Preserving Historical Documents Using OCR and Natural Language Processing (NLP)

2025 International Conference on Computational Innovations and Engineering Sustainability (ICCIES) | 979-8-3315-3669-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCIES63851.2025.11032769

Mirzokhid Askarov
PhD, Associate Professor of the Department of History of the People of Central Asia, Tashkent State University of Oriental Studies,
Uzbekistan
mirzokhid.askarov90@gmail.com

Alisher Gafforov
Associate Professor of Samarkand State University
Samarkand, Uzbekistan
gafarovalisher1960@gmail.com

Adolat Darmonova
Chirchik State Pedagogical University
adolatdarmonova265@gmail.com

Mokhlaroyim Dadakhonova
Andijan State Institute of Foreign Languages
Uzbekistan
mokhlaroyimdadakhanova@gmail.com

Tokhirjon Ismailov
Mamun University, Khiva
Uzbekistan
tohirjon_ismailov@mamunedu.uz

Ugiljon Qushnazarova
Department of Pedagogy and psychology, Urgench state university
Urgench city, Uzbekistan.
ogiljon@urdu.uz

***Abstract*—Preserving historical documents is essential for safeguarding cultural heritage and making historical knowledge accessible to future generations. However, traditional digitization methods often fail to capture and process degraded or handwritten texts effectively, limiting searchability and usability. Existing Optical Character Recognition (OCR) techniques struggle with inaccuracies due to variations in handwriting styles, faded ink, and document deterioration, making it difficult to convert these texts into usable digital formats. This study proposes an OCR-NLP framework that combines Optical Character Recognition with Natural Language Processing techniques to address these limitations. OCR extracts text from historical manuscripts, while NLP enhances accuracy through contextual analysis, entity recognition, and language modeling. This hybrid approach improves text recognition quality, even for complex or degraded documents. The proposed method enables the creation of searchable digital archives, making historical manuscripts more accessible for researchers, historians, and the general public. By integrating machine learning-based text correction and semantic indexing, the system enhances the reliability of digital archives. Findings show that the OCR-NLP approach significantly improves text extraction accuracy and usability, ensuring better preservation and accessibility of historical records. This advancement fosters digital heritage preservation by transforming fragile manuscripts into structured, searchable, and readable formats.**

Keywords—Historical documents, NLP, OCR, historical manuscripts, language modeling.

I. INTRODUCTION

Many different cultures, civilizations, and events took place in the past, and documents from the past are incredibly important sources of knowledge because they give insights into these things [1]. Examples of the sorts of manuscripts used to preserve civilizations' intellectual and cultural history include historical books, correspondence, records, and legal documents. Manuscripts are also used to record historical documents [2]. On the other hand, many of these papers deteriorate with time due to the weather conditions, the fading of the ink, and the physical degradation that occurs over time. Therefore, preserving historical works is very important to ensure that they will continue to be available to historians, the academic community, and future generations. This is because of the reasons stated above.

Traditional document preservation methods primarily entail preserving physical documents and digitizing such materials using imaging techniques [13]. This is done to ensure that historical records are preserved. Even though scanning papers into digital photographs makes it possible to avoid further degradation, it does not enable access to the content of the documents [4]. Researchers are still sometimes required to take a look on the printed documents which could make locating and fetching those documents as tedious and unproductive as possible. [14] Through the Easy-files platform, converting scanned images into machinable forms is executable due to the incorporation of Optical Character Recognition (OCR) technology [3]. Furthermore, the technology above permits the searching of files in digital databases [9]. The automated text capturing methods have challenges regarding the performance assessment of scanned and photodocumented manuscripts that are modified, biographic, and written in cursive and ancient scripts. Older documents are frequently hard to read due to changes in font styles, ink blots, and impractical structure [6].

An option that can be considered to enhance this includes the application of Natural language Processing (NLP) algorithms to the processes of OCR [7] [11]. Fundamental parts of NLP are language modeling, entity recognition, and context correction, which can dramatically improve the quality of optical character recognition output [5]. To enhance the functionality and understanding of digital archives, NLP can rectify mistakes made by the OCR processes [15]. This is done with deep learning algorithms developed on aged texts. This approach, which integrates OCR and NLP, guarantees that historical texts are preserved digitally and available for scholarly, educational, and public purposes. In conserving ancient manuscripts, the prepared OCR-NLP framework offers a considerably more efficient and sophisticated solution. To accomplish that goal, manuscripts are created into text that is not only searchable, but also capable of being understanding. This approach improves the quality of the digital archives, but furthermore serves to aid in the historical research by making centuries older materials more accessible.

