates that the considered line of research has recently grown considerably and is currently acquiring a relevant dimension in the scientific literature, especially since year 2015 (Figure 3 shows how after 2015 there is a constant increasing trend). Graphical representations of key features are shown in Figures 4, 5 and 6.

INSERT TABLE 1

INSERT FIGURE 4

INSERT FIGURE 5

INSERT FIGURE 6

3.2 Thematic areas and most influential publications (Q2)

This section discusses the results in terms of their discipline and domain, as well as the most influential relevant publications. Figure 7 shows the distribution of published research articles on AI in heritage conservation grouped by their discipline and domain, as reported by the Scopus database. 27 relevant thematic domains were identified, of which Computer Science was the one with the largest volume of scientific articles, 1131, a 23.8% of all publications. Engineering followed with 962, followed by Medicine with 493, Mathematics with 348 publications. In addition, Table 2 shows the most relevant publications ordered by their impact, based on the number of citations received.

INSERT FIGURE 7

INSER TABLE 2

The publication with the highest number of citations is from *Haykin, S..*, 2005, in which the author analyzed the emergent behavior of cognitive radio. In addition, based on the cognitive radio, the author addressed three fundamental cognitive tasks: the analysis of the radio-scene; estimation of the state of the channel and predictive modeling and finally, the control of transmission power and dynamic spectrum management. *Phillips, S.J., Anderson, R.P., Schapire, R.E*, 2006, presented the use of the maximum entropy method (Maxent) to model geographic distributions of species with presence data only. Maxent is a general-purpose machine learning method with a simple and precise mathematical formulation. *Raissi, M., Perdikaris, P., Karniadakis, G.E.,* 2019, introduced physics-informed neural networks – neural networks that are trained to solve supervised learning tasks while respecting any given laws of physics described by general nonlinear partial differential equations. The authors demonstrated the effectiveness of the proposed framework through a collection of classical problems in fluids, quantum mechanics, reaction–diffusion systems, and the propagation of nonlinear shallow-water waves.

Stockwell, D.R.B., Peterson, A.T., 2002, provided a method to determine the size that a sample of biodiversity data must have in order to have the capacity to model ecological niches and predict optimal geographic distributions, reducing the cost of treating large sample sizes.

Linsker, R., 1998, explored the emergence of a feature-analyzing function from the development rules of simple, multilayered networks. His study showed that even a single developing cell of a layered network exhibits a remarkable set of optimization properties that are closely related to issues in statistics, theoretical physics, adaptive signal

processing, the formation of knowledge representation in artificial intelligence, and information theory.

Chan, R.H., Ho, C.-W., Nikolova, M., 2015, proposed a two-stage scheme to eliminate salt-and-pepper impulse noise. In terms of edge preservation and noise suppression, images restored using this method show significant improvement compared to those restored using only nonlinear filters or regularization methods only.

Zhang, L., Wu, X.., 2006, proposed a new edge-guided nonlinear interpolation technique through directional filtering and data fusion.

Miotto, R., Li, L., Kidd, B.A., Dudley, J.T., 2016, presented a novel unsupervised deep feature learning method to derive a general-purpose patient representation from EHR data that facilitates clinical predictive modeling.

Lu, J., Behbood, V., Hao, P., (...), Xue, S., Zhang, G., 2015, examined computational intelligence-based transfer learning techniques and clusters related technique developments into four main categories: neural network-based transfer learning; Bayes-based transfer learning; (fuzzy transfer learning, and applications of computational intelligence-based transfer learning.

On the other hand, Darwiche, A., Pearl, J., 1997, showed that the AGM postulates are too weak to ensure the rational preservation of conditional beliefs during belief revision, thus permitting improper responses to sequences of observations.

Apparently, after a review of the collected articles, some of them were discarded, as they were not related to the main objective of this research. In this way, it was possible to conclude that there is a main line of investigation, focused on the development of different methodologies based on machine learning for data processing.

