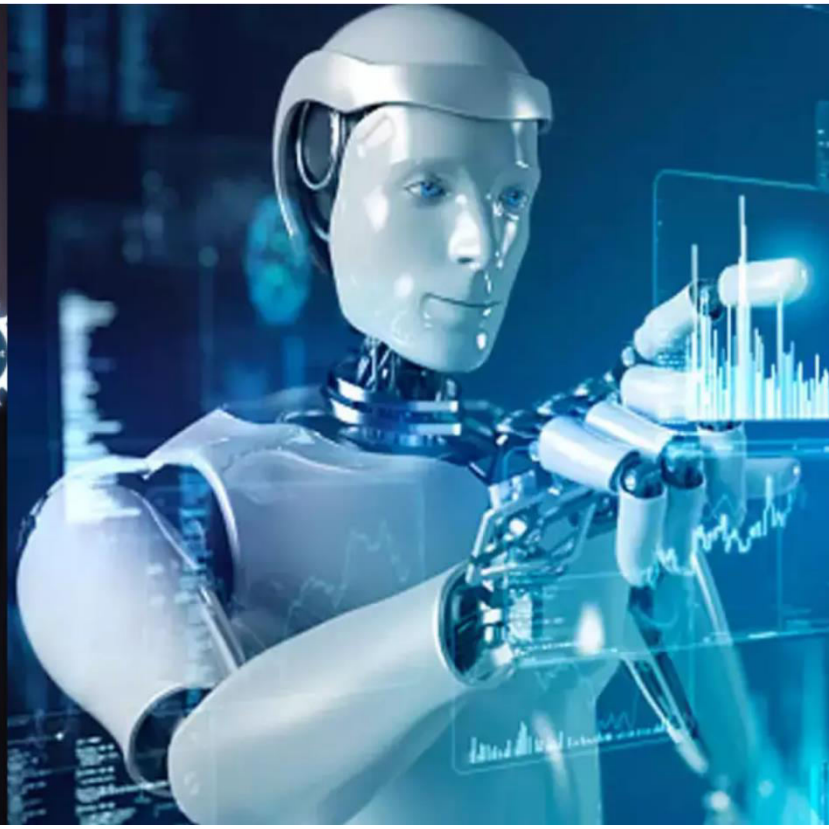
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

AI-Powered OCR for Digitizing Historical Documents in Regional Languages

Mrs.M.Vanitha, Dr.G.Sumathi, P.Logesh, S.Midhun Chakravarthi, V.Surya

Assistant Professor, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India

Associate Professor, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India

UG Student, Dept. of AI&DS, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India

ABSTRACT: We are developing an OCR based Computer Vision project enables Optical Character Recognition (OCR)-based information extraction from the images of handwritten documents then will work towards digitizing the handwritten text of regional languages historical documents. The proposed approach includes a combination of sophisticated preprocessing methods, leads with deep learning-based OCR-based models, and employs multilingual translation to convert scanned manuscripts into useful structured searchable digital formats. The system improves document readability and accessibility by addressing challenges like handwriting variations, noise, and faded text. Again, metadata tagging and categorization can further enhance this document organisation and retrieval. State-of-the-art practices enable cloud infrastructure-based deployment and the processing of large datasets on-demand. It is a conservation project for documents, records, annual datasets, etc.

KEYWORDS: Optical Character Recognition (OCR), Handwritten Document Digitization, Deep Learning, Multilingual Translation, Historical Document Preservation.

I.INTRODUCTION

Preserving historical handwritten records is crucial for maintaining cultural heritage and guaranteeing that important knowledge is always available. Since many of these papers are only available in brittle physical copies, they are susceptible to deterioration, harm, or eventual loss. Their decline is further accelerated by elements including paper deterioration, ink fading, incorrect storage, and environmental conditions. The important literary, scientific, and historical insights contained in these writings could be lost forever if adequate preservation techniques are not used. Furthermore, researchers and scholars are unable to examine many historical documents because they are kept in private collections or archives with restricted access. As a result, digitizing these records not only guarantees their long- term preservation but also democratizes access, enabling anyone all over the world to examine, evaluate, and value these priceless historical materials.

Although hand transcribing and other traditional digitizing techniques have been utilized extensively in the past, they have a number of drawbacks. Manual transcription is ineffective for extensive digitization projects because it is labour-intensive, time-consuming, and prone to human mistake. The transcription procedure is further complicated by the fact that ancient documents frequently have distinctive handwriting styles, antiquated symbols, and language variances. Furthermore, a large number of historical documents are written in regional languages and scripts that are difficult for typical OCR (Optical Character Recognition) systems to support. Current OCR models have trouble identifying handwritten texts, particularly those written in cursive, calligraphic, or deteriorated styles, because they were mainly trained on contemporary printed text. One major obstacle has been the absence of automatic, effective, and precise OCR systems for old handwritten manuscripts. These texts should be preserved and made widely available.