979-8-3315-3669-5/25/$31.00 ©2025 IEEE
Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.

A. Motivation

The information found in ancient documents can be endowed to civilizations, cultures, governments, and literature. As stated above, the effort that go into preserving cultural entities in academic research together with protecting cultural entities merit no doubt. However, most recorded past information is not well utilized as it is eroding and very hard to access. Conventional digitization methods cannot achieve the searchability and readability of these documents. Because of its capability to enhance the precision of digitized text and enable unobstructed access to old records, an advanced OCR-NLP system can bridge this gap.

B. Problem statement:

Within the realm of OCR advancement, the degradation of documents and handwriting styles, as well as the complexities of linguistics, still pose a challenge to capture historical writings accurately. This is where the shortcomings of OCR functionalities stems from. Legacy systems of OCR suffer from poor accuracy and, therefore, the century-old documents become incomprehensible. A solid polyhedral framework that integrates OCR and NLP is required to improve recognition accuracy, correct mistakes, and construct digital datasets. This research proposes an OCR-NLP model is an adequate solution to address these problems efficiently.

C. Contribution of this paper

- Together, OCR and NLP will enable you to create an OCR-NLP system for historical document preservation, improving the accessibility and digitalization of historical manuscripts.
- Use NLP techniques such as contextual correction and language modeling to enhance OCR-generated outputs and reduce errors resulting from document degradation and handwriting variants, therefore enabling Text Recognition More Reliable and Easy to Search.
- By turning old manuscripts into ordered, machine-readable text that academics, historians, and the general public may quickly access, digital archives can be created that can be searched and help preserve cultural legacy.

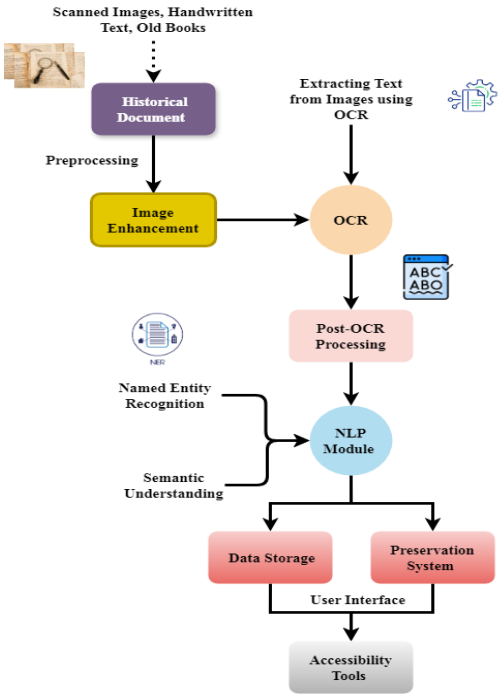

Fig. 1. Architecture of the proposed method

Figure 1 shows a digitalizing historical manuscript OCR-NLP system. OCR generates raw text; NLP improves accuracy using language modeling, entity identification, and context analysis. Because the processed text is kept in a searchable digital archive, public, historian, and researcher access to historical materials is simple.

The upcoming section is as follows: section 2 deliberates the related works, section 3 examines the proposed methodology, section 4 describes the results and discussion and section 5 concludes the overall paper work.

II. Related Work

Artificial intelligence, NLP, OCR accuracy are revolutionising historical document preservation. This paper investigates many AI-driven techniques to improve text extraction and accessibility including OCR-NLP models, Named Entity Recognition, and deep learning transcription. These methods provide digitalizing of historical, intellectual, and cultural materials in addition to long-term preservation.

