Journal of Artificial Intelligence Research, vol. 1, no.2, pp. 15-21, NOVEMBER 2024

Application of AI in the Field of Documentary Heritage: A Review of the Literature

Yaohan Lu[a]

[a]School of Information Resource Management, Renmin University of China, Beijing, China

ARTICLE INFO

Keywords:
artificial intelligence
documentary heritage
conservation research

ABSTRACT

The advent of artificial intelligence has precipitated a transformation in the manner in which documentary heritage is researched, safeguarded and utilized. This paper presents a systematic review of the literature on the use of artificial intelligence technology in the core database of Web of Science for the protection of documentary heritage. It reviews all the studies on the use of artificial intelligence technology in the core database of Web of Science to participate in the preservation of documentary heritage, analyzes the changing trend of the number of published documents, uses VOSviewer to draw keyword co-graph, author contribution graph, citation analysis graph and coupling analysis graph, and analyzes author groups according to Lotka's law and Price's law. Finally, the paper summarizes the characteristics and shortcomings of artificial intelligence research in the field of document protection, and looks forward to the possible development direction of this field in the future.

1. Introduction

It is commonly accepted that the production of metal tools, the advent of writing and the formation of states are the three principal indications of the progression of human society towards civilization [42]. A document is defined as an entity comprising information content and its carrier in digital or analogue format. Documentary heritage is composed of words, symbols and codes, which can be preserved and copied. Documentary heritage is a mirror of human memory, encompassing the recording and transmission of human knowledge and thoughts, as well as the representation of cultural, linguistic, and national diversity. However, due to the inadequate preservation environment, all forms of documentary heritage are susceptible to continuous deterioration, including fracture, moths, mold pollution, and other deterioration processes.

In order to safeguard the collective memory of human civilization, countries around the world are engaged in ongoing endeavors to safeguard cultural heritage. As early as 1992, UNESCO initiated the "Memory of the World" project, with the objective of fostering public awareness of the importance and necessity of the preservation and utilization of documentary heritage. In the same year, N. C. Burckel conducted a study on the utilization of national documentary heritage [8]. In 1995, A. Abid sorted out the selection criteria and process for the "Memory of the World" and proposed that financing methods should be appropriately enhanced to facilitate more optimal conditions for the preservation and utilization of heritage [1]. In 2017, the Norwegian technology company Piql inaugurated the world's inaugural World Data Archive in the Arctic archipelago of Svalbard. This facility is utilized for the storage of optical film copies of assorted historical documents, which can be preserved for a minimum of 500 years, or even 1,000 years in the case of the optimal storage conditions.

With the development of modern science and technology, the safeguarding of cultural heritage has undergone a progressive transition from a primary focus on the physical preservation of artefacts to a more encompassing approach that encompasses the prevention and protection of deterioration in digital formats. On the one hand, it is because the original has been damaged and only digital copies or digital restoration will cause less damage to the original. On the other hand, it is hoped that through the restoration of digital copies, the original appearance of cultural relics can be more clearly displayed, and the interaction between audiences and researchers and documentary heritage can be increased [52]. In the process of restoring digital copies, it was found that artificial intelligence provided new ideas for restoration. Artificial intelligence (AI) uses data, algorithms, and other technologies to gain knowledge [14]. With the development of modern science and technology, the application of artificial intelligence technology in all aspects of society is becoming more mature.

Although the evidence presented in previous studies indicating the potential applications of artificial intelligence in the field of cultural heritage protection [25]. However, documentary heritage, as a significant element of cultural heritage, encompasses not only the extensive and nationally representative assortment of material and intangible cultural heritage, but also exhibits a heightened distinctiveness due to its intrinsic "archive" quality, serving as a repository of information and a medium for recording. Therefore, it is necessary to use an article to study the integration and application of document heritage protection and artificial intelligence technology.

The protection of documentary heritage requires the collaboration of art galleries, libraries, archives, museums (GLAM) and other institutions. What are the characteristics of artificial intelligence research in the field of document protection? What are the research directions that scholars are

DOI: https://doi.org/10.70891/JAIR.2024.110005
ISSN of JAIR: 3078-5529

concerned about? How much attention is paid to this field by scholars around the world? What advantages can artificial intelligence play in the field of document protection in the future? These are the problems that this paper tries to study.