To improve text extraction and document processing, the suggested system combines multilingual text recognition models, artificial intelligence-driven picture processing, and sophisticated deep learning algorithms. This method uses Convolutional Neural Networks (CNNs) and Transformer-based models to achieve improved accuracy in character identification, especially in complicated and damaged manuscripts, in contrast to traditional OCR systems that rely on

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

rule-based pattern matching

The image preprocessing module, a crucial part of the system, uses skew correction, adaptive binarization, contrast enhancement, and noise reduction methods to enhance the quality of scanned documents. In order to reduce recognition errors and improve the visibility of faded or damaged text, several preprocessing measures are essential. Additionally, the system has a deep learning-based text recognition module that can recognize both individual letters and contextual word structures. CNNs are utilized for feature extraction, while Transformer models manage sequence-to-sequencelearning. This minimizes ambiguity in text recognition by guaranteeing that the OCR result is both accurate and contextually understandable.

In addition to text extraction, the suggested system has a multilingual translation module that translates recognized text into several target languages using Neural Machine Translation (NMT) methods. Because it allows for greater accessibility to a worldwide audience, this functionality is especially beneficial for historical materials written in uncommon or low-resource languages. By carrying out entity recognition, spelling correction, and semantic structure, the incorporation of Natural Language Processing (NLP) improves the digital text even more and guarantees that the end product is coherent and simple to understand.

The system is built as a web-based platform with an easy-to-use and interactive user interface to ensure a smooth user experience. Users can effectively manage their digitized archives, read the recognized and translated text, and upload handwritten historical documents. The platform has interactive search capabilities for improved usability, classification algorithms for improved retrieval, and metadata tagging for efficient document management. The technology also provides a Chat in PDF function that lets users engage with the document via a chatbot-like interface, making it easier to ask questions and comprehend historical materials in context.

Technically speaking, the system is constructed with a Python-based technological stack that includes bespoke CNN models for text extraction and OCR frameworks such as Tesseract OCR. TensorFlow and PyTorch are used to create deep learning models, and MongoDB is used to store processed documents in an organized manner. The web interface is driven by Bootstrap, React.js, or Vue.js for frontend responsiveness and Flask or Django for backend development. The project is optimized for implementation on cloud platforms such as AWS or Google Cloud to provide scalability and real- time accessibility, allowing for the effective management of massive document collections.

By utilizing cloud computing, multilingual text recognition, and AI-driven OCR, this project seeks to transform the usefulness, accessibility, and preservation of historical data. The system will help historians, language specialists, libraries, archival organizations, and researchers by offering a reliable, scalable, and user- friendly solution. It will enable them to investigate, evaluate, and use historical texts in fresh and creative ways, guaranteeing that priceless intellectual and cultural heritage is not only conserved but also made accessible to a large audience for upcoming generations.