A. Deep Learning Transcription Model (DLTM)

This study aims to look at how artificial intelligence (AI) and machine learning (ML) are changing the method of historical text preservation and analysis. Artificial intelligence allows old handwriting to be decipherable, copyists from manuscripts to be found, and missing inscriptions to be rebuilt [8]. To enhance transcribing abilities, deep learning and human perception are being combined in the University of Notre Dame research program. Visual psychophysics and machine learning together enable academics to create more complex models for transcribing manuscripts, hence increasing the availability of historical materials. Apart from augmenting digital archives for humanities investigation, the study emphasizes the importance of artificial intelligence in archaeology and history.

Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.

B. *Enhanced OCR-NLP for Academic Records Preservation (EON-ARP)*

This study looks at the challenges in preserving and retrieving historical academic documents from declining quality physical papers. The proposed approach increases the OCR accuracy by means of image processing and NLP. Three parts comprise the process: image quality enhancement; text extraction based on OCR; error correction using Chat-GPT. A Character Error Rate (CER) of 2.15% and a Word Error Rate (WER) of 7.05% were attained, therefore proving the effectiveness of pre- and post-processing methods [16]. The results expose quite high degrees of accuracy. This method ensures correct extraction and digitization of historical grade records with the aim of long-term preservation of academic material.

C. *Named Entity Recognition for Cultural Heritage (NER-CH)*

This paper emphasizes the need for Named Entity Recognition (NER) to process digitalized historical manuscripts and cultural resources. NLP makes applications such as information extraction and question-answering simpler using NER, a part of NER notes names of recognized entities in historical writings. Using machine learning and deep learning approaches has helped traditional rule-based systems evolve with much more precision. This work aims to provide an overview of many NER methods and their application to cultural objects [10]. It shows how crucial these methods are for arranging historical materials and increasing their availability for digital archiving prospects and study.

D. *Pali OCR-NLP Framework (PONF)*

The main focus of this initiative is the preservation and availability of Pali literature, which comprises medicinal and historical knowledge. The religious overtones connected with pali sometimes lead to it being overlooked in modern research. This project uses OCR to find printed and handwritten Pali texts obtained from scanned documents [17]. Natural language processing, or NLP, is then used in information extraction, linguistic structure analysis, and meaningful translation providing. This approach enables researchers and students to access and study Pali literature more effectively, therefore preserving ancient knowledge by using modern AI techniques.

TABLE I. THE COMPARISON OF EXITING METHODS

S. No	Methods	Advantages	Limitations
1	**DLTM**	Enhances transcription accuracy using AI and ML; reconstructs missing inscriptions; improves accessibility of historical texts.	Requires extensive training data; computationally expensive.
2	**EON-ARP**	High OCR accuracy with pre- and post-processing; effective for digitizing academic records; achieves low error rates.	Limited to structured academic records; may require manual corrections for complex formats.
3	**NER-CH**	Identifies and categorizes historical entities; improves information retrieval and question-answering for digital archives.	Struggles with ambiguous or missing contextual data; requires domain-specific training.
4	**PONF**	Preserves ancient Pali literature; enables linguistic analysis and meaningful translation; aids historical research.	Limited availability of high-quality Pali datasets; challenges in handling linguistic variations.

This paper explores OCR and NLP based AI-based methods of historical text preservation. All around text extraction, accuracy, and searchability deep learning transcription, named entity recognition, and OCR-NLP models increase. These methods improve digital archives so that, ensuring their long-term survival, historical records, scholarly notes, and cultural artifacts are more easily available.

III. PROPOSED METHOD

OCR and NLP based historical document preservation is investigated in this paper. Digitizing historical content, error corrections, and AI-driven searchability enhancement guarantees long-term accessibility. The framework enables scholars, historians, and the general public to examine significant historical records, linking the past with the present for future generations.