2. Methods

In this study, data were retrieved from the Web of Science in order to form a dataset, which was then analyzed with the assistance of VOSviewer in order to study the characteristics of the author, key words, literature citations, coupling network, and so forth.

2.1. Searching the Literature

The dataset employed in this study, derived from the Web of Science Core Collection, encompassed all editions. Take "Documentary Heritage OR Ancient text* OR Ancient manuscript* OR Ancient book* OR Ancient writing*" as the topic, and "AI OR Artificial intelligence OR Large language model OR Image processing OR Deep learning OR Machine learning OR Natural language processing" was searched in the full text. The 117 search results comprise all articles, review articles, early accesses and other types of articles from the database, dated on or before 31 October 2024.

2.2. Analyzing Data

2.2.1. Publication Time Analysis

The number of publications on the same topic in different years usually reflects the characteristics of the era of research on the topic. The search results will present a line chart of the number of publications, with the horizontal axis representing the publication time and the vertical axis representing the number of publications. As illustrated in Figure 1, the number of articles utilizing artificial intelligence in the domain of documentary heritage is exhibiting a general upward trend.

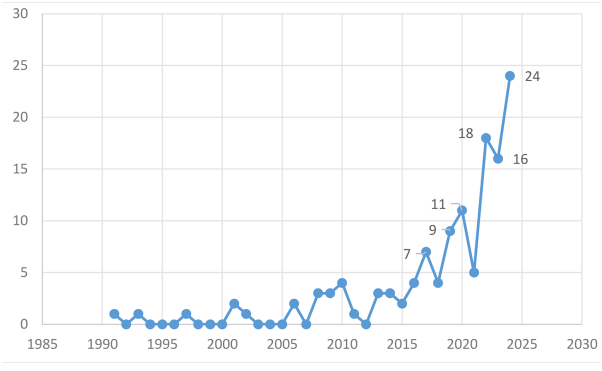

Figure 1: Line chart of the number of issued documents.

Prior to the 21st century, there were few articles, the earliest of which came from 1991, in which the authors studied and designed a system to buffer moisture and facilitate the transportation of documents [9]. Since 2017, the annual number of published papers is almost 5 or more, and the annual number of published papers between 2022 and 2024 is more than 15. These developments suggest that scholars engaged in the field of document protection are directing increasing attention towards the advancement and implementation of artificial intelligence (AI) technology. This trend also reflects the advancement of AI technology and its broader applicability. In the contemporary era, the field of artificial intelligence is experiencing a period of significant growth. It is projected that the number of documents pertaining to this domain will continue to fluctuate or reach a plateau at a high level in the near future.

2.2.2. Keyword Analysis

The key words are the refining of the core content and research direction of the article. The main research content of artificial intelligence in the field of document protection and utilization can be summarized through the keywords of the research literature. By establishing a threshold, the keywords are classified into three categories, represented by red, blue, and green. The size of the nodes in the graph is indicative of the frequency of occurrence of the word in question. The lines between the nodes show that the words are related. See Figure 2.

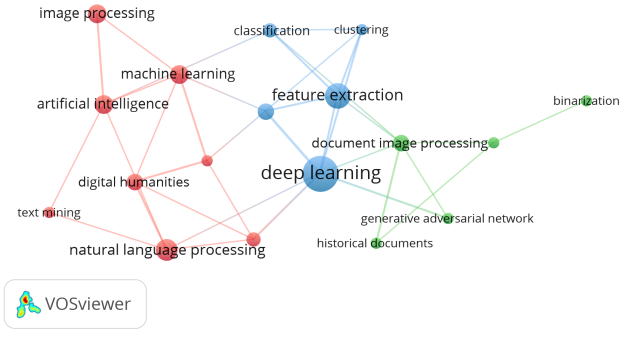

Figure 2: Keywords cluster analysis graph.

The first part of keyword clustering is represented by the red section of the figure, which encompasses eight keywords, including "artificial intelligence", "natural language processing", and so on. Most of this part is an overview of artificial intelligence in the field of document protection and research. M. Pennacchiotti employed natural language processing tools to examine ancient Italian texts and assess their utility, as well as to evaluate the potential for enhancing the performance of natural language processing tools through a series of tests and training iterations [39]. S. Chadha used artificial intelligence to train on existing and self-generated data sets, thereby developing a prediction and translation system that exhibited greater accuracy than that of traditional translation software [11]. Artificial intelligence technology is also often applied to the reading of ancient books, and these articles make a close connection between digital technology and human history. E. Kogkitsidou applied five geographical named entity recognition tools to build maps displaying the locations mentioned in ancient

Table 1
List of the top 10 countries with the highest number of publications.