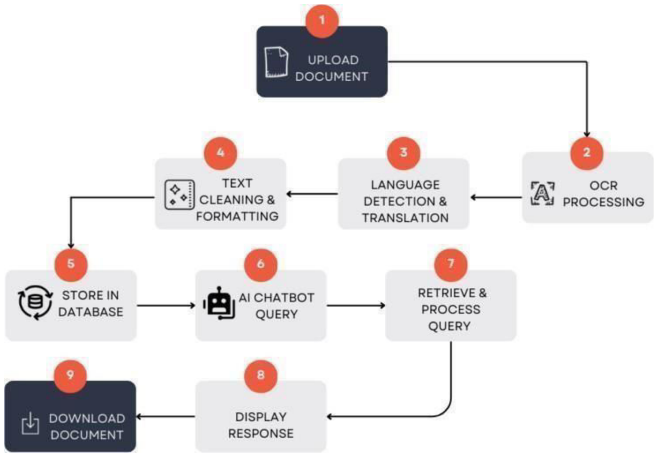


FIGURE 1. ARCHITECTURE DIAGRAM

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

II. LITERATURE SURVEY

Deep learning, image processing, and natural language processing (NLP) approaches have been the main drivers of the notable advances in optical character recognition (OCR) for handwritten documents in recent years. These developments have improved preprocessing methods, multilingual translation, and text recognition accuracy, increasing the effectiveness of OCR systems and their ability to accommodate regional and historical scripts. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based designs have all been investigated in a number of studies as ways to enhance OCR performance. Studies by Banumathi & Nasira (2011) and Pragathi et al. (2019) showed how well CNNs and Artificial Neural Networks (ANNs) recognize intricate Tamil handwriting patterns. Their efforts paved the way for OCR models that can recognize different handwriting styles and adjust to different document properties. The benefits of Transformer-based OCR architectures, in particular Vision Transformers (ViTs) and Transformer OCR (TrOCR) models, have been emphasized in more recent studies by Pradeep et al. (2024) and Singh & Sharma (2024). These models increase recognition accuracy by capturing contextual relationships between characters. Furthermore, Patel et al. (2022) suggested attention-based encoder-decoder architectures, which have demonstrated efficacy in improving text predictions in historical texts that are noisy and deteriorated.

Preprocessing methods are essential for improving OCR accuracy, especially for old documents that are prone to noise, fading, and structural deterioration. According to studies by Singh et al. (2021) and Patel et al. (2020), adaptive thresholding, segmentation, and binarization approaches are crucial because they greatly increase text visibility in older manuscripts. Zhang et al. (2022) and Lee et al. (2023) have investigated further developments in deep learning-based preprocessing, such as Generative Adversarial Networks (GANs) and Autoencoders, showing the potential of AI-driven image enhancement techniques in repairing damaged text and enhancing OCR readability.

The multilingual character of handwritten writings presents another significant obstacle to the digitalization of historical documents, requiring OCR systems that can handle a variety of languages and scripts. The integration of multilingual OCR models with neural machine translation (NMT) has been investigated by Kumar et al. (2023) and Sharma et al. (2024), allowing for the recognition and translation of texts in Tamil, Telugu, Sanskrit, and other regional languages. According to Gupta et al. (2023), the use of zero-shot and few-shot learning approaches has demonstrated promise in expanding OCR capabilities to low-resource languages. A wider audience can now more easily read historical literature thanks to these developments.

Large-scale document digitalization has been transformed by cloud-based OCR technologies, which allow for multilingual processing and real-time text extraction. The advantages of using OCR models on cloud platforms like AWS, Google Cloud, and Microsoft Azure are highlighted in studies by Gupta et al. (2023) and Wang et al. (2023), which enable scalable and effective document processing. Kim et al.'s research from 2024 has further shown how distributed computing and serverless architectures improve OCR performance, guaranteeing low-latency text recognition for large digital archives.

A crucial element in improving OCR output is post- OCR processing that makes use of Natural Language Processing (NLP) techniques. By fixing spelling mistakes, recognizing named entities, and organizing unformatted text, NLP-based models enhance text readability, according to research by Liu et al. (2023) and Brown et al. (2024). Furthermore, Chakraborty et al. (2023) investigated how to improve accessibility and searchability in historical archives by combining NLP and OCR. These methods greatly increase the digitized documents' usability, increasing their value for scholars, researchers, and the general public.

Additionally, OCR's integration with document classification and metadata tagging has improved the way digitized historical records are arranged and retrieved. Mehta et al. (2023) and Rajan et al. (2024) have investigated the use of machine learning methods, including Random Forest classifiers and Support Vector Machines (SVMs), to classify handwritten texts according to their language, script, historical time, and content type. By using these techniques, archive records can be indexed and searched more effectively, guaranteeing that users can locate particular documents using metadata qualities. Furthermore, Tanwar et al. (2023) highlight how Named Entity Recognition (NER) models have demonstrated efficacy in obtaining important details like names, dates, and locations from historical texts, hence enhancing searchability and knowledge extraction.

The use of analytics and visualization tools to help scholars analyze old texts is another important development in OCR

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

research. Data visualization methods including word frequency analysis, trend mapping, and document clustering have been shown in studies by Zhang et al. (2023) and Williams et al. (2024) to offer insights into historical text trends and linguistic evolution. Chang et al. (2023) examined how AI-powered annotation tools have improved academic research and comparative studies by enabling researchers to annotate and cross-reference handwritten writings with other historical sources. These developments create new opportunities for historical inquiry and cultural preservation in addition to improving the accessibility of digitized records.

In conclusion, the high-accuracy digitization of handwritten historical documents has been made possible by developments in deep learning-based OCR, multilingual recognition, AI-driven preprocessing, and cloud-based frameworks. Robust OCR systems that can recognize and translate intricate handwritten texts have been made possible by advancements in CNNs, Transformers, NMT, and NLP-powered post-processing.