3.3 Greater contributions and international cooperation networks (Q3 and Q4)

This section presents the results of the productivity of the authors, institutes, countries and journals, as well as their international cooperation networks. International cooperation networks make it possible to understand the relationships between researchers and the dissemination of knowledge (Chen 2006), while collaborations allow the generation of new high-impact research by generating synergies that contribute to the exchange of ideas (Acedo et al. 2006). In the international cooperation maps (Figure 8), the size of the circles indicates the number of published scientific papers, the colors indicate the cooperation clusters, and the distance refers to the frequency of co-authored

publications. Table 3 shows the 10 most productive authors in the line of research in the period 2008-2021, as well as their main characteristics.

INSERT TABLE 3

As seen in Table 3, the most productive authors are *Wu, D* from United States with 10 published articles; *Niyato D* and *Tang C.* from Singapore and China, respectively, since they are the authors with the most publications (6 published articles each), followed by *Kang, J.* of Australian origin, with 5 published articles. Regarding the dissemination of the results, *Wu, D., with 475 citations, followed by Niyato and D, Kang, J.* with 475 and 469 citations, respectively. Regarding the average number of citations per article, *Kang, J* stands out, with 94 citations per article, followed by *Niyato, D,* with 78.8 citations per article, the latter being the one with the most H Index, with 83. It is noteworthy that of the ten most productive authors, most have carried out their work in recent years, beginning in 2017 their research work in this area of study.

Figure 8 shows the relevant international cooperation networks of the co-authors. The total of 5189 authors, was selected with a minimum interaction of 4 published articles, a total of 98 authors were obtained, of which 86 of them make up the 9 international cooperation cluster in the line of research.

INSERT FIGURE 8

It stands out that, among the 10 most productive authors in the research of interest, some of the authors cooperate. In this sense, it is worth highlighting *Niyato, D.*, and *Kang J.*, both belong to the same network composed of a total of 11 co-authors; followed by *Wu D*, with an international cooperation network made up of 9 co-author. *Liu X. and Tang C,* both belong to the same network composed of a total of 7 co-authors.

On the other hand, it should be noted that of these 10 most productive authors, 5 do not have an international collaboration network, *Casolla, G.*, and *Cuomo, S.* from Italy; *Zhang, M.L*. from China; *Budka, M.*S from United Kingdom and *Ghosh, I.,* from India.

According to these data, it can be affirmed that in the line of research related to heritage conservation and AI, there is a wide network of international cooperation, centered around United States, China and Singapore. Furthermore, Table 4 shows the ranking of the 10 most productive relevant institutes, along with the rates of international cooperation in the period between 2008 and 2021, a period of expansion of the scientific production.

INSERT TABLE 4

The most productive institutes are the *Ministry of Education China* and the *Nanyang Technological University* of Singapore with 32 and 23 articles published respectively. The *Nanyang Technological University* of Singapore is the one with the greater impact, as revealed by the highest number of citations (1108). However, the Stanford University is the one with the highest number of citations per article (64.5). It is followed, in terms of the average number of citations per article, by the Nanyang Technological University, with 48.2 citations per article. Far behind, it is followed by *Massachusetts Institute of Technology*, with 34.1 and the rest of the institutes are well below this number. The *Chinese Academy of Sciences* and *Southeast University*, both in China, are the institutes with the lowest citation averages, with 11.8 and 9.4, respectively. Regarding the H index, *Ministry of Education China* is the one with the highest value, achieving an H index of 8. With regard to international cooperation, it should be noted that all of the ten most productive institutes publish in international cooperation. Of these, the *Nanyang Technological University* of Singapore, stand out as one of publish more articles in international cooperation compared to those with national co-authors. This institute present a cooperation index (CI) of 78.3%. It is followed by the *Southeast University*, *University of Electronic Science and Technology and Xidian University* of China. These institutes present the same cooperation index CI of 53.3%. The University of Central Florida of United States is the institute that have the lowest rates of international cooperation, with a cooperation index (CI) of 7.1%,

Figure 9 shows the international cooperation networks of institutes in the relevant research. For a total of 4054 identified institutes, an interaction of at least 2 published research articles was selected, and 98 international institutes were identified.

INSERT FIGURE 9

The figure 9 shows an international cooperation network made up of 7 institutes grouped into 2 international cooperation clusters. One of them formed by 4 institutes and the other by three. It highlights that none of the ten most productive institutes are part of an international cooperation network.

