International Journal of Linguistics Applied Psychology and Technology
https://ijlapt.strjournals.com/index.php/ijlapt
vol. 02 issue 03 (2025)
ISSN: 3048-4529

AI-Based OCR for Digitizing Ancient Indian Texts: Preserving Linguistic Heritage and Overcoming Script Challenges

Shivraj Gaikwad[1], Renu Kachhoria[2], Gitanjali yadav[3]

[1]Undergraduate student (AI & DS), Vishwakarma Institute of Information Technology, Pune - 411048, India, E-mail: shivrajgaikwad91@gmail.com

[2]Associate Professor (AI & DS), Vishwakarma Institute of Information Technology, Pune - 411048, India, E-mail: renu.kachhoria@viit.ac.in

[3]Associate Professor (AI&DS), Vishwakarma Institute of Information Technology, Pune - 411048, India, E-mail: gitanjali.yadav@viit.ac.in

Abstract

To preserve India's language and culture, traditional literature must be preserved. Due to linguistic differences, script degradation, and a lack of digital copies, many antique manuscripts are unintelligible. This effort digitizes ancient Indian manuscripts using AI-OCR. The initiative dissolves texts, intricate language systems, and scripts. AI-powered optical character recognition (OCR) systems improve Devanagari, Tamil, Grantha, and Brahmi text recognition using deep learning models and NLP. AI-driven optical character recognition (OCR) with extensive pre-processing, character segmentation and language modelling is used to solve recognition problems in the study. According to this program, AI may help indigenous knowledge, language studies, and classic literature. AI is essential for cultural preservation and ancient text studies, the study found.

Keywords: *AI-based OCR, ancient Indian texts, script digitization, linguistic heritage, deep learning, natural language processing, manuscript preservation, historical texts, script recognition, cultural preservation.*

1. INTRODUCTION

Ancient manuscripts in Devanagari, Tamil, Grantha, Brahmi, and others have preserved India's diverse linguistic heritage. The manuscripts have survived generations. Physical degradation, script intricacy, and language development hinder old text preservation and accessibility. Traditional digitization processes may not recognize and understand these scripts, thus more advanced technology is needed. OCR powered by AI is a revolutionary solution to these limits. Ancient manuscripts are digitized using deep learning, NLP, and machine learning. Researchers can improve script identification by combining AI-driven optical character recognition (OCR) with pre-processing, character segmentation, and contextual language modeling. This project examines how artificial intelligence-based optical character recognition can protect India's extensive linguistic heritage, overcome the technological challenges of

digitizing ancient manuscripts, and improve accessibility for scholars, researchers, and the public.

Objectives

- To examine how the complexity of historical records, language discrepancies, and script degradation complicate the digitization of ancient Indian literature.
- To assess the identification and processing capacities of Devanagari, Tamil, Grantha, and Brahmi scripts via AI-driven OCR.
- To examine how NLP and deep learning might enhance the precision of OCR for deteriorating manuscripts and intricate ancient scripts.
- To evaluate the influence of AI-driven OCR on scholarship, accessibility, and the conservation of ancient Indian writings.
- To propose improvements for AI-driven OCR models aimed at cultural preservation, linguistic precision, and script identification.

2. LITERATURE REVIEW

Nandhini Pradeep et al. (2021) suggest digitizing Tamil Palm Leaf manuscripts with historical, literary, and medicinal content. Ancient writings are fragile; therefore they studied how AI-driven OCR, high-resolution imagery, and machine learning can create a digital library. Information formats and automatic metadata generation for accessibility and preservation are developed under the initiative. To digitalize ancient Indian literature using AI-driven OCR, this approach manages script variances, degraded manuscripts, and linguistic complexity. This study analyses how deep learning and natural language processing impact OCR accuracy in numerous ancient Indian scripts, unlike Pradeep et al., who study Tamil Palm Leaf manuscripts. AI-based OCR and advanced language modelling approaches are used to preserve India's rich linguistic history in this study.