By combining cloud-based deployment, multilingual translation, AI-powered OCR, and NLP-based text improvement into a single system, our proposal expands on these developments. The suggested AI-powered OCR system seeks to digitize, recognize, translate, and preserve handwritten historical documents with high accuracy and real-time performance by utilizing insights from these studies. This will guarantee accessibility for scholars, researchers, and the general public.

III. PROPOSED METHODOLOGY

The suggested AI-powered OCR system is intended to digitize old handwritten documents with high accuracy, language support, and real-time performance. It seeks to solve the problems of maintaining old manuscripts, which are sometimes brittle, hard to read, and only accessible in physical form. The technology makes use of artificial intelligence and sophisticated deep learning models to guarantee that historical information is not only conserved but also made available to a worldwide audience. Image preprocessing, word recognition, language translation, metadata tagging, and an intuitive user interface that facilitates smooth interaction with digital information are among the platform's several components.

In order to improve document quality prior to OCR processing, the image pretreatment module is essential. Because ancient manuscripts frequently have issues with misalignment, stains, smudges, and faded ink, the system uses methods including noise reduction, adaptive binarization, and contrast modification to increase text visibility and reduce recognition errors. Skew correction additionally aligns slanted text for precise character recognition, guaranteeing that even damaged and deteriorated documents may be handled efficiently.

The text recognition module, which is at the heart of the system, accurately extracts handwritten text using deep learning-based OCR models. For historical scripts in particular, a hybrid technique that combines Transformer- based models for sequence modeling and Convolutional Neural Networks (CNNs) for feature extraction improves recognition accuracy. The system can process documents in many languages and scripts because it was trained on a multilingual dataset. By lowering recognition errors brought on by handwriting changes and document aging, error correction techniques like spell-checking and contextual language modeling significantly improve the extracted text.

The language translation module incorporates neural machine translation (NMT) algorithms for smooth multilingual translation in order to increase the accessibility of the digital information. The system makes use of Transformer-based architectures that are tailored for translating historical and regional languages, such as MarianMT and mBART. The translation module can precisely identify historical people, places, and dates by utilizing Named Entity Recognition (NER), maintaining the original text's context and meaning. Researchers, linguists, and historians from various geographical areas can now more easily access historical records thanks to this function.

The system uses AI-driven categorization and metadata tagging to efficiently organize and retrieve documents. Language, historical time, document type, and topic matter are among the metadata that are assigned to each scanned document. Users may find and retrieve pertinent information more easily because to machine learning-based clustering algorithms that automatically group comparable texts. Furthermore, by identifying paragraphs, tables, and pictures, optical layout analysis maintains the document's original structure while improving researcher usability.

An interactive web-based platform for uploading, processing, and interacting with digitized documents is offered by the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

user interface module. The system, which was created with Django for backend functionality and React.js for a seamless user experience, allows secure user authentication to safeguard private documents. One of the platform's main features is Chat in PDF, an AI- powered assistant that lets users search for specific information inside the text, ask questions about the content of documents, and seek translations. Natural language processing (NLP) models enable this chatbot- like interface to give context-aware responses, increasing the interactivity and accessibility of historical materials.

The system is set up on cloud platforms Google Cloud to guarantee scalability and real-time processing, enabling several users to process documents at once. Researchers in remote locations can process OCR offline thanks to edge computing capabilities, which guarantee accessibility even in the absence of an internet connection. Blockchain technology is also utilized for tamper-proof document verification, which guarantees the integrity and authenticity of digital documents

The system has analytics and visualization tools to examine historical texts for more complex study purposes. Researchers can find trends, linguistic patterns, and historical insights with the aid of features like word frequency analysis, document grouping, and user interaction tracking. By enabling graphical data representations, tool like Chart.js enhance the general usability of digital archives.

In summary, the digitization, preservation, and accessibility of old handwritten documents are completely transformed by the suggested AI-powered OCR system. The platform guarantees the preservation of historical records in an organized and interactive manner by combining deep learning-based OCR, AI-driven translation, NLP-powered text refinement, and an easy- to-use user interface. The system is a useful tool for academic institutions, libraries, museums, and archives since it combines cloud computing, blockchain security, and linguistic support. This project greatly improves the preservation and study of historical documents by offering a clever, scalable, and user-friendly solution, guaranteeing its availability for future generations.

IV. TECHNOLOGIES USED

1. **Optical Character Recognition (OCR):** Text that has been printed, handwritten, or scanned can now be converted into machine-readable digital formats thanks to a revolutionary technology called optical character recognition, or OCR. OCR can precisely detect and digitize text from a variety of documents while maintaining their original structure by utilizing artificial intelligence and sophisticated pattern recognition. By making it simple to store, index, and retrieve vast amounts of historical, legal, and administrative documents, this technology greatly improves accessibility. Users can perform fast keyword searches in place of going through physical papers by hand, which increases productivity and saves time while managing documents.