Regarding the most productive countries and their cooperation results, Table 5 presents the results of the top 10 in the line of research of interest in the period 1985-2021.

INSERT TABLE 5

United States and China are the countries with the highest production, with 402 and 305 articles published, respectively. They are also the ones that achieve a greater impact, based on the number of citations, with 23132 in the case of United States and 5033 for

the articles from China. It is noteworthy that although the number of articles published by these two countries is the similar magnitude (China 24% less than the United States) the impact of those from the United States is 78.2% higher than those from China. With respect to the H index, these two countries are also the ones that stand out the most compared to the rest with values 45 and 83 respectively.

With respect to the ratio of total citations received per article, Canada and United States are the ones that stand out the most compared to the rest with values 177.3 and 57.5 respectively. Canada stands out especially for the work of *Haykin, S*. 2005 that reaches 9991 citations. The rest of the countries are all above a threshold of 30, with the exception of Germany with 29 and Spain with 25.

Finally, with regard to international cooperation, the data analyzed show a high level of international collaboration. The United Kingdom stands out with an index of 67.9%, followed by Australia with 66.3%, France with 66% and Canada with 60%. Of the 10 countries with most articles, practically all, except Italy with 35.79% and India with 31.3%, are above to 40%. The United States with 56 and the United Kingdom with 44 are the countries with the highest number of collaborators.

INSERT FIGURE 10

Figure 10 shows the international cooperation networks of the countries in the line of research of interest, since the first article was published in 1985. 166 countries were identified. An interaction of at least five published articles was considered, which gave rise to eight international cooperation networks made up of 51 countries. The networks shown in the figure confirm the high level of international cooperation in this line of research.

Finally, Table 6 shows the 10 most productive journals in the line of research of interest, and their main characteristics since the first article was published.

INSERT TABLE 6

IEEE Access is the most productive journal in the publication of research articles in the study domain, with a total of 59 publications, followed by *Lecture Notes In Computer Science* with 32 and *Neurocomputing* and *Information Sciences* with 30 and 21 published articles each. However, it is *Scientific Reports* that obtains the greatest impact per article, 49.24 citations per article. Regarding the impact of the articles, *Scientific Reports* stands out as being the journal with the highest number of citations, 837, closely followed by *Expert Systems With Applications*, with 656. *Lecture Notes In Computer Science*

Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics has the highest H Index (400).

3.4 Current and future research topics in sustainable crafts (Q5)

This section shows the results of the keyword co-occurrence analysis, which has been analyzed for the period 1985-2021. According to Weinberg (1974), co-occurrence is based on the fact that records that share the same keywords are related; they are representative of the content of the research articles, and create an image of the research line (Comerio and Strozzi, 2019). According to Park and Nagy (2018), VOSviewer, elaborates the matrix of keywords based on the extraction and calculation of the frequency.

For the 805 research articles published between 1985 and 2021, 10565 keywords have been identified. After an interaction of at least 20 co-occurrences, a total of 121 keywords were obtained. Next, a filtering process was performed. In this process, the keywords included in the search and those words that are not related to the investigation were eliminated, in order to avoid reaching erroneous conclusions. As a result of this process, a final number of 82 keywords grouped into four clusters was obtained, as shown in Figure 11.

INSERT FIGURE 11

#1. Red

This cluster is made up of 40 keywords. The search shows 206 documents. The contribution with the most citations is that of *Lu, J., Behbood, V., Hao, P., Xue, S., and Zhang, G.,* i2015., in which the authors examined computational intelligence-based transfer learning techniques and clusters related technique developments into four main categories: (a) neural network-based transfer learning; (b) Bayes-based transfer learning; (c) fuzzy transfer learning, and (d) applications of computational intelligence-based transfer learning.

For their part, *Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., and Muys, B.*, 2010, and in the field of Forestry science, in order to establish the most suitable model in the relationship between stand productivity and abiotic and biotic site characteristics, such as climate, topography, soil and vegetation, these authors compared and evaluated five different modeling techniques: multiple linear regression (MLR), classification and

regression trees (CART), boosted regression trees (BRT).), generalized additive models (GAM), and artificial neural networks (ANN).