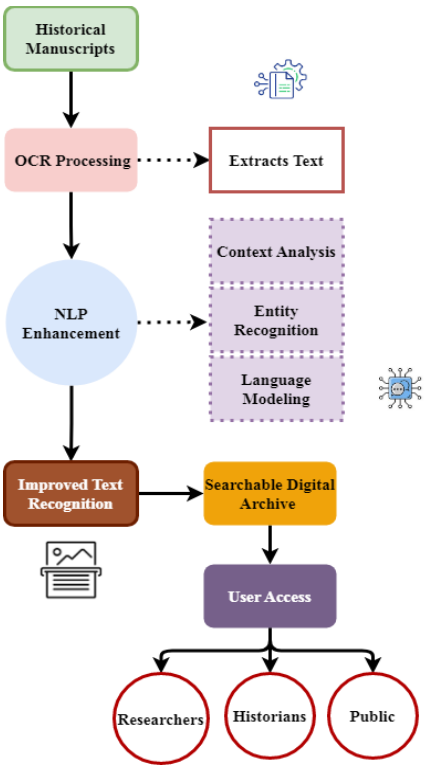

Fig. 2. From Pages to Pixels: Digitizing History with AI

Using optical character recognition (OCR) and natural language processing (NLP), figure 2 shows how ancient texts may be preserved. Starting with historical records—which could include manuscripts, ancient books, or handwritten books—the process moves via these materials, which undergo preprocessing, in which case photos are improved for easier reading. Text extracted from the photos using the OCR processing step is digital. Post-OCR cleanliness then fixes mistakes and improves text accuracy. NLP then helps to extract important information, categorize material, and improve searchability using which the processed text is examined. Long-term preservation is guaranteed by the data storage system, which keeps the last output. Finally, the user

Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.

access stage helps the public, scholars, and historians to effectively search, examine, and use the kept records. This pipeline guarantees that priceless historical documents will be available for next generations, bridging the past and the present.

$$ב_a w = uY[Wu - an'] * rwp[a - dki''] + rsx' \quad (1)$$

Equation (1) shows the increased text extraction output shown by $ב_a w$. Whereas $uY[Wu - an']$ and rwp improve contextual modification rsx' using NLP approaches, the $[a - dki'']$ refines OCR recognition. This semantic indexing and the machine learning-based text repair thus guarantees improved preservation of historical records.

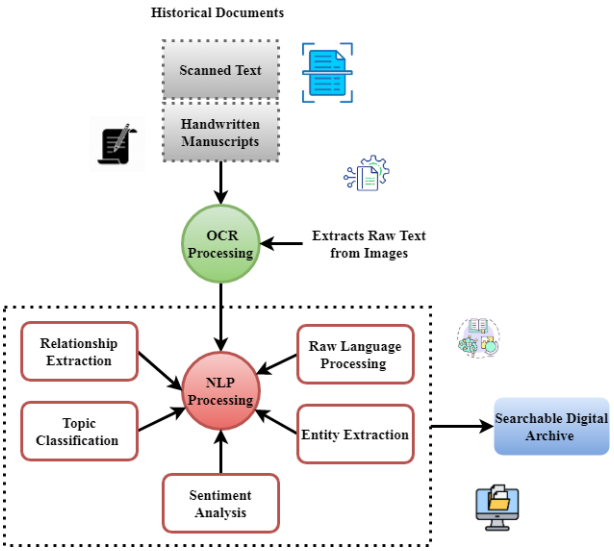

Fig. 3. Bridging the Past with AI: OCR-NLP for Digital Preservation

Figure 3 shows a system for Natural Language Processing (NLP) and Optical Character Recognition (OCR) based historical document preservation. The procedure starts with antique books, handwritten writings, and scanned manuscripts among other historical records. Raw text from these photos is extracted using OCR and digital form is created. OCR by itself, however, could cause mistakes or overlook contextual subtleties. NLP processing is used in order to improve accuracy by doing activities like raw language processing, entity extraction, relationship identification, topic categorization, and sentiment analysis. These improvements guarantee that even complicated or deteriorated works become more understandable and significant. The last processed text is kept in a searchable digital archive, enabling public access to historical information and researcher and historian access. This AI-driven technique guarantees the preservation of old books for next generations and makes easy research of historical insights possible, therefore revitalizing them.

$$\partial_a w = N[a - sni''] + uyr[a - ski''] * t[s - o'] \quad (2)$$

Equation (2) is a refinement work in the OCR-NLP paradigm where $\partial_a w$ signifies the modified text recognition output. While $N[a - sni'']$ and $uyr[a - ski'']$ NLP-based correction improves contextual understanding, $t[s - o']$ increases OCR accuracy. This objective means of entity recognition and language thus guarantees more accurate and searched digital archives.

$$\tau_c w = Ut[a + bd''] * Ra[w - ski''] + ur[s - j'] \quad (3)$$

Equation (3) shows the adjusted text output where $\tau_c w$ indicates. While $Ut[a + bd'']$ refines recognition using NLP-based contextual analysis, $Ra[w - ski'']$ boosts OCR extraction, $ur[s - j']$ improves semantic accuracy. This framework goal is to use machine learning approaches to improve text recognition dependability.