No.	Country	Documents	Citations
1	People's Republic of China	39	101
2	Italy	15	112
3	England	10	99
4	India	8	61
5	America	8	122
6	Australia	7	84
7	Algeria	5	15
8	France	5	13
9	Saudi Arabia	5	5
10	Netherlands	4	38

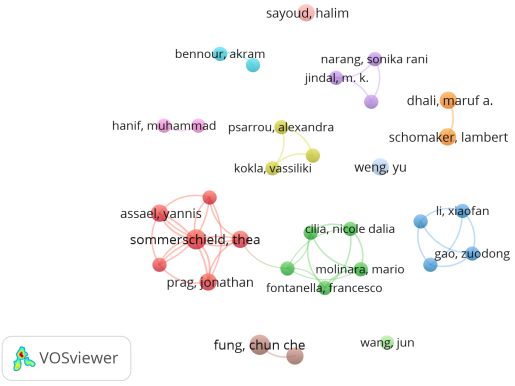

Figure 3: Author cooperation network.

texts [30]. L. X. Wang expanded the CABC ontology to digitally model the catalog of ancient Chinese books [50].

In recent years, a plethora of many machine learning models [21] such as deep learning and graph learning models have been developed [53], which provide crucial technical support for artificial intelligence in the identification of documentary heritage. The blue section of the figure primarily comprises five keywords, including "deep learning", which represents a set of techniques for digital text processing by computers. In the field of document conservation, deep learning is often used to read ancient scripts, such as Greek inscriptions [3], ancient Yi [15], ancient Aegean scripts [18], Arabic [2], etc.

Cluster three is the green part, including "document image processing", "binary", and so on. R. Chamchong attempted to develop an image processing system for the purpose of extracting text and characters from manuscripts through the utilization of binarisation techniques [13].

2.2.3. Author Analysis

Among the 117 articles used in the study, the authors came from all over the world, with China, Italy and the United Kingdom being the top three countries. See Table 1.

The most prolific author is T. Sommerschield, who has published four articles on this topic between 2019 and 2023, which have been cited 83 times. Furthermore, he is the most cited author in this field. He mainly uses artificial intelligence techniques such as deep learning [3], machine learning [47], and neural networks [4], to recover ancient texts, especially Greek and Latin texts. As a highly cited author, his research provides a good idea for artificial intelligence to recognize ancient characters.

There were 428 authors in the study, and the authors were usually collaborative. A group will be formed comprising three to five core authors. The group centered on the T. Sommerschield, and the group centered on N. D. Cilia, represent the two largest academic groups in this field, shown in red and green. See Figure 3. The research conducted by N. D. Cilia et al. concentrated on the utilization of deep learning techniques for the identification of digital image of ancient manuscripts [17].

In bibliometrics, Lotka's law and Price's law are frequently employed to ascertain the nucleus of authors within a given field and to delineate the attributes of the research community therein [29].

Lotka's study revealed a correlation between the maturity of a field and the ratio of authors publishing N articles to those publishing one article. Specifically, the number of authors publishing N articles was found to be $1/N^2$ of the number of authors publishing one article. In other words, the number of authors who published two articles accounted for one-quarter of the total number of authors who published one paper. The number of authors who published three papers accounted for one-ninth of the total number of authors who published one paper. The number of authors who published four papers accounted for one-sixteenth of the number of authors who published one paper. What's more, the number of authors who published one paper accounted for 60% of all authors in this field [5].

A total of 428 authors have published papers in the field of artificial intelligence in document protection, according to the statistics of VOSviewer. Of these authors, 32 have published more than one paper, 22 have published two papers, eight have published three papers, and two have published four papers. See Table 2. Following the calculation, it was found that authors who had published one article accounted for 92.5% of the total number of authors. In contrast, authors who had published four articles accounted for only 0.51% of the number of authors who had published one article, which is a proportion significantly below the 6.25% predicted by Lotka's law. Furthermore, the number of authors who had published two articles accounted for 5.5% of the number of authors who had published one article, which is also a proportion significantly below the 25% predicted by Lotka's law. The proportion of authors with one paper who were accounted for by three authors was 2%, a figure that is far less than the 11.1% calculated by Lotka. It can thus be concluded that the publication rule of artificial intelligence application in the field of document protection does not conform to Lotka's law. This indicates that, despite the considerable interest shown by scholars in this field, the formation of a

Table 2
Statistics on the number of published documents.