Regarding the energy efficiency of buildings, *Chou, J.-S., and Bui D.-K.,* 2014, estimated their energy performance using various data mining techniques, including support vector regression (SVR), artificial neural network (ANN), classification and regression tree, chi-squared automatic interaction detector, general linear regression, and ensemble inference model. They showed that the ensemble approach (SVR +ANN) and SVR were the best models for predicting cooling load (CL) and heating load (HL). Along the same lines, *Deng, H., Fannon, D., and Eckelman, M.J.* 2018., tested different predictive modeling approaches for estimating energy use Intensity for both United States commercial office and end-users buildings. Among the analyzed approaches stand out Support Vector Machine, Random Forest and machine learning algorithms. With a similar concern regarding energy consumption, and considering buildings as the main responsible for this consumption, *Khosravani, H.R., Castilla, M.D.M., Berenguel, M., Ruano, A.E., and Ferreira, P.M.* 2016, compared a neural network model that was designed using statistical and analytical methods, with a group of neural network models designed benefiting from a multi-objective genetic algorithm. In addition, the neural network models were compared to a basic naive autoregressive model.

Li, H., Parikh, D., He, Q., (...), Fang, D.,and Hampapur, A., 2014, explored several avenues to machine learning techniques including distributed learning and hierarchical analytical approaches, applied to increase the speed of the rail network without compromising safety and trying to avoid interruptions or delays, taking advantage of the large volume of data provided by the extensive network of mechanical condition detectors implemented on the tracks, such as temperature, deformation, vision, infrared, weight, impact, etc.

On the other hand, *Sudakov, O., Burnaev, E., and Koroteev, D.,* 2019, demonstrated the applicability of machine learning for image-based permeability prediction. To do this, they relied on a training set containing 3D scans of Berea sandstone subsamples imaged with X-ray microtomography and corresponding permeability values simulated with Pore Network approach. Also, they used Minkowski functionals and Deep Learning-based descriptors of 3D images and 2D slices as input features for predictive model training and prediction.

Regarding this cluster, it could be concluded that it is about valuing the use of machine learning techniques in different areas.

#1. Green

This cluster is made up of 19 keywords. The search shows 156 documents. The contributions that received the most citations is the articles published by *Chan, R.H., Ho, C.-W. and Nikolova,* M., 2005 and *Zhang, L. and Wu, X, 2006;* whose works have already been mentioned previously. In both cases related to image processing. In the same line, regarding the treatment of images, *Ren, Z., Gao, S., Chia, L. T., and Tsang,* (2013 applied Laplacian sparse coding (LSc) to feature quantization in Bag-of-Words image representation. With this technique, these authors achieve good performance regarding the solution of the image classification problem as well as successfully solving the problem of semi-automatic image labeling. Following with the treatment of images, *Brunet, D., Vrscay, E.R., Wang, Z.,* 2012, build a series of normalized and generalized metrics based on the advantages presented by the structural similarity index (SSIM) to evaluate the quality of the images and the performance of algorithms and systems image processing. According to these authors, the developed work expands the potential of SSIM both in theoretical development and in practical applications. For their part, *Brown, M.S. and Seals, W.B.*, 2004, presented a method to restore images of deformed documents. According to these authors, their method is designed to be used in the digitization of very old and damaged manuscripts, generally existing in libraries and museums.

On the other hand, *Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., and Ribeiro, L.*, 2014, apply the Random Forest (RF) method to improve decisions related to watershed management, establishing an accurate predictive model of the occurrence of pollutants. RF is a powerful data-driven method of machine learning. Key advantages of RF include: parametric nature; high predictive accuracy; and ability to determine the importance of variables. The prediction results show the ability of RF to build accurate models with strong predictive capabilities.

Sarma, P., Durlofsky, L.J., Aziz, K. 2008, described a new approach to create an efficient, general and differentiable parameterization of large-scale non-Gaussian, non-stationary random fields (represented by multipoint geostatistics) capable of reproducing complex geological structures. Such parameterizations are appropriate for use with gradient-based algorithms applied, for example, to historical matching or uncertainty propagation.

In summary, it can be concluded that in this cluster, fundamentally, those works related to methodologies for the treatment of images and their restoration are collected.