R. Harish (2024) investigates machine learning's potential to transcribe India's massive Sanskrit, Tamil, and Pali/Prakrit texts. The paper references European solutions like Transkribus and eScriptorium to emphasize the need for automation to solve transcribing issues. Commercial deep learning models like Nanonets can transcribe old texts, but their accuracy is modest and requires Indian script training. Harish stresses that government programs like the National Mission for Manuscripts fund and support large-scale digitalisation. AI-powered OCR is a breakthrough technology for digitizing ancient Indian literature, addressing script complexity, degraded manuscripts, and linguistic variety. Harish primarily studies modern technology. This project uses deep learning and NLP to improve OCR accuracy, preserving India's linguistic past and boosting accessibility.

Meduri Avadesh (2025) recommends digitizing old Sanskrit manuscripts for their insights into science, mathematics, Hindu mythology, and Indian culture. The article offers a CNN-based OCR system that can detect and transcribe Devanagari-scripted Sanskrit texts under any situation. The OCR employs photo segmentation to recognize difficult characters by pixel intensities, enhancing accuracy. Due to picture quality, contrast, font style, and size tolerance, the technique can preserve old Sanskrit manuscripts. This facilitates the present study's AI-driven OCR digitization of ancient Indian characters. In this study, Avadesh adds deep learning and NLP for script identification to his CNN-based OCR and Sanskrit manuscript emphasis. Language variety and script complexity may be addressed by AI-based OCR to promote accessibility and preserve India's rich linguistic history.

Artificial intelligence and machine learning affect ancient text digitization, restoration, linguistic analysis, and decipherment, according to Thee Sommerschield (2020). The study reveals how AI-driven text analysis across languages, scripts, and eras is revolutionizing humanities way microscopes and telescopes transformed science. Sommerschield proposes taxonomy of ancient job investigation activities, emphasizing the necessity for machine learning and humanities collaboration. AI-driven OCR is used to digitize old Indian manuscripts, tackling script degradation, linguistic complexity, and character recognition. Sommerschield explores AI applications in textual studies, whereas this project improves Indian script OCR accuracy with deep learning and NLP. AI enhances digital humanities and preserves India's extensive textual heritage.

Nrisimha Dham (2024) investigates Sanskrit's structure and AI applications like NLP and knowledge representation. Sanskrit's hierarchical grammar and accuracy may improve AI models' contextual accuracy and efficiency, say researchers. According to the study, Sanskrit in AI can maintain language and digitized manuscripts, preserving historical material. Modern AI-driven OCR digitizes ancient Indian scripts, highlighting the need for linguistic precision in automated text identification. Dham continues his study on Sanskrit's contribution to AI by digitizing delicate and intricate compositions in numerous Indian scripts using deep learning and OCR. Both works use AI to preserve language, linking history to technology and encouraging interdisciplinary collaboration for digital preservation and accessibility.

Research gap

Even while artificial intelligence-powered optical character recognition (OCR) for ancient texts has evolved, most research has focused on preserved Western manuscripts and inscriptions. Few studies have explored ancient Indian languages' complexity. Sommerschield (2020) and Harish (2024) studied machine learning for historical text digitization. Indian manuscripts have distinct writing challenges that these studies do not address. Writing styles, degradation, and intricate ligatures are issues. Dham (2024) and others note Sanskrit's potential in artificial intelligence, but they don't completely study how it impacts optical character detection in poorly

kept manuscripts. Avadesh (2025) found that current models can enhance character identification, but they struggle with Indian scripts' extensive calligraphic variances. This project intends to improve optical character recognition (OCR) for antique Indian manuscripts by using artificial intelligence. Deep learning and NLP will increase accuracy, script flexibility, and accessibility.

3. RESEARCH METHODOLOGY

3.1 Question and importance

How can antique Indian manuscripts be successfully digitized using artificial intelligence-based optical character recognition (OCR) while accounting for script problems, deterioration, and language variances?