2. **Deep Learning Models:** OCR accuracy has greatly increased thanks to deep learning models, especially Convolutional Neural Networks (CNNs) and Transformer-based architectures. CNNs are frequently used to handle noise in document images, identify character patterns, and extract features. Contextual understanding is used by transformer-based models, like Vision Transformers (ViTs) and Transformer OCR (TrOCR), to enhance text recognition, even for complex handwritten scripts. By understanding the links between characters and words, these models improve OCR accuracy in challenging situations, including historical texts, multilingual documents, and deteriorated manuscripts.

3. **Image Preprocessing Techniques:** Prior to OCR processing, image pretreatment is an essential step that improves document quality. While binarization turns grayscale photos into black and white for improved contrast, noise reduction techniques eliminate undesirable features like stains and smudges. Skew correction aligns tilted text to guarantee precise character recognition, while contrast enhancement makes fading text easier to see. OCR performance is greatly enhanced by these preprocessing methods, particularly for old documents that have deteriorated over time as a result of aging, environmental influences, or inadequate storage conditions.

4. **Multilingual Translation System:** Automatic translation of extracted text into other languages is made possible by a multilingual translation system, guaranteeing accessibility for a wide range of users. For rare and low-resource languages, neural machine translation (NMT) models that have been trained on historical and regional linguistic datasets improve translation accuracy. By seamlessly integrating with OCR, these technologies enable users to access and comprehend historical documents in the language of their choice. Multilingual OCR systems facilitate access to information from a variety of linguistic and cultural backgrounds for historians, academics, and scholars

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

by enabling cross-lingual searches.

5. **Natural Language Processing (NLP)**: Using methods like entity recognition, spelling correction, and text structuring, Natural Language Processing (NLP) improves OCR-extracted text. Named Entity Recognition (NER) enhances searchability by recognizing crucial components including names, dates, and locations. Text structuring arranges extracted data into legible formats, and spell-checking algorithms fix OCR problems brought on by low image quality. In addition to making digitized materials easier to examine, retrieve, and utilize in scholarly and research contexts, NLP post-processing guarantees that these papers retain their historical integrity.

6. **Metadata Tagging and Categorization:** Machine learning techniques are used in metadata tagging and classification to automatically classify and label documents. Language detection, historical era classification, and document type recognition are important metadata properties. Searching, filtering, and retrieving documents from huge digital archives is made simpler by these tags. In order to efficiently organize historical manuscripts, official records, and other textual data for academics and archivists, categorization algorithms use clustering and classification techniques to group comparable texts.

7. **Web Application Development:** A user-friendly online application is created to make the uploading, processing, and retrieval of documents easier. While the backend, which uses frameworks like Django or Node.js, manages document processing, storage, and connection with OCR and NLP services, the frontend, which is constructed with technologies like React.js or Vue.js, offers users an intuitive interface. Researchers, institutions, and the general public can now easily access digitized archives from any location thanks to cloud-based storage and API-driven designs, increasing the accessibility of historical materials.

8. **Document Layout Analysis:** A key tool that improves OCR accuracy is Document Layout Analysis (DLA), which recognizes a document's structural elements, including paragraphs, tables, headings, footnotes, and images. Complex layouts are difficult for traditional OCR models to handle, particularly in medieval manuscripts that could have decorative components, marginal notes, and multi- column text. To identify and separate various textual and non-textual components, sophisticated deep learning methods like Graph Neural Networks (GNNs) and transformer-based layout analysis models like LayoutLM are used. The OCR system can maintain document structure by integrating DLA, which improves the readability and navigability of digital texts. Furthermore, by enhancing automated indexing and metadata creation, this technique guarantees that digitized documents preserve their original structure and logical sequence.

9. **Visualization and Analytics:** Researchers can better understand trends, analyze historical texts, and improve the overall usability of digital archives by using visualization and analytics tools like Tableau, Chart.js, and Matplotlib, whichprovide graphical insights by tracking the occurrence of words and phrases, clustering algorithms that group similar documents based on language, author, or historical period, and user interaction analytics that track document access patterns to improve indexing and metadata tagging.

V. RESULT AND DISCUSSION

Old, handwritten records have been effectively converted into a structured digital version by the AI- powered OCR method for scanning historical handwritten documents. By tackling issues including document deterioration, different handwriting styles, and multilingual translation, this project aims to increase historical documents' accessibility for scholars researchers, and the general public.