#3. Blue

This cluster is made up of 15 keywords. The search shows 21 documents. The contributions that received the most citations is the articles published by *Susto, G.A., Schirru, A., Pampuri, S., McLoone, S.* 2016. In this study, a functional learning paradigm is exploited in a supervised fashion to derive continuous smooth estimates of time-series data (yielding aggregated local information), while simultaneously estimating a continuous shape function yielding optimal predictions.

Tapete, D., and Cigna, F., 2018, evaluated the potential of the multispectral constellation Sentinel-2 of the European Commission Earth observation program Copernicus to detect prominent features and changes in heritage sites, during both ordinary times and crisis. **They** test the 10 m spatial resolution of the 3 visible spectral bands of Sentinel-2 for substantiation of single local events. By screening long Sentinel-2 time series, the authors demonstrated that changes of textural properties and surface reflectance can be logged accurately in time and space and can be associated to events relevant for conservation.

Uhl, J.H., Leyk, S., Chiang, Y.-Y., Duan, W., Knoblock, C.A. 2020, proposed "an automated machine-learning based framework to extract human settlement symbols, such as buildings and urban areas from historical topographic maps in the absence of training data, employing contemporary geospatial data as ancillary data to guide the collection of training samples. These samples are then used to train a convolutional neural network for semantic image segmentation, allowing for the extraction of human settlement patterns in an analysis-ready geospatial vector data format".

Macher, H., Landes, T., Grussenmeyer, P., Alby, E., 2014, presented a semi-automatic approach for creating a 3D model from point clouds.

Lamas, A., Tabik, S., Cruz, P., (...), Cruz, T., Herrera, F., 2021, based on the idea that an automatic system based on an image can help identify the architectural style or to detect the architectural elements of a monument and consequently to improve knowledge in art and history, presented the MonuMAI framework (Monument with Mathematics and Artificial Intelligence).

This cluster highlights the use of machine learning and artificial intelligence in different applications for heritage conservation.

#3. Yellow

This cluster is made up of 7 keywords. The search shows 27 documents. The contributions that received the most citations is the articles published by *Phillips, S.J., Anderson,*

R.P., Schapire, R.E. 2006. in which, as we referred previously, presented the use of the maximum entropy method (Maxent) to model geographic distributions of species with only presence data. Maxent is a general-purpose machine learning method, which according to these authors, consists of a simple and precise mathematical formulation, which makes it suitable for species distribution modeling. Following the same line, in the field of biodiversity, *Stockwell, D.R.B., and Peterson, A.T.,* 2002, explored sample size needs for accurate modeling for three predictive modeling methods via re-sampling of data for well-sampled species, and developed curves of model improvement with increasing sample size; being the models based on machine learning methods the most successful.

Sarkar, A., Lathia, N., Mascolo, C., 2015, examined how the online maps which were built to inform users of the state of urban bicycle sharing systems can be used to analyze, compare and predict mobility across the growing number of cities that are adopting these forms of transport. To do this, these authors used random forests and neural networks to compare the accuracy of forecasting how many bicycles will be at a given station and time to two baselines, and evaluate how prediction accuracy varies across cities.

Oonk, S., Spijker, J. 2015, applied data fusion of multi-element XRF results from archaeological feature soils and regional background soils to assess the complementary value of geochemistry and machine learning in predictive modeling in archaeology. These authors integrated multiple data sources, trained learning models for archaeological soil and background soil classification, and compared model predictions for three validation areas with current archaeological interpretation and established predictive models.

Villarin, M.C., Rodriguez-Galiano, V.F., 2019, showed the application of machine learning (ML) methods to build a predictive model of water demand in the city of Seville, Spain, at the census tract level. Tehy uised a classification and regression trees (CART) and random forest (RF), a multivariate, spatially nonstationary and nonlinear ML approach. In a similar vein, *Granata, F.,and Di Nunno, F.h.* 2021, developed several different forecasting models in order to predict the tide level of the city of Venice. Each model was built in three variants, varying the implemented machine learning algorithm: M5P Regression Tree, Random Forest and Multilayer Perceptron.

This cluster is focused on the development of predictive models with different technologies based on AI and with application in different contexts, one of them being cultural heritage.