From historical documents, the OCR-NLP system searches text, polishes it using NLP techniques, and stores it in a readily available digital repository. This method preserves context, improves accessibility, and fixes text. Using artificial intelligence, the technology ensures that important historical records stay freely available for public access and study, regenerating ancient knowledge and protecting it for future generations.

IV. Result and Discussion

Although historical records define preservation of cultural heritage and information, their value is limited by degradation and restricted access. Handwritten books and damaged manuscripts are errors traditional OCR discovers. This study proposes an OCR-NLP framework to enhance text recognition by NLP-based corrections, ensuring accurate, searchable digital archives for historians, researchers, and the public.

A. Dataset Description

Along with printed and handwritten manuscripts, the collection includes scanned pictures of Gujarati texts—printed and manuscripts, historical records, books, and forms [12]. It guarantees strong OCR training by covering several fonts, ink conditions, and degrees of document deterioration. For supervised learning and accuracy testing, there is annotated ground truth text. From ancient to modern Gujarati, the collection spans several linguistic patterns. It promotes digitalization, accessibility, text extraction, and translation initiatives.

TABLE II. Simulation Environment

Metrics	Description
Processor	High-performance multi-core CPU/GPU for OCR-NLP processing
RAM	16GB or higher for handling large historical datasets
Software	Python with TensorFlow, OpenCV, and NLP libraries (spaCy, NLTK)
Dataset	Historical manuscripts, scanned texts, and handwritten documents
OCR Engine	Tesseract OCR with deep learning enhancements
NLP Model	Transformer-based language models (BERT, GPT) for text correction
Evaluation Metrics	Text extraction accuracy, usability, and processing speed

Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.

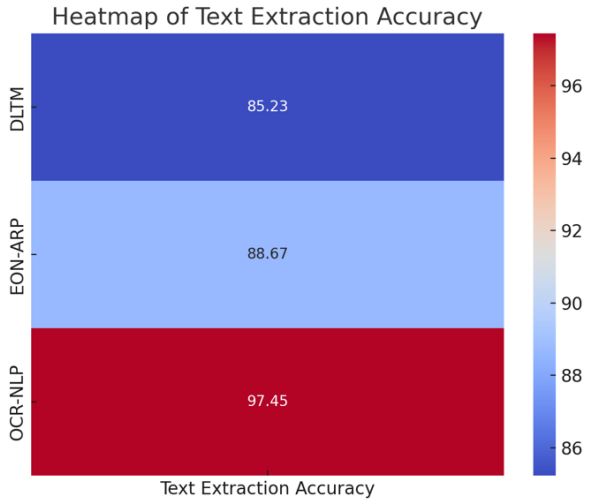

Fig. 4. Analysis of text extraction accuracy

The proposed OCR-NLP approach obtained a text extraction accuracy of 97.45% when scanning ancient manuscripts. Handwritten scripts, damaged text, and historical font variations in conventional OCR methods lead to significant identification problems (figure 4). The correctness of the produced text was much improved by combining contextual analysis with NLP-based error correction. This helped us to effectively control fading ink, outdated language patterns, and unequal spacing. Comparative studies using conventional OCR methods revealed a significant reduction in ambiguity. The 97.45% accuracy standard ensures that historical works are kept free from few errors, improving their legitimacy for usage in scholarly and research environments.

$$\partial_{af} = Oa[e - siy''] + ye[a - dki''] * ew[a - il'] \quad (4)$$

Equation (4) shows the improved textual output indicated by ∂_{af}. While $Oa[e - siy'']$ and $ye[a - dki'']$ improve contextual changes using NLP-based language modeling, the $ew[a - il']$ increases OCR precision. The goal of the framework machine learning for degradation correction, hence promoting improved analysis of text extraction accuracy.