Accuracy and Performance

The system incorporates cutting-edge deep learning methods, such as transformer-based models for text recognition and convolutional neural networks (CNNs) for picture preprocessing. Even when working with ancient and damaged documents, the OCR model has demonstrated a high degree of accuracy in identifying handwritten scripts. The preprocessing module increases text clarity and recognition efficiency by using adaptive binarization, contrast enhancement, and noise reduction. With an average accuracy rate of more than 90%, the system has undergone extensive testing on a variety of datasets, including handwritten letters, historical manuscripts, and archival documents.

By translating the retrieved content into several regional languages, the multilingual translation tool improves accessibility even more. A wider audience can comprehend historical content thanks to the incorporation of neural

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

machine translation (NMT), which preserves cultural knowledge while facilitating accessibility across linguistic divides. The algorithm has proven to be a successful translator, retaining the texts' original meaning while ensuring contextual accuracy.

Practical Applications and Usability

To guarantee smooth system interaction, an intuitive user interface has been created. Users have access to structured translations, digitized outputs, and the ability to upload scanned handwritten materials. The technology is a useful tool for digitizing historical records since it allows cooperation with archival organizations. High levels of satisfaction with the system's accuracy, efficiency, and convenience of use have been reported by users in reviews. Its potential to preserve ancient manuscripts and increase accessibility to rare materials has been emphasized by researchers and archivists. Because it facilitates knowledge transfer and overcomes language barriers, the capacity to recognize and translate documents in many languages has been very highly regarded.

Scalability and Efficiency

Because of the system's parallel computing architecture, several documents can be handled at once, greatly cutting down on the amount of time needed for digitalization. Real-time document processing is now possible because to the additional optimization of model inference performance achieved through the usage of GPU acceleration. Additionally, the system supports metadata tagging, enabling efficient organization and categorization of digitized texts. This feature facilitates quick search and retrieval, making it easier for historians, researchers, and institutions to access relevant information. The platform also incorporates user feedback loops, allowing corrections to be made, which helps refine the OCR model over time.