Finally, Figure 12 shows the network of keywords of this period, ordered from darker to lighter color, to highlight future research trends. It is clearly observed how future research trends in the line of research of interest are mainly focused on predictive analytic, the learning systems and machine learning.

INSERT FIGURE 12

4. Conclusions

This article proposed a scientometric methodology to study the application of AI in the conservation of cultural heritage, including research articles in the period ranging between 1993 and 2021. A bibliometric review of 845 relevant research articles retrieved from the Scopus database was carried out for the selected period, generating four main conclusions, summarized below.

C1. The main characteristics of the domain of interest show a strong growth as of 2008, which indicates the interest in the academic community. The high number of topics presented in Figure 7, as well as the weight of each of them, shows a high multidisciplinary in this area of knowledge.

C2 In the top-ten list of most relevant publications, none appears directly related to the conservation of cultural heritage. Machine learning and AI are increasingly applied as robust pattern recognition methods in virtually all scientific domains. Still, many of the included publications, although specifically related to cultural heritage, they do not relate to AI but only to computer vision or another technical domain. Only very recent works specifically target the bridging of AI with cultural heritage.

C3. The most prolific authors in this research line have been *Wu, D* from United States and *Niyato D* and *Tang C*. from Singapore and China; while *Ministry of Education China* and the *Nanyang Technological University* of Singapore are the most productive institutes. *United States* is the most prolific country, while the most productive journal is *IEEE Access*.

C4. In the line of research related to heritage conservation and AI, there is a wide network of international cooperation, centered on United States, China and Singapore.

C5. Multiple research themes have been identified. On the one hand, works related to methodologies for the treatment of images and their restoration have been identified. On the other hand, the application of machine earning in different areas stands out. A third line could be focused on the development of predictive models with different technologies

based on AI and with application in different contexts, one of them being cultural heritage. Finally, the use of machine learning and artificial intelligence in different applications for heritage conservation.

C6. Regarding future research trends, they are mainly focused on predictive analytics, the learning systems and machine learning.

Our findings have shown that although there is a very important increase in the academic literature in relation to technologies such as AI, ML and cultural heritage, the publications that specifically deal with these topics are neither popular (they are not the most cited) nor the main authors have delved into this line of research, therefore, this fact is striking, taking into account the importance of cultural heritage and its conservation, in the development of the territories, in the generation of employment and wealth (Del Barrio-García and Prados-Peña, 2019).

5. Limitations

We should note that this study has certain limitations that should be considered for future research. In our case, data retrieval experiments in the Scopus database turned out that creating and executing a meaningful query, which would be expected to present relevant results is not a trivial task. Small changes in the search criteria results in significantly different numbers of corresponding publications. Common criteria in all searches were that the publications are journal articles, published in English language, any time before (excluding) the year 2022. The rest of the criteria involved the selection of key phrases to find in the articles' title, abstract and keywords, and any possibly needed subject area exclusion filters.

Thus, in future works it would be interesting to include in the analysis all the types of documents that can be retrieved from Scopus, as well as the incorporation of the search in Web of Science or Google Scholar. Finally, the bibliometric analysis methodology does not take into account that the citations require time to be analyzed. Therefore, a systematic content analysis could provide a complementary method to evaluate research in the future.

References

Acedo, F. J., Barroso, C., Casanueva, C., and Galán, J. L. 2006. "Co-authorship in management and organizational studies: An empirical and network analysis". *Journal of Management Studies, 43(5),* 957-983. doi:10.1111/j.1467-6486.2006.00625.x

Aertsen, W., Kint, V., Van Orshoven, J., Özkan, K., and Muys, B. 2010. “Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests”. *Ecological modelling*, *221*(8), 1119-1130.

Baas, J., Schotten, M., Plume, A., Côté, G., and Karimi, R. 2020. “Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies”. *Quantitative Science Studies*, *1*(1), 377-386.

Brown, M. S., and Seales, W. B. 2004. “Image restoration of arbitrarily warped documents”. *IEEE Transactions on pattern analysis and machine intelligence*, *26*(10), 1295-1306.