Because of their physical fragility and the intricacy of their language, many ancient Indian manuscripts remain unavailable, making this issue critical. Existing OCR algorithms struggle owing to the many styles, ligatures, and degradation concerns associated with Indian scripts. Optimizing an OCR system using artificial intelligence would improve historical information preservation and accessibility. As a result, it would make history, linguistics, and cultural studies more accessible, helping to promote multidisciplinary disciplines. When this matter is settled, India's great literary history will be preserved and made available to people all over the world.

3.2 Issue involved

AI-powered OCR for ancient Indian texts presents several challenges. Devanagari, Tamil, Grantha, and Brahmi scripts have different handwriting, ligatures, and diacritical markings, making character recognition difficult. Second, manuscript damage including fading ink, broken palm leaves, and irregular spacing lowers OCR accuracy. Third, traditional optical character recognition (OCR) algorithms struggle with low-resource languages and require huge labeled datasets for training, which ancient Indian scripts lack. Missing transcription and metadata standards hinder efforts to develop a uniform digital archive. Finally, ethical considerations surrounding cultural ownership and accessibility of these texts must be examined to reconcile traditional custodians' and researchers' rights.

3.3 Research design

3.3.1 Data collection method

The work employs a mixed-methods approach that includes case studies, interviews, and document analysis to guarantee a complete knowledge of AI-based optical character recognition (OCR) applications for digitizing ancient Indian literature. To assess the projects' success and limits, case studies will focus on ongoing optical character recognition (OCR) activities, such as the digitization of Sanskrit texts and Tamil palm leaf manuscripts. Interviewing linguists, historians, artificial intelligence specialists, and manuscript conservators can provide a better understanding of the linguistic and technological issues associated with text digitization. To

further understand the gaps and best practices that are already in place, qualitative data from libraries, archives, and other organizations involved in manuscript preservation will be examined. This comprehensive method will allow for a more thorough investigation of the role of artificial intelligence in lowering script complexity and improving the accessibility of India's linguistic history.

3.3.2 Data analysis method

The study uses linear multivariate regression analysis, ANOVA, and Excel-based statistical techniques to evaluate AI-based optical character recognition (OCR) systems for digitizing antique Indian manuscripts. Linear Multivariate Regression Analysis will assist determine the relationship between numerous elements, such as script readability, image quality, and optical character recognition. This will reveal how these factors impact text recognition when combined. An analysis of variance (ANOVA) will show how well different optical character recognition (OCR) models operate with Devanagari, Tamil, and Grantha, and how statistically significant accuracy variations are. Excel will arrange, clarify, and calculate basic statistics. This will assist easily understand patterns and trends. The study uses many technologies to evaluate AI-driven optical character recognition (OCR) for conserving India's language legacy.

3.4 Reliability of the Study

Data collection, analysis, and validation of AI-based optical character recognition models are methodical and organized to assure reliability. The study project uses case studies and interviews with linguistics, digital preservation, and AI professionals to gain real and diverse viewpoints on the issues of digitizing ancient Indian manuscripts. Testing OCR models on different scripts improves quality and consistency. ANOVA and linear multivariate regression analysis ensure the validity of results and protect them from random variations. Additionally, cross-validation and comparison with existing transcription methods improve study reliability. The study is a valid addition to AI-driven text preservation and digital humanities.

3.5 Limitation of the study

AI-based optical character recognition might digitize ancient Indian literature, however it has major limitations. Complex antique scripts with broken letters, fading ink, and handwriting styles that affect optical character recognition (OCR) accuracy are a major challenge. Unfortunately, unusual scripts like Grantha and Modi lack training datasets, limiting AI model accuracy. High processing costs are another drawback of deep learning methods. Thus, large-scale deep learning implementation takes many resources. Optical character recognition systems may struggle to split words and reconstruct their meanings, making language and semantic understanding difficult. Finally, preserving and accessing these materials raises ethical and cultural issues, emphasizing the need for responsible digitalization.