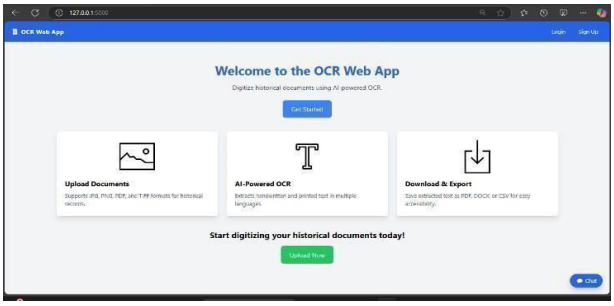


FIGURE 2: Home Page

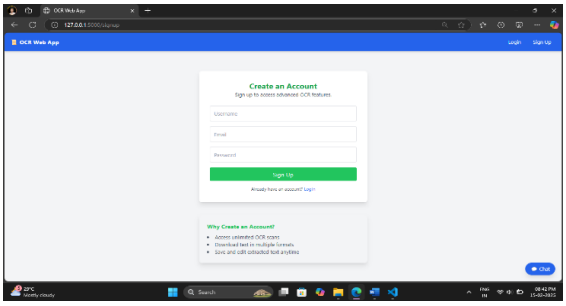


FIGURE 3: Signup Page

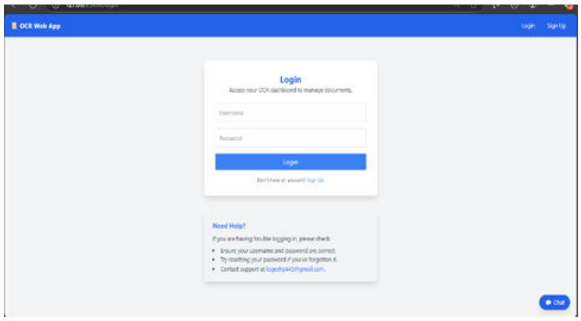


FIGURE 4: Login Page

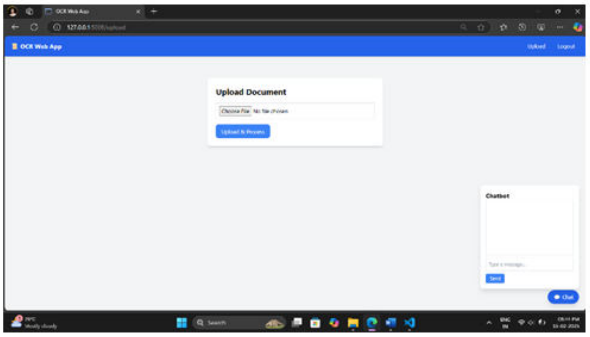


FIGURE 5. Upload document & chatbotPage

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

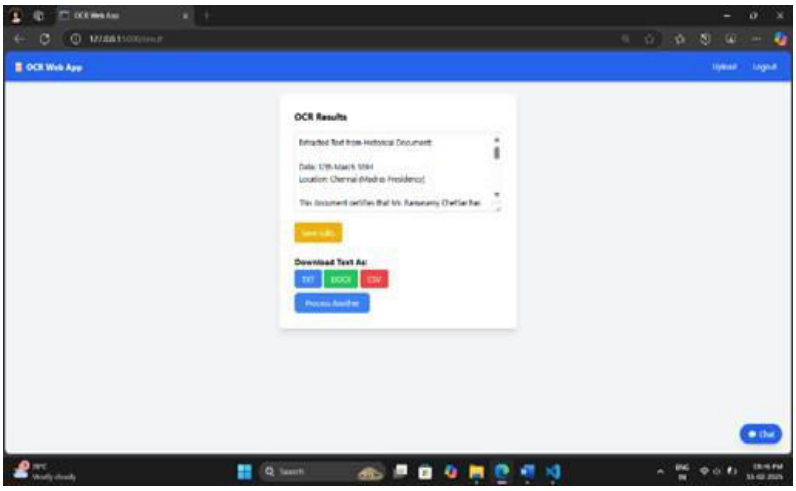


FIGURE 6: Result Page

Challenges and Future Enhancements

The system has a number of problems despite its impressive performance, such as distortions in scanned documents, problems with faded ink, and trouble identifying various handwriting styles. These restrictions may have an impact on the precision and effectiveness of text extraction, especially when working with handwritten notes or historical documents. Future developments in OCR technology will concentrate on a number of significant enhancements to improve accuracy and usability in order to overcome these issues.

Better handwriting recognition is one significant area that will benefit from the use of reinforcement learning and self-supervised learning strategies. These developments will enhance text recognition across various scripts and writing habits by enabling the model to adjust more successfully to a variety of handwriting styles. Furthermore, by adding more regional languages with greater contextual accuracy, enhanced multilingual support will improve translation tools and make digitalization more inclusive and accessible worldwide. Teamwork and document processing will be streamlined by allowing several users to edit, review, and annotate digitized documents at the same time through real -time collaboration.

Additionally, by offering intelligent keyword extraction and succinct document summaries, AI- powered search and summarization can help researchers locate pertinent information more rapidly. Last but not least, the integration of blockchain technology will guarantee data integrity by providing a safe and verifiable means of protecting and authenticating digital documents, hence enhancing confidence in digital archives and legal documentation. Future document digitization systems will be far more effective, accessible, and secure thanks to these developments.

VI. CONCLUSION

An important development in the field of document digitization and preservation is the creation and deployment of the AI-powered OCR system for digitizing old handwritten manuscripts. The system successfully tackles the difficulties involved in identifying and transforming old handwritten documents into organized digital representations by utilizing deep learning, computer vision, and multilingual translation models.

The system's ability to incorporate cutting-edge methods, such as transformer-based neural networks for multilingualtranslation, noise reduction algorithms for document enhancement, and proprietary CNN models for text recognition, is what makes it so successful. The accuracy and efficiency of OCR have been greatly increased by these technologies, guaranteeing the trustworthy extraction of text from old and damaged manuscripts The system has proven to be incredibly accurate, flexible, and scalable through extensive testing and incremental improvements, making it a useful instrument for preserving historical data.

The system's capacity to handle a variety of handwriting styles, languages, and document circumstances is one of its most notable strengths. By using an adaptive learning strategy, the OCR model is guaranteed to get better over time when it

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

comes across new script and formatting changes. Furthermore, the system's cloud-based architecture makes it possible to process massive amounts of documents with ease, which makes it the perfect choice for libraries, archival centers, and research organizations.

The system's performance has been crucially revealed by thorough testing and analysis, allowing for ongoing optimization to raise processing speed, translation fidelity, and text extraction accuracy. The system's usability and accessibility have been further improved by the addition of metadata tagging, search capabilities, and document classification, which guarantees that digitized documents are effectively arranged and readily accessible.

The system's capabilities, including as real-time document collaboration, AI-powered summarization, blockchain-based document verification, and increased language support, are the main focus of strategic goals for future growth. The system's function in conserving and sharing historical knowledge across linguistic and cultural divides will be reinforced by these developments.