Number of publications	Number of authors
1	396
2	22
3	89
4	2

sustainable, large-scale scientific research group has yet to occur.

$$N_{min} = 0.749 \times \sqrt{N_{max}} \quad (1)$$

Price's law reveals that a researcher can be repudiated as a core author in the field, mainly by comparing the number of publications with the authors who published the most. See Equation 1. N_{min} represents the minimum number of publications by core authors, and N_{max} represents the maximum number of publications by core authors. The minimum number of publications of the core authors of research utilising artificial intelligence technology in the field of document protection is 1.498. Authors with two or more publications may be considered core authors in this field. In total, these core authors published 76 articles, representing 16% of the total number of articles. This figure is below the 50% threshold that Price suggests should be published by the highly productive author group [56]. This suggests that the field has not yet established a stable core group of authors. The identification of two or more publications as core authors also indicates that the depth of research on the application of artificial intelligence in the field of document protection requires strengthening.

2.2.4. Citation Analysis

In the context of academic research, the term "citation analysis of literature" is used to describe the process of determining the frequency with which an article is referenced by other academic works. A higher number of citations is indicative of a greater importance and status within the field of study. This may be attributed to the article's proposal of a novel concept for future investigation or the fact that it has garnered the attention of renowned scholars in this field.

The larger nodes in the citation analysis diagram indicate a greater frequency of citations of the article in question. See Figure 4. In 2019, Rehman's study on identifying target text images using deep transmission convolutional neural networks (CNN) represents a significant advancement in the field of literature research, exemplifying the utility of deep learning techniques [43]. The 2022 paper by Y. Assael on the use of deep neural networks for the restoration of ancient books and the determination of their attribution attracted greater attention due to its publication in the journal Nature. This paper introduces Ithaca, a neural network for the restoration of ancient Greek inscriptions, which has enhanced the efficiency and accuracy of the study of ancient Greek inscriptions by historians. It constitutes a successful example of the deep integration of digital technology and the humanities [4].

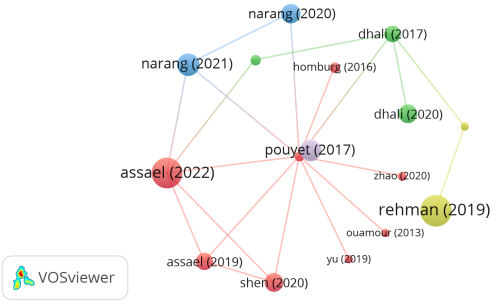

Figure 4: Article citation analysis.

In the context of academic literature, the term "literature coupling analysis" is used to describe the process of identifying common references cited by two or more articles. A greater coupling strength indicates a higher degree of reference duplication between two articles. The literature with high coupling degree is roughly divided into 5 categories. See Figure 5.

Some of the research is based on traditional techniques. In the early stage, people need to manually scan documents to form black and white images, and computers extract image features and classify them to realize text recognition [44]. J. Philips introduced the main stages of historical document processing (HDP), pointing out that ancient documents should be converted into digital formats that facilitate data mining and information retrieval systems [40]. In 2021, based on the practice of optical character recognition (OCR), convolutional neural network (CNN) and other technologies by predecessors, S. R. Narang proposed that CNN technology could be used to identify Sanskrit manuscripts, which is another extended application of convolutional neural network technology [34].

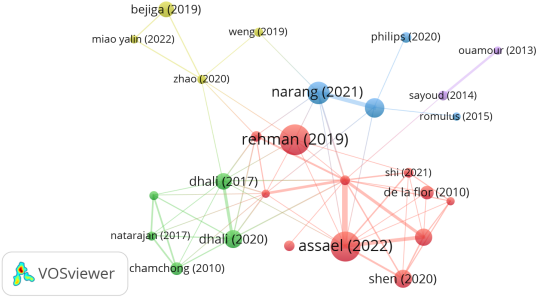

Figure 5: Literature coupling analysis.

Most of the articles in the red area propose new models or ideas that can be used for text or language recognition and cite each other with articles in other color areas, such as digital images [23] and blank language models [45]. At the same time, T. Sommerschield has studied the literature for

nearly 20 years to summarize and demonstrate the feasibility of machine learning in the recognition of ancient characters [48].