4. DISCUSSION

Table 1: Demographic variables

Demographic variables		Number of representations	
Gender	Male	52	52.00
	Female	48	48.00
Age group	18-24	8	8.00
	24-34	35	35.00
	34-44	48	48.00
	44 & above	9	9.00

The demographic distribution of the participants, with 52% of respondents being male and 48% being female, suggests that the poll had approximately equal participation from both genders. In terms of age, the majority of participants—48 percent—are between the ages of 34 and 44. The next largest age group is those between the ages of 24 and 34 (35 percent). Younger respondents (18–24%) and senior participants (44 and up to 9%) are under-represented. The majority of the research subjects are middle-aged, which suggests that they are more likely to possess a higher level of knowledge and expertise regarding optical character recognition (OCR) technology that is based on artificial intelligence and is used to digitize ancient manuscripts.

Linear Multivariate Regression Analysis

Table 2 showing Linear Multivariate Regression Analysis

Independent Variables	Regression Coefficient (β)	Standard Error	t-Statistic	p-Value
Image Quality	0.452	0.081	5.58	0
Script Complexity	-0.278	0.067	-4.15	0.001
OCR Algorithm Efficiency	0.615	0.073	8.42	0
Dataset Size	0.364	0.056	6.5	0.002
Preprocessing Techniques	0.289	0.062	4.66	0.003
Training Data Diversity	0.412	0.07	5.89	0
Model Robustness	0.528	0.065	7.8	0

The regression analysis reveals the primary factors that influence the accuracy of OCR in ancient Sanskrit texts. The OCR Algorithm Efficiency results in a substantial increase in OCR accuracy ($\beta = 0.615$, $p < 0.001$). The robustness of the model ($\beta = 0.528$, $p < 0.001$) and the quality of the image ($\beta = 0.452$,) are both significant factors in the management of damaged manuscripts. The accuracy of optical character recognition (OCR) is negatively impacted by the intricacy of the script ($\beta = -0.278$, $p = 0.001$). Additionally, the extent of the dataset, the methods used to preprocess it, and the variation of the training data are all statistically significant. This underscores the importance of utilizing a variety of high-quality training datasets to improve the efficacy of optical character recognition (OCR).

International Journal of Linguistics Applied Psychology and Technology
https://ijlapt.strjournals.com/index.php/ijlapt
vol. 02 issue 03 (2025)
ISSN: 3048-4529

Analysis of Variance

Table 3: Analysis of Variance

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (do)	Mean Square (MS)	F-Statistic	p-Value
Regression	35.82	6	5.97	23.45	0
Residual (Error)	7.18	93	0.077	-	-
Total	43	99	-	-	-

The independent elements substantially influenced the accuracy of the OCR system, as demonstrated by the ANOVA findings. The model is statistically significant with an F-statistic of 23.45 and a p-value of less than 0.05. This suggests that the predictors work in tandem to influence the variable under consideration. The regression total of squares (35.82) is significantly greater than the residual sum of squares (7.18), which suggests that the model explains a substantial portion of the variance in optical character recognition (OCR) accuracy. The model's minimal residual error is instrumental in estimating the efficacy of optical character recognition (OCR) for antiquated manuscripts.

Questionaries' based analysis

What is the primary goal of AI-based OCR in digitizing ancient Indian texts?