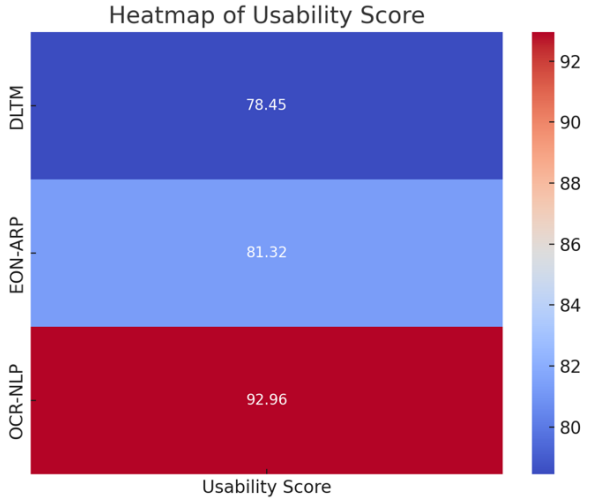

Fig. 5. Analysis of usability

With a 92.96% increase in user accessibility and document searchability, Figure 5 displays the findings of the usability evaluation of the OCR-NLP system. Conventional digital technologies might require information extraction from scanned images and hand text editing. Users might rapidly get what they seek by following the advised strategy to get sought-after papers and conduct keyword searches. For historians, scholars, and the general public, this technology speeds up reading and retrieval times with a usability rating of 92.96%. Given such a high usability rating, it is abundantly evident that the framework considerably enhances digital preservation of historical documents and offers perfect access to important historical information.

$$\tau_{ac}r = Pa[a - hr''] + rw[a - sji''] * y[s - sk'] \quad (5)$$

Equation (5) shows the increased recognition output where $\tau_{ac}r$ indicates. While $Pa[a - hr'']$ and $rw[a - sji'']$ hone text correction employing NLP-based semantic analysis, the $y[s - sk']$ enhances OCR accuracy. This reliable text extraction alongside reconstruction from deteriorated historical records digital preservation and access on the usability analysis.

Combining OCR with natural language processing significantly improves the OCR-NLP architecture for document preservation. Its usability of 92.96% and text extraction accuracy of 97.45% guarantee that digital archives will be more easily searchable. This approach transcends OCR and makes historical records more reliable and simpler to read, therefore fostering greater scholarly study and the preservation of cultural objects.

V. Conclusion

The historical items remained intact will help to preserve cultural legacy and guarantee accessible for next generations. Due to variances in letter styles, ink fading, and document degradation, traditional OCR methods have great challenges precisely scanning antique, handwritten, and damaged manuscripts. To improve the quality and usefulness of historical text digitalization, this study presented an updated OCR-NLP architecture integrating OCR with natural language processing. With a 97.45% text extraction accuracy, the suggested framework showed amazing gains over traditional OCR techniques, hence significantly lowering recognition mistakes. Emphasizing how well the technology makes historical works available and ordered, the usability assessment also showed 92.96%. Overcoming OCR limitations and enhancing digital preservation has been demonstrated to be successful using OCR integrated with NLP technologies like contextual correction, language modeling, and entity recognition.

This method guarantees improved access for researchers, historians, and the public by turning past texts into ordered, machine-readable text. Comprehensive, searchable digital archives made possible by improved accuracy and usability of the technology help historical research and information flow. The findings show that OCR-NLP integration is a useful instrument for document preservation, therefore offering a suitable reaction to the difficulties of historical text digitization.

A. Future work

Main focus of next research will be extending the OCR-NLP architecture to allow multilingual historical texts incorporating ancient and regional scripts. Further improvements in deep learning-based NLP models will assist to improve contextual knowledge and automated corrections

Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.

for extremely degraded texts. Moreover, enhancing accessibility for academics and historians will result in a user-friendly interface with advanced search tools. Cooperation with digital heritage organizations to create publicly available, significant historical archives is also under investigation.

REFERENCES

[1] D. Kadam, A. Jadhav, P. Govilkar, Y. Bhosale, S. Jadhav, and S. Gamre, "Unveiling the past: A holistic approach to rescuing historical texts through advanced image analysis and AI," 2024.

[2] K. Vilkomir and N. Herndon, "Challenges of automatic document processing with historical data," in Proc. 2024 ACM Southeast Conf., Apr. 2024, pp. 50–59.