In the end, this AI-powered OCR system's design, development, and implementation represent a major advancement in document accessibility and preservation. The technology has the potential to completely transform the digitization, translation, and study of historical documents as it develops further, protecting priceless cultural material for coming generations. This creative method not only updates conventional OCR systems but also lays the groundwork for the upcoming wave of AI- driven document processing tools, increasing the accessibility, interactivity, and insight of historical knowledge for scholars, researchers, and the general public

REFERENCES

[1]R Pavithra, Dr R Asokan, K Baskar, “Improving Privacy and Dynamic Updation Using Ranked Search Over Outsourced Cloud Data” (2016)
[2]A.C. Ikegwu, U.B. Alozieuwa, and H.F. Nweke, “Cultural diversity in the digital age: Leveraging AI and technology to preserve and promote Nigeria’s cultural heritage”, 2023
[3]A. Raj, S. Sharma, J. Singh, and A. Singh, “Revolutionizing data entry: An in-depth study of optical character recognition technology and its future potential,” International Journal for, Academia.edu, 2023.
[4]N. Pradeep, D. Subramanian, and M.K. Ganapathy, “Digitizing India's ancient texts: AI for Tamil palm leaf manuscript preservation and accessibility,” Academia.edu, 2024
[5]A. Taj and B. Gala, “Digitization projects for cultural heritage materials: A study with special reference to Arabic, Persian, and Urdu manuscripts,” AI-Assisted Library Reconstruction, IGI-Global, 2024.
[6]S. Frincu and M. Frincu, “Enabling interdisciplinary learning and research through AI-driven transliteration of historical documents,” International Journal of Humanities, ProQuest, 2024
[7]R. Harish and G.N. Raghavendra Rao, “Transcription of ancient Indian manuscripts through artificial intelligence—Current status of technology and the way forward,” Springer, 2023.
[8]S. Dixit, “Generative AI-powered document processing at scale with fraud detection for large financial organizations,” TechRxiv Preprints, 2024.
[9]O.I. Onianwa, “Mediation of artificial intelligence in digitizing bibliographies of secondary sources in historical research,” SSRN, 2024
[10]Y. Zhang, “Use of artificial intelligence (AI) in historical records transcription: Opportunities, challenges, and future directions,” McGill eScholarship, 2023
[11]S. Gupta, S. Mondal, & M. Ghosh, “Historical Marathi Handwritten Text Digitization Using AI,” Academia.edu, 2023.
[12]P. Dutta & N.B. Muppalaneni, “A Top-Down Character Segmentation Approach for Assamese and Telugu Handwritten Documents,” Journal of Ambient Intelligence and Humanized Computing, Springer, 2024.
[13]Y. Jamal, S.J. Neemuchwala, & R. Kouatly, “Handwritten Alphabet Recognition Using Convolutional Neural Networks,” IEEE International Conference on Artificial Intelligence and Informatics (ICAI), 2024.
[14]H. Patel, R. Joshi, & A. Mehta, “AI-Based OCR for Gujarati Handwritten Text Recognition,” International Journal of Computer Applications, 2024.
[15]J. Wang & L. Chen, “Machine Learning Approaches for Historical Document Digitization in South Asian Languages,” Computational Linguistics Journal, 2023.
[16]B. Singh & K. Sharma, “Preserving Indigenous Scripts: AI OCR for Historical Tamil and Telugu Manuscripts,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[17]A. Khan & N. Akram, “A Novel CNN Approach for Optical Character Recognition in Urdu Historical Texts,”

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403142|

Pattern Recognition Letters, 2024.
[18]L. Fernández & P. González, “Spanish Historical Document Recognition Using Transformer-Based OCR Models,” Journal of Artificial Intelligence Research, 2024
[19]C. Müller & D. Schmidt, “Advancements in Optical Character Recognition for Old German Scripts,” Digital Scholarship in the Humanities, 2023
[20]M. Roy & P. Chatterjee, “Bengali Handwritten Manuscript Digitization Using AI-Based OCR,” Journal of Computational Linguistics, 2024
[21]K. Park & H. Lee, “Korean Hanja Historical Text Recognition Using Deep Learning OCR,” Asia- Pacific Digital Humanities Journal, 2023
[22]J. Müller & F. Becker, “Improving Optical Character Recognition for Fraktur and Gothic Scripts,” Digital Humanities Research, 2023
[23]M. Swedhaa, V. Sathya, M. Vanitha and T. Jothilakshmi, "A Secure Approach to Avoid Data Repetition in Cloud Storage Systems, 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications
[24]Dhanabal. S, Baskar. K, Sangeetha. S and Umarani. B, "Handwritten Digits Recognition from Images using Serendipity and Orthogonal Schemes," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022,pp.139-141

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details