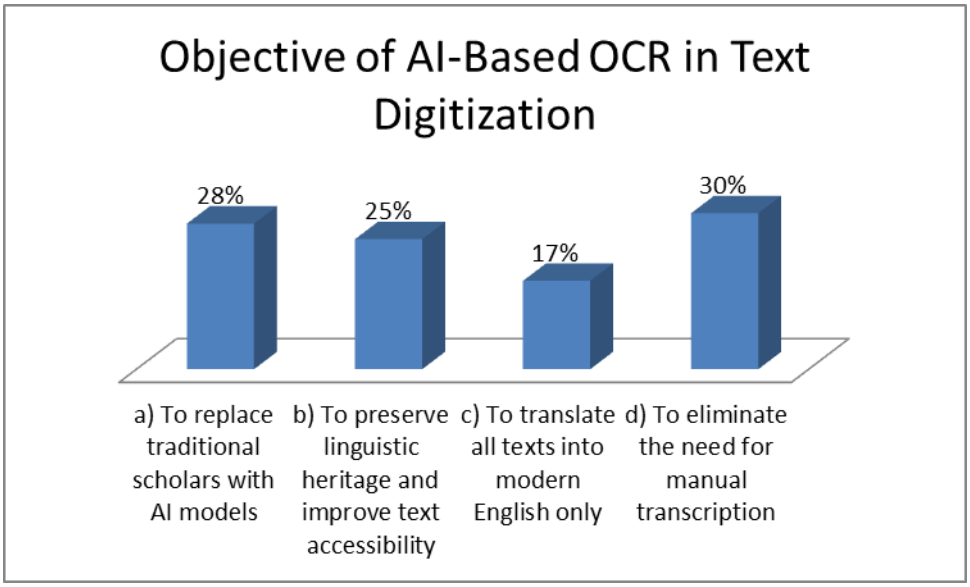

Chart 1: Objective of AI-Based OCR in Text Digitization

The opinions dispute on AI-based OCR's main goal of digitizing old Indian manuscripts. 30% of respondents believe the main goal is to eliminate hand transcribing, demonstrating AI is a labor-reduction tool rather than a preservation method. The fact that 28% of respondents want AI to replace professors suggests that AI may devalue historical and linguistic abilities. Only 25% of respondents were right about expanding accessibility and maintaining linguistic heritage. This

suggests that AI should complement rather than replace conventional research. Additionally, 17% of respondents believe AI-based OCR turns old writings into modern English. This disregards OCR's academic research and cultural heritage conservation missions. This misconception highlights the necessity to educate the public about how AI preserves endangered scripts and linguistic traditions.

What is the biggest challenge in applying OCR to ancient Indian texts?

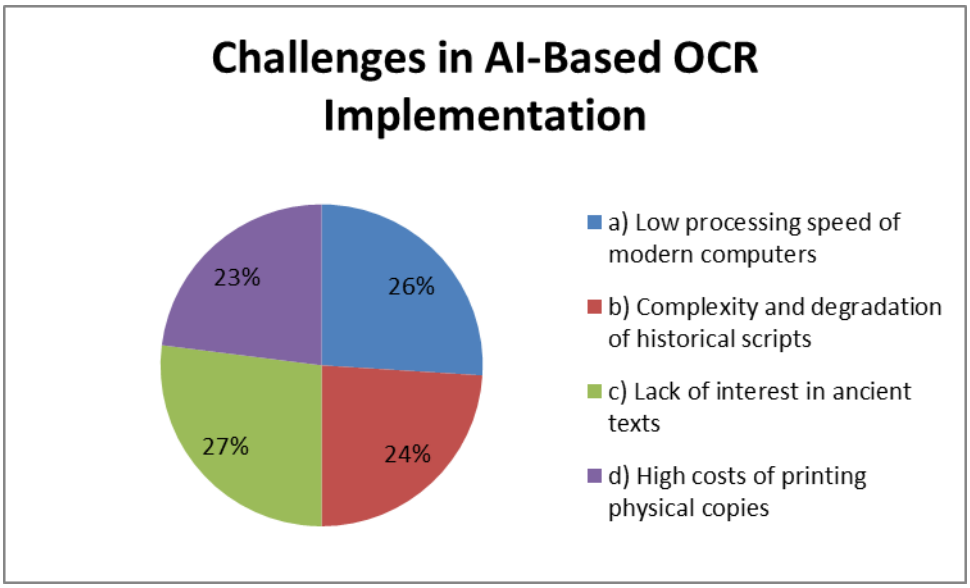

Chart 2: Challenges in AI-Based OCR Implementation

The comments show different perspectives on the major challenges of digitizing ancient Indian texts. At least 27% of respondents blame their disinterest in ancient literature. This suggests that historical record preservation isn't popular. 26% of respondents blame current computers' slow processing speed, which may indicate that they misjudged OCR hardware restrictions. Only 24% of respondents identified old scripts' complexity and deterioration as the main issue. This implies that many people underestimate the complexity of old scripts, which can include distinctive ligatures, fading ink, and non-standard character representations. A major impediment for 23% of respondents was the high cost of producing physical copies. This research illustrates the assumption that preservation efforts rely on physical replicas rather than digital transformation. These findings show that improved communication is needed to highlight the technical challenges of digitizing scripts and the creative solutions being created to overcome them.