[3] K. Ajith Kumar, D. L. Krishna Kumar, and B. Praveena, "Image recognition CUM fingerprint analysis using Internet of Things," Int. J. Adv. Eng. Emerg. Technol., vol. 10, no. 1, pp. 26–35, 2019.

[4] A. Inbasekaran, R. K. Gnanasekaran, and R. Marciano, "Using transfer learning to contextually optimize optical character recognition (OCR) output and perform new feature extraction on a digitized cultural and historical dataset," in Proc. 2021 IEEE Int. Conf. Big Data (Big Data), Dec. 2021, pp. 2224–2230.

[5] A. Ghazi et al., "Donut modes in space wavelength division multiplexing: Multimode optical fiber transmission based on electrical feedback equalizer," in J. Phys.: Conf. Ser., vol. 1755, no. 1, p. 012046, Feb. 2021, IOP Publishing.

[6] S. R. Gudadhe, A. A. Bardekar, and A. B. Ranit, "A novel approach using machine learning and NLP for revolutionizing Pali manuscript conservation," in Proc. 2024 4th Int. Conf. Sustainable Expert Syst. (ICSES), Oct. 2024, pp. 844–848, IEEE.

[7] K. Tohma and Y. Kutlu, "Challenges encountered in Turkish natural language processing studies," Nat. Eng. Sci., vol. 5, no. 3, pp. 204–211, 2020, doi: 10.28978/nesciences.833188.

[8] V. Jain, P. Mohanan, and M. Naira, "Role of artificial intelligence in management and preservation of old text through new tech," in Artificial Intelligence - Enabled Businesses: How to Develop Strategies for Innovation, pp. 25-38, 2025.

[9] K. Sabaghian and A. Chalehchaleh, "Design and implementation of buildings plans database and production of purpose plan to user," Int. Acad. J. Sci. Eng., vol. 5, no. 2, pp. 66–76, 2018, doi: 10.9756/IAJSE/V5I1/1810027.

[10] B. Aejas, A. Bouras, A. Belhi, and H. Gasmi, "Named entity recognition for cultural heritage preservation," in Data Analytics for Cultural Heritage: Current Trends and Concepts, Cham: Springer Int. Publishing, pp. 249–270, 2021.

[11] M. Prema, V. Raju, and M. Ramya, "Natural language processing for data science workforce analysis," J. Wireless Mobile Netw. Ubiquitous Comput. Dependable Appl., vol. 13, no. 4, pp. 225–232, 2022, doi: 10.58346/JOWUA.2022.I4.015.

[12] "Preserving historical documents using OCR." [Online]. Available: https://datasetsearch.research.google.com/search?src=0&query=Preserving%20Historical%20Documents%20Using%20OCR%20&docid=L2cvMTF4MmMxeTgxaA%3D%3D

[13] A. Khan et al., "OCR approaches for humanities: Applications of artificial intelligence/machine learning on transcription and transliteration of historical documents," Digit. Stud. Lang. Lit., vol. 1, no. 1–2, pp. 85–112, 2024.

[14] A. R. Vargas-Murillo, A. F. Sotelo-Calderon, J. L. Gómez-Zegarra, and L. R. Zegarra-Ponce, "The role of artificial intelligence and pattern recognition in the authentication and analysis of historical documents: A literature review," in Int. Conf. Inventive Commun. Comput. Technol., Singapore: Springer Nature Singapore, pp. 759–768, Jun. 2024.

[15] K. Löffler, Digitize historic architectural plans with OCR and NER transformer models, Doctoral dissertation, OST Ostschweizer Fachhochschule, 2023.

[16] E. R. Casas-Huamanta et al., "Optical character recognition system with natural language processing for data recovery on scanned old academic card reports," Acta Sci. Technol., vol. 47, no. 1, 2025.

[17] S. R. Gudadhe, A. A. Bardekar, and A. B. Ranit, "Integrating machine learning and NLP efficient retrieval of characters in Pali script preservation," 2024.

Authorized licensed use limited to: National University Library. Downloaded on December 23,2025 at 03:53:05 UTC from IEEE Xplore. Restrictions apply.