Brunet, D., Vrscay, E. R., and Wang, Z. 2011. “On the mathematical properties of the structural similarity index”. *IEEE Transactions on Image Processing*, *21*(4), 1488-1499.

Carbone, F. 2016. “An insight into cultural heritage management of tourism destinations”. *European Journal of Tourism Research*, *14*, 75-91.

Chan, R. H., Ho, C. W., and Nikolova, M. 2005. “Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization”. *IEEE Transactions on image processing*, *14*(10), 1479-1485.

Chen, C. 2006. “CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature”. *Journal of the American Society for information Science and Technology*, *57*(3), 359-377.

Chou, J. S., and Bui, D. K. 2014. “Modeling heating and cooling loads by artificial intelligence for energy-efficient building design”. *Energy and Buildings*, *82*, 437-446.

Comerio, N., and Strozzi, F. 2019. “Tourism and its economic impact: A literature review using bibliometric tools”. *Tourism economics*, 25(1), 109-131. doi: 10.1177/1354816618793762

Darwiche, A., and Pearl, J. 1997. “On the logic of iterated belief revision”. *Artificial intelligence*, *89*(1-2), 1-29.

Del Barrio-García, S., and Prados-Peña, M. B. 2019. “Do brand authenticity and brand credibility facilitate brand equity? The case of heritage destination brand extension”. *Journal of Destination Marketing & Management*, *13*, 10-23.

Deng, H., Fannon, D. and Eckelman, M. J. 2018.”Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata”. *Energy and Buildings*, *163*, 34-43.

Di Pietro, L., Mugion, R. G., Mattia, G., and Renzi, M. F. 2015. “Cultural heritage and consumer behaviour: A survey on Italian cultural visitors”. *Journal of Cultural Heritage Management and Sustainable Development*, 5(1), 61-81. https://doi.org/10.1108/JCHMSD-03-2013-0009.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., and Lim, W. M. 2021. “How to conduct a bibliometric analysis: An overview and guidelines”. *Journal of Business Research*, *133*, 285-296.

Granata, F., and Di Nunno, F. 2021. “Intelligence models for prediction of the tide level in Venice”. *Stochastic Environmental Research and Risk Assessment*, *35*(12), 2537-2548.

Gyr, U. 2010. *The history of tourism: Structures on the path to modernity*. Notes, 2, 1–18.

Haykin, S. 2005. “Cognitive radio: brain-empowered wireless communications”. *IEEE journal on selected areas in communications*, *23*(2), 201-220.

Herrod, R. A., and Papas, B. C. 1989. “Industrial applications of artificial intelligence”. In *Conference Record of Annual Pulp and Paper Industry Technical Conference*, (pp. 86-90). IEEE.

Jayapalan, N. 2001. *Introduction to tourism*. Atlantic Publishers & Dist.

Khosravani, H. R., Castilla, M. D. M., Berenguel, M., Ruano, A. E., and Ferreira, P. M. 2016. "A comparison of energy consumption prediction models based on neural networks of a bioclimatic building". *Energies*, *9*(1), 57.

Kumar, S., Lim, W. M., Pandey, N., and Christopher Westland, J. 2021. "20 years of electronic commerce research". *Electronic Commerce Research*, *21*(1), 1-40.

Lamas, A., Tabik, S., Cruz, P., Montes, R., Martínez-Sevilla, Á., Cruz, T., and Herrera, F. 2021. "MonuMAI: Dataset, deep learning pipeline and citizen science based app for monumental heritage taxonomy and classification". *Neurocomputing*, *420*, 266-280.

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., and Hampapur, A. 2014. "Improving rail network velocity: A machine learning approach to predictive maintenance". *Transportation Research Part C: Emerging Technologies*, *45*, 17-26.

Linsker, R. 1988. "Self-organization in a perceptual network". *Computer*, *21*(3), 105-117.

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., and Zhang, G. 2015. "Transfer learning using computational intelligence: A survey". *Knowledge-Based Systems*, *80*, 14-23.

Macher, H., Landes, T., Grussenmeyer, P., and Alby, E. 2014. "Semi-automatic segmentation and modelling from point clouds towards historical building information modelling". In *Euro-Mediterranean Conference* (pp. 111-120). Springer, Cham.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. 2016. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". *Scientific reports*, *6*(1), 1-10.