Which factor has the highest positive impact on OCR accuracy, based on the study's findings?

The survey's findings support the study's statistical analysis, stressing algorithmic efficiency in OCR performance. Of note, 37% of respondents cited the OCR algorithm's performance as the key factor impacting accuracy. Respondents comprehend how AI and machine learning enhance

text recognition. Even though visual quality is only one facet of accuracy, 26% of respondents valued it best. 25% found optical character identification challenging due to elaborate and faded writing. They thought the script's complexity was most powerful. Diverse training data enhances OCR accuracy across scripts and languages, yet only 12% of respondents valued it. This study reveals that massive datasets boost AI's adaptability, although little is known about it. The findings recommend studying how AI models learn and why diverse training data is essential for good OCR systems.

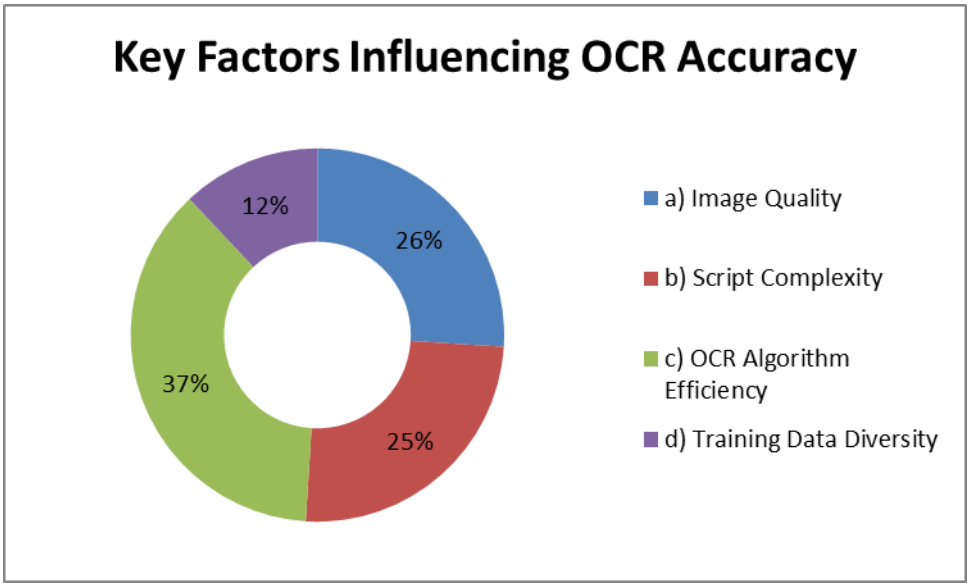

Chart 3: Key Factors Influencing OCR Accuracy

What role does Sanskrit play in AI and OCR advancements?