The articles in the green area study the content information of the documentary heritage such as author and writing style by means of binarisation [13, 35], pattern recognition [19], and feature extraction [20].

3. Results

The research of artificial intelligence in the field of documentary heritage can be characterized by the following features:

(1) In the study of documentary heritage, scholars have paid attention to the universality of documents. First of all, different document carriers. The research on historical documents not only focuses on paper documents, but also exists in the carrier documents such as oracle bones [51], palm leaf manuscripts [12] and stone carvings. For example, based on deep neural network technology in computer vision, X. R. Fu inserted a pseudo-label into the backbone network to predict inscriptions and identify ancient texts [24]. Secondly, different civilizations and cultures. The recognition of writing involves ancient civilizations on multiple continents, such as the Herculaneum scrolls [7], Chinese ancient books [58], ancient texts [38], Polish ancient manuscripts [16], Akkadian word [26], and so on. These articles show that this field has aroused the attention and attempts of global scholars, and has a large audience and broad prospects for development.

(2) The application of artificial intelligence to the research of documentary heritage is mostly used to read the text or content of documents. For example, ChatGPT is used to read literature and extract information to facilitate the classification and sorting of literature [32]. Designing a multi-task model aimed at deciphering the terminology of etiquette and customs in Chinese ancient books to provide a basis for the extraction of etiquette information in ancient books and the automatic construction of a knowledge base [46]. Lingdan model is based on large language model, extensively integrates ancient Chinese medicine books and clinical data, and makes the model prescribe medicine according to electronic medical records through training [27]. Identify ancient inscriptions and predict damaged characters on artifacts to generate restored images [22]. These studies relate to literature, archaeology, medicine, sociology, and many other fields, and they underscore the significance of leveraging artificial intelligence to facilitate research on the preservation of documentary heritage.

(3) The artificial intelligence technology used in the article reflects the characteristics of The Times and is becoming more and more diverse. These technologies can be broadly divided into three categories. The first category is image recognition and processing technology, such as P. Romulus using optical characters to recognize ancient Batak characters [44], H. Y. Ma used the OCR system to infer the internal sequences of ancient texts [33], and Shuo Yu proposed a generalized multiagent hypergraph-learning framework [57]. The second category is deep learning technology. Deep learning is capable of dealing with complex geometric properties [25], and it is able to automatically extract features from raw data without any prior knowledge [17]. Therefore, many scholars are trying to apply deep learning models to the recognition of basic information of ancient texts and documents. For example, Narang and Sonika Rani used a deep learning model as a feature extractor and classifier to identify 33 types of characters in Sanskrit manuscripts, with an accuracy rate of 93.73% [34]. Akram Bennour developed a deep learning model and, through experimentation, found it to be superior to the state of the art in accurately identifying authors of historical documents. This study demonstrates the potential of deep learning to solve complex problems in document analysis and authorship, and offers new ways for historians to study authors [6].

(4) Nevertheless, the majority of articles concentrate on the implementation of artificial intelligence in digital formats, and there is a paucity of research concerning the protection and utilization of document entities. In fact, the protection of document entities cannot be ignored, in long-term storage, paper, disk, CD and other document carriers are easy to receive dust [41], microorganisms [10, 36], insects, and other erosion. This part of the study may refer to mural protection and restoration. In the early days, researchers restored murals by building collaborative virtual environments [31]. Latter, the seven murals in the Zhao Yigong Tomb, a Tang Dynasty tomb, were integrated into an augmented reality (AR) platform utilizing 2D and 3D modelling in conjunction with AR technology. Thereafter, a usability evaluation of the target audience of the project was conducted, the results of which demonstrated that the AR application was widely accepted [60]. Some researchers have established ultra-high-definition data sets for the Dunhuang murals, which are facing the threat of deterioration. They have preserved the completeness and details of the murals with high-resolution images and introduced a digital restoration framework supporting damage segmentation and digital restoration. This framework is designed to protect, restore, and display the Dunhuang murals by digital means [54]. These are techniques for digitally scanning or modelling entities, which are then protected and restoration in a digital environment. These methods may provide ideas for document protection.