Oonk, S., and Spijker, J. 2015. "A supervised machine-learning approach towards geochemical predictive modelling in archaeology". *Journal of archaeological science*, *59*, 80-88.

Park, J. Y., and Nagy, Z. 2018. "Comprehensive analysis of the relationship between thermal comfort and building control research-A data-driven literature review". *Renewable and Sustainable Energy Reviews, 82*, 2664-2679. doi: 10.1016/j.rser.2017.09.102.

Paul, J., Lim, W. M., O'Cass, A., Hao, A. W., and Bresciani, S. 2021. "Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR)". *International Journal of Consumer Studies*, 45. doi:10.1111/ijcs.1269

Phillips, S. J., Anderson, R. P., and Schapire, R. E. 2006. "Maximum entropy modeling of species geographic distributions". *Ecological modelling*, *190*(3-4), 231-259.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. 2019. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". *Journal of Computational physics*, *378*, 686-707.

Ren, Z., Gao, S., Chia, L. T., and Tsang, I. W. H. 2013. "Region-based saliency detection and its application in object recognition". *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(5), 769-779.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., and Ribeiro, L. 2014. "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain)". *Science of the Total Environment*, *476*, 189-206.

Sacco, P. L., Ferilli, G., Blessi, G. T., and Nuccio, M. 2013. "Culture as an engine of local development processes: System-wide cultural districts I: Theory". *Growth and change*, *44*(4), 555-570.

Sarkar, A., Lathia, N., and Mascolo, C. 2015. ·Comparing cities' cycling patterns using online shared bicycle maps". *Transportation*, *42*(4), 541-559.

Sarma, P., Durlofsky, L. J., and Aziz, K. 2008. “Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics”. *Mathematical Geosciences*, *40*(1), 3-32.

Stockwell, D. R., and Peterson, A. T. 2002. “Effects of sample size on accuracy of species distribution models”. *Ecological modelling*, *148*(1), 1-13.

Sudakov, O., Burnaev, E., and Koroteev, D. 2019. “Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks”. *Computers & geosciences*, *127*, 91-98.

Susto, G. A., Schirru, A., Pampuri, S., and McLoone, S. 2015.”Supervised aggregative feature extraction for big data time series regression”. *IEEE Transactions on Industrial Informatics*, *12*(3), 1243-1252.

Tapete, D., and Cigna, F. 2018. “Appraisal of opportunities and perspectives for the systematic condition assessment of heritage sites with copernicus Sentinel-2 high-resolution multispectral imagery”. *Remote Sensing*, *10*(4), 561.

Timothy, D. J., and Boyd, S. W. 2006. “Heritage tourism in the 21st century: Valued traditions and new perspectives”. *Journal of Heritage Tourism*, *1*(1), 1-16.

Uhl, J. H., Leyk, S., Chiang, Y. Y., Duan, W., and Knoblock, C. A. 2019. “Automated extraction of human settlement patterns from historical topographic map series using weakly supervised convolutional neural networks”. *IEEE Access*, *8*, 6978-6996.

Unesco 2014. “Index of development of a multidimensional framework for the sustainability of Heritage”. Available at https://es.unesco.org/creativity/sites/creativity/files/digital-library/cdis/Patrimonio.pdf. Acceso 20/10/2017.

Villarin, M. C., and Rodriguez-Galiano, V. F. 2019. “Machine learning for modeling water demand”. *Journal of Water Resources Planning and Management*, *145*(5), 04019017.

Weinberg, B. H. 1974. “Bibliographic coupling: A review”. *Information Storage and Retrieval, 10(5-6),* 189-196. doi: 10.1016/0020-0271(74)90058-8.

Zhang, L., and Eichmann-Kalwara, N. 2019. “Mapping the scholarly literature found in Scopus on “research data management”: A bibliometric and data visualization approach”. *Journal of Librarianship and Scholarly Communication*, *7*(1).

Zhang, L., and Wu, X. 2006. “An edge-guided image interpolation algorithm via directional filtering and data fusion”. *IEEE transactions on Image Processing*, *15*(8), 2226-2238.