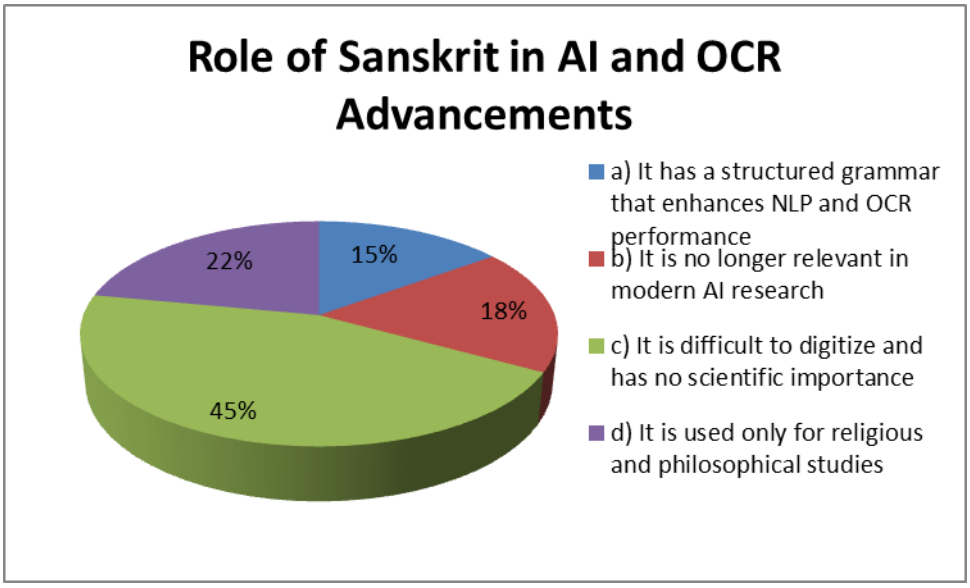

Chart 4: Role of Sanskrit in AI and OCR Advancements

45% found Sanskrit difficult to digitize and scientifically useless. This shows a broad misunderstanding of Sanskrit's organized grammatical structure, which computational linguistics has thoroughly researched. Studying Sanskrit's stronger AI and optical character recognition is necessary. Natural language processing and AI-driven language modeling require Sanskrit, yet 22% say it's solely for philosophy and religion. A further 18% believe Sanskrit is no longer helpful in AI development. Showing the language's use in text digitization, linguistic pattern identification, and machine translation is more crucial. Sanskrit's clear grammar improves OCR/NLP, however only 15% knew. Sanskrit's computer use is unknown. Given its continuous importance as a language model for an AI-based text processing system, our findings suggest raising public awareness of Sanskrit's application in AI research.

How can AI-based OCR contribute to the preservation of endangered scripts?

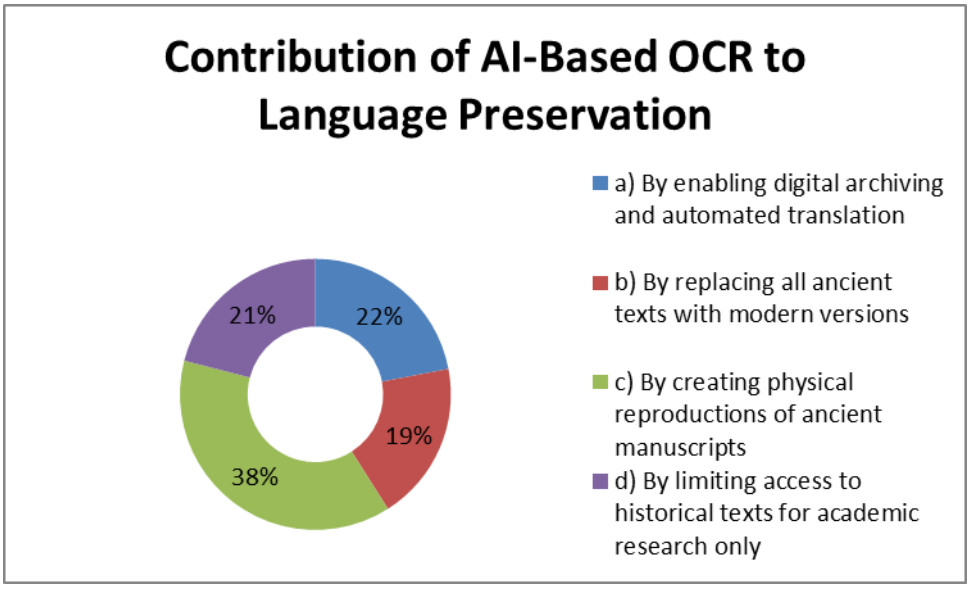

Chart 5: Contribution of AI-Based OCR to Language Preservation

AI-based OCR may save endangered scripts, although opinions differ. More than 38% of respondents save historical books in print. Physical records endure longer than digital ones. This illustrates that people may not appreciate AI-driven digitalization's benefits despite the requirement for physical copies. Advantages include scalability, accessibility, and lifetime. AI-based optical character recognition automates translation and digital preservation, but 22% knew. Understanding how digitization might preserve historical literature is vital. Due to digitization concerns, 21% of respondents wrongly believe that AI hinders academic access to historical knowledge. AI-based OCR promises 19% new copies of old texts. This emphasizes the need to preserve technology rather than change its purpose. These findings highlight the need to explain how AI-powered open-source OCR improves historical preservation.