4. Discussion and Conclusion

It is certain that the application of AI in the field of document protection will definitely develop in a diversified direction. Electronic health records are now embedded in health care to predict disease [49]. It is conceivable that in the future, GLAM institutions may draw inspiration from this example and establish a useful and easy-to-use disease database of documentary heritage, predict potential diseases of documentary heritage on a digital platform, and formulate

more comprehensive preservation and protection plans in advance [59].

With regard to the presentation and utilization of documentary heritage, further investigation could be conducted into the reading and recovery of digital documents through graph enhancement techniques [28] and graph convolutional networks. From the perspectives of emotional design [55], digital narration [37], etc., the text in ancient books and documents can be "brought to life".

Acknowledgement

The authors declare that there is no funding and no conflict of interest.

References

[1] Abid, A., 1995. Memory of the world–preserving the documentary heritage. IFLA Journal 21, 169–174.

[2] Al-homed, L.S., Jambi, K.M., Al-Barhamtoshy, H.M., 2023. A deep learning approach for Arabic manuscripts classification. Sensors 23.

[3] Assael, Y., Sommerschield, T., Prag, J., 2019. Restoring ancient text using deep learning: A case study on Greek Epigraphy, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6368–6375.

[4] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., de Freitas, N., 2022. Restoring and attributing ancient texts using deep neural networks. Nature 603, 280.

[5] Bapte, V.D., Gedam, J.S., Bejalwar, S., 2024. Authorship pattern and authorship productivity in library and information science literature. Annals of Library and Information Studies 71, 180–189.

[6] Bennour, A., Boudraa, M., Siddiqi, I., Al-Sarem, M., Al-Shaby, M., Ghabban, F., 2024. A deep learning framework for historical manuscripts writer identification using data-driven features. Multimedia Tools and Applications .

[7] Booras, S.W., Chabries, D.M., 2001. The herculaneum scrolls, in: Proceedings of the Image Processing, Image Quality, Image Capture, Systems Conference, pp. 215–218.

[8] Burckel, N.C., 1992. Review: Using the nations documentary heritage: the report of the historical documents study. The American Archivist 55, 490–492.

[9] Cains, A., 1991. The book of kells–the exhibition and transport of an ancient manuscript, in: Science, Technology and European Cultural Heritage. Butterworth-Heinemann, pp. 338–340.

[10] Cappitelli, F., Sorlini, C., 2005. From papyrus to compact disc: The microbial deterioration of documentary heritage. Critical Reviews in Microbiology 31, 1–10.

[11] Chadha, S., Gupta, N., Anil, B.C., Chauhan, R., 2022. A novel framework for ancient text translation using artificial intelligence. Advances in Distributed Computing and Artificial Intelligence Journal 11, 411–425.

[12] Chamchong, R., Fung, C.C., 2014. A combined method of segmentation for connected handwritten on palm leaf manuscripts, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, IEEE. pp. 4158–4161.

[13] Chamchong, R., Fung, L.C.C., Wong, K.K.W., 2010. Comparing binarisation techniques for the processing of ancient manuscripts, in: Proceedings of the International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation.

[14] Chaovalitwongse, W.A., Yuan, Y., Zhang, Q., Liu, J., 2022. Special issue: Innovative applications of big data and artificial intelligence. Frontiers of Engineering Management 9, 517–519.

[15] Chen, S., Yang, Y., Liu, X., Zhu, S., 2022. Dual discriminator GAN: Restoring ancient Yi characters. ACM Transactions on Asian and Low-Resource Language Information Processing 21.

[16] Choros, K., Jarosz, J., 2018. Most frequent errors in digitization of Polish ancient manuscripts, in: Proceedings of the 10th Asian Conference on Intelligent Information and Database Systems, Springer, Dong Hoi City, Vietnam. pp. 170–179.

[17] Cilia, N.D., De Stefano, C., Fontanella, F., Marrocco, C., Molinara, M., Freca, A.S.d., 2020. An experimental comparison between deep learning and classical machine learning approaches for writer identification in medieval documents. Journal of Imaging 6.

[18] Corazza, M., 2022. Unsupervised deep learning for ancient Aegean scripts: From deciphered to undeciphered. Lingue E Linguaggio 21, 311–331.

[19] Dhali, M.A., He, S., Popovic, M., Tigchelaar, E., Schomaker, L., 2017. A digital palaeographic approach towards writer identification in the Dead Sea scrolls, in: Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, Porto, Portugal. pp. 693–702.