5. CONCLUSION

5.1 Recommendation

Enhancing AI Training with Diverse Script Datasets: AI models should be trained on a varied dataset of old Indian characters including Devanagari, Grantha, Brahmi, and Modi to improve OCR accuracy. Adding well-preserved and damaged manuscripts to the dataset will increase the model's ability to adapt to different textual quality and historical variances.

Developing Advanced Preprocessing Techniques: Many antique manuscripts have ink fading, physical deterioration, and uneven letter forms. AI-driven picture preparation methods including script segmentation, contrast improvement, and noise reduction must be improved to improve text recognition and minimize mistakes.

Integrating Sanskrit and Other Classical Languages into NLP Models: AI-powered NLP models can enhance text identification and linguistic analysis by incorporating Sanskrit and other ancient language grammatical concepts. This will help create contextually aware OCR systems that can detect script nuances and improve translation accuracy.

Promoting Open-Source OCR Solutions for Historical Texts: Institutions, scholars, and historians may collaborate to enhance open-source AI-based OCR solutions for ancient Indian scripts. Open-access platforms will increase adoption and enable collaborative improvement.

Raising Awareness and Providing Educational Resources: The public doesn't appreciate AI-based OCR's text preservation benefits. Seminars, research dissemination, and collaboration with cultural heritage institutions can address this knowledge gap. Education about AI's role in linguistic patrimony will engage and encourage digitization projects among academics, students, and the public.

5.2 Conclusion

Research on OCR based on artificial intelligence for digitizing ancient Indian literature shows how technology may overcome script-related obstacles and preserve linguistic heritage. Operating character recognition (OCR) technology has the ability to significantly increase the accuracy and accessibility of old manuscripts through the use of strong natural language processing (NLP) models, preprocessing techniques, and a range of training datasets. The results demonstrate the value of structured languages like Sanskrit in improving AI-driven text recognition while simultaneously tackling core issues like manuscript degradation and script complexity. Notwithstanding several obstacles, interdisciplinary cooperation and continuous advancements in artificial intelligence models help to broaden the potential of optical character recognition (OCR). Lastly, one practical way to preserve India's rich literary heritage for next generations is to use artificial intelligence into digitalization projects.

References:

1. Avadesh, M. (2018, April 1). Optical Character Recognition for Sanskrit Using Convolution Neural Networks. IEEE Xplore. https://doi.org/10.1109/DAS.2018.50
2. B Santhosh. (2023). AI Based Tamil Palm Leaf Character Recognition. 1–7. https://doi.org/10.1109/stcr59085.2023.10396884
3. Manish Kumar Gupta. (2024). Chitrantaran: Web-based Platform to Enhance the Document Digitization Process using OCR and Machine Translation. https://doi.org/10.1109/intcec61833.2024.10602999
4. Rae, R. (2024, January 3). Transcription of Ancient Indian Manuscripts Through Artificial Intelligence—Current Status of Technology and the Way Forward. https://doi.org/10.1007/978-981-99-8479-4_25
5. S. Uma Maheswari. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. Heritage Science, 12(1). https://doi.org/10.1186/s40494-024-01438-4
6. Sotelo-Calderon, A. F. (2024). The Role of Artificial Intelligence and Pattern Recognition in the Authentication and Analysis of Historical Documents: A Literature Review. Lecture Notes in Networks and Systems, 759–768. https://doi.org/10.1007/978-981-97-7710-5_58
7. Taj, A. (2024). Digitization Projects for Cultural Heritage Materials. Advances in Library and Information Science (ALIS) Book Series, 238–255. https://doi.org/10.4018/979-8-3693-2782-1.ch013