[20] Dhali, M.A., Jansen, C.N., de Wit, J.W., Schomaker, L., 2020. Feature-extraction methods for historical manuscript dating based on writing style development. Pattern Recognition Letters 131, 413–420.

[21] Dorneanu, B., Zhang, S., Ruan, H., Heshmat, M., Chen, R., Vassiliadis, V.S., Arellano-Garcia, H., 2022. Big data and machine learning: A roadmap towards smart plants. Frontiers of Engineering Management 9, 623–639.

[22] Duan, S., Wang, J., Su, Q., 2024. Restoring ancient ideograph: A multimodal multitask neural network approach.

[23] de la Flor, G., Luff, P., Jirotka, M., Pybus, J., Kirkham, R., Carusi, A., 2010. The case of the disappearing ox: Seeing through digital images to an analysis of ancient texts, in: Proceedings of the 28th Annual CHI Conference On Human Factors In Computing Systems, ACM, Atlanta, GA, USA. p. 473.

[24] Fu, X., Zhou, R., Yang, X., Li, C., 2024. Detecting oracle bone inscriptions via pseudo-category labels. Heritage Science 12.

[25] Girbacia, F., 2024. An analysis of research trends for using artificial intelligence in cultural heritage. Electronics 13.

[26] Homburg, T., Chiarcos, C., 2016. Akkadian word segmentation, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, pp. 4067–4074.

[27] Hua, R., Dong, X., Wei, Y., Shu, Z., Yang, P., Hu, Y., Zhou, S., Sun, H., Yan, K., Yan, X., Chang, K., Li, X., Bai, Y., Zhang, R., Wang, W., Zhou, X., 2024. Lingdan: Enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models. Journal of the American Medical Informatics Association 31, 2019–2029.

[28] Jin, S., Chen, Z., Yu, S., Altaf, M., Ma, Z., 2023. Self-augmentation graph contrastive learning for multi-view attribute graph clustering, in: Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering, Association for Computing Machinery, Ottawa, ON, Canada. pp. 51–56.

[29] Kastrin, A., Hristovski, D., 2021. Scientometric analysis and knowledge mapping of literature-based discovery (1986-2020). Scientometrics 126, 1415–1451.

[30] Kogkitsidou, E., Gambette, P., 2020. Normalisation of 16th and 17th century texts in French and geographical named entity recognition, in: Proceedings of the 4th ACM Sigspatial International Workshop on Geospatial Humanities, pp. 28–34.

[31] Li, X., Lu, D., Pan, Y., 2000. Virtual Dunhuang mural restoration system in collaborative network environment. Computer Graphics Forum 19, C331.

[32] Lin, D., Zou, R., 2024. Applications, risk, challenges, and future prospects of ChatGPT in electronic records management. Journal of Artificial Intelligence Research 1, 1–9.

[33] Ma, H., Huang, H., Liu, C., 2024. Reading between the lines: Image-based order detection in OCR for Chinese historical documents, in: Proceedings of the 38th AAAI Conference on Artificial Intelligence, AAAI Press, Vancouver, Canada. pp. 23808–23810.

[34] Narang, S.R., Kumar, M., Jindal, M.K., 2021. DeepNetDevanagari: A deep learning model for Devanagari ancient character recognition. Multimedia Tools and Applications 80, 20671–20686.

[35] Natarajan, J., Sreedevi, I., 2017. Enhancement of ancient manuscript images by log based binarization technique. AEU-International Journal of Electronics and Communications 75, 15–22.

[36] Nitiu, D.S., Mallo, A.C., Saparrat, M.C.N., 2022. Pigments synthesized by dark fungi and their impact on the deterioration of documentary heritage on paper. Boletin De La Sociedad Argentina De Botanica 57, 169–184.

[37] Niu, L., Zeng, J., Wu, F., Wang, K., 2023. Digital storytelling: A new opportunity for the archival documentary heritage of Suzhou silk. Library Trends 71.

[38] Ouamour, S., Sayoud, H., 2013. Authorship attribution of ancient texts written by ten Arabic travelers using character N-Grams, in: Proceedings of the International Conference on Computer, Information and Telecommunication Systems, IEEE.

[39] Pennacchiotti, M., Zanzotto, F.M., 2008. Natural language processing across time: An empirical investigation on Italian, in: Proceedings of the Advances in Natural Language Processing, p. 371.

[40] Philips, J., Tabrizi, N., 2020. Historical document processing: A survey of techniques, tools, and trends, in: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 341–349.

[41] Prajapati, C.L., 2003. Accumulation of solid particles on documents, a threat for preservation of documentary heritage: The example of the national archives of India. Restaurator: International Journal for the Preservation of Library and Archival Material 24, 46–54.

[42] Qu, F.G., Dong, Y.H., 2005. The study on the city sustainable development, in: Proceedings of the International Conference on Management Science and Engineering, pp. 303–308.

[43] Rehman, A., Naz, S., Razzak, M.I., Hameed, I.A., 2019. Automatic visual features for writer identification: A deep learning approach. IEEE Access 7, 17149–17157.

[44] Romulus, P., Maraden, Y., Purnamasari, P.D., Ratna, A.A.P., 2015. An analysis of optical character recognition implementation for ancient Batak characters using K-nearest neighbors principle, in: Proceedings of the International Conference on Quality in Research, IEEE. pp. 47–50.

[45] Shen, T., Quach, V., Barzilay, R., Jaakkola, T., 2020. Blank language models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Assoc Computat Linguist. pp. 5186–5198.

[46] Siriguleng, Lin, M., Guo, Z., Shujun, Z., Li, B., Gao, Y., 2024. Multi-task learning for ancient ritual literature etiquette entity recognition. Data Analysis and Knowledge Discovery 8, 56–66.

[47] Sommerschield, T., 2020. Restoring ancient text using machine learning: A case-study on Greek and Latin epigraphy. Papers of the British School at Rome 88, 387–388.

[48] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., de Freitas, N., 2023. Machine learning for ancient languages: A survey. Computational Linguistics 49, 703–747.

[49] Tang, T., Han, Z., Cai, Z., Yu, S., Zhou, X., Oseni, T., Das, S.K., 2024. Personalized federated graph learning on non-IID electronic health records. IEEE Transactions on Neural Networks and Learning Systems 35, 11843–11856.

[50] Wang, L., Wang, J., Wei, T., 2023. Modeling Chinese ancient book catalog, in: Proceedings of the 16th International Conference on Knowledge Science, Engineering and Management, Guangzhou, China. pp. 355–367.

[51] Wang, P., Zhang, K., Wang, X., Han, S., Liu, Y., Wan, J., Guan, H., Kuang, Z., Jin, L., Bai, X., Liu, Y., 2024. An open dataset for oracle bone character recognition and decipherment. Scientific Data 11.

[52] Wang, Z., Wang, Y., 2024. Digital library book recommendation system based on tag mining. Journal of Artificial Intelligence Research 1, 10–16.

[53] Xia, F., Chen, X., Yu, S., Hou, M., Liu, M., You, L., 2024. Coupled attention networks for multivariate time series anomaly detection. IEEE Transactions on Emerging Topics in Computing 12, 240–253.

[54] Xu, Z., Yang, Y., Fang, Q., Chen, W., Xu, T., Liu, J., Wang, Z., 2024. A comprehensive dataset for digital restoration of Dunhuang murals. Scientific Data 11.

[55] Yan, H.B., Li, Z., 2022. Review of sentiment analysis: An emotional product development view. Frontiers of Engineering Management 9, 592–609.

[56] Yang, W., Shen, L., Huang, C.F., Lee, J., Zhao, X., 2024. Development status, frontier hotspots, and technical evaluations in the field of AI music composition since the 21st century: A systematic review. IEEE Access 12, 89452–89466.

[57] Yu, S., Huang, H., Shen, Y., Wang, P., Zhang, Q., Sun, K., Chen, H., 2024. Formulating and representing multiagent systems with hypergraphs. IEEE Transactions on Neural Networks and Learning Systems .

[58] Zhang, M., Ma, S.P., Jiang, Z., Huang, K., 2001. Statistical learning and analyses on Chinese ancient books for information retrieval, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE, Tucson, AZ, USA. pp. 869–873.

[59] Zhang, M., Qin, C., Qiang, F., 2024. Leveraging artificial intelligence to assess physicians' willingness to share electronic medical records in a hierarchical diagnostic ecosystem. Journal of Artificial Intelligence Research 1, 27–35.

[60] Zheng, S., 2024. Intangible heritage restoration of damaged tomb murals through augmented reality technology: A case study of Zhao Yigong Tomb murals in Tang Dynasty of China. Journal of Cultural Heritage 69, 135–147.