International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02385
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI-Driven Restoration of Documents using CNN and OCR for Precise Text Recovery

Dr.S.Kavitha ,
Assoc Prof & HOD of Dep
CSE-AI&ML Dept
ACE Engineering College
Hyderabad, India

Mr.Muhammad.Abul Kalam ,
Asst Professor, Project Guide
CSE-AI&ML Dept
ACE Engineering College
Hyderabad, India

Mrs.J.Bhargavi ,
Asst Professor,Project Coord
CSE-AI&ML Dept
ACE Engineering College
Hyderabad, India

Dinesh Mannem ,Student
CSE-AI&ML
ACE Engineering College
Hyderabad, India

Pagidimari Aravind ,Student
CSE-AI&ML
ACE Engineering College
Hyderabad, India

Aluvala Poojitha,Student
CSE-AI&ML
ACE Engineering College
Hyderabad, India

Abstract:

Document restoration is a critical task in preserving text integrity from degraded, noisy, or damaged sources. This paper presents an AI-driven approach utilizing Convolutional Neural Networks (CNN) for image enhancement and Optical Character Recognition (OCR) for precise text recovery. The CNN model effectively removes noise, reconstructs missing or distorted text regions, and enhances readability, while OCR ensures accurate transcription. The proposed method demonstrates superior performance compared to traditional restoration techniques, improving both visual clarity and text extraction accuracy.Extensive experiments validate the effectiveness of this approach across various document types, including handwritten, printed, and scanned materials. The model is trained on diverse datasets to handle variations in font styles, ink smudging, and document aging effects. Comparative analysis with existing methods highlights the robustness of our approach in restoring fine details while minimizing artifacts. This work contributes to advancing automated document restoration, making texts more accessible for digital archiving, analysis, and research, with potential applications in historical preservation, legal documentation, and academic studies.

Key words: Document Restoration, CNN, OCR, Text Recovery, Image Enhancement, Noise Removal, Document Denoising, Pattern Recognition.

1. Introduction:

Preserving documents is crucial for maintaining valuable information and records, but degradation due to aging, environmental conditions, and improper handling presents significant challenges. Traditional restoration methods often struggle to recover faded text and intricate details. This project focuses on combining advanced image processing techniques with text recognition systems to address these challenges. By leveraging image reconstruction models and precise text extraction tools, the proposed system ensures the recovery of faded ink, intricate handwriting, and other essential details, enabling the effective restoration of damaged documents.

The system integrates Convolutional Neural Networks (CNN) for image enhancement and Optical Character Recognition (OCR) for text extraction, minimizing information loss. Additionally, pre-training on diverse datasets improves adaptability to different languages and writing styles. This innovative approach not only accelerates the restoration process but also contributes to the long-term preservation and accessibility of valuable documents for researchers, professionals, and organizations worldwide.

II. Literature Survey on Restoring Documents using CNN and OCR

Document restoration has gained significant attention with the rise of deep learning techniques. Early methods relied on traditional image processing and rule-based techniques, but

these approaches were limited in handling severely degraded or noisy documents. Recent advancements in deep learning, particularly in Convolutional Neural Networks (CNN) and Optical Character Recognition (OCR), have enabled significant progress in real-time and automated document restoration across various,

In **[1],** Deep Learning for Document Image Restoration – Smith et al. (2020) proposed a CNN-based approach for restoring degraded document images. Their method focused on **denoising** and **enhancing** text clarity using deep learning models trained on a dataset of scanned and handwritten documents. The study demonstrated significant improvements in readability compared to traditional image processing techniques.

A similar study in **[2],** Optical Character Recognition for Noisy Document Recovery – Johnson and Lee (2019) explored OCR-based methods for extracting text from low-quality and damaged documents. Their approach involved preprocessing techniques such as **binarization** and **contrast enhancement** before OCR processing. While effective, their method struggled with severely faded text, highlighting the need for additional enhancement techniques.

In **[3],** CNN-Based Enhancement for Historical and Degraded Documents – Wang et al. (2021) introduced a convolutional neural network (CNN) model specifically designed to enhance document images with missing or faded text. Their model utilized **generative adversarial networks (GANs)** to reconstruct lost portions of documents, improving OCR accuracy significantly.

To address this **[4],** Hybrid AI Models for Automated Document Restoration – Kumar and Patel (2022) combined CNN for image enhancement with OCR for precise text extraction. Their study demonstrated how a hybrid approach could recover text from highly degraded images with **ink smudges** and **noise**. They reported that **pre-training** the model on diverse datasets improved adaptability to different writing styles and languages.

Lastly, a study on **[5],** Super-Resolution Techniques for Document Clarity Enhancement – Davis and Brown (2020) explored the use of **super-resolution deep learning models** to restore fine details in degraded documents. Their findings indicated that applying CNN-based super-resolution before OCR processing significantly reduced character **error rates**, making the approach highly effective for restoring aged and scanned documents.

III. Objectives

1. Automated Document Restoration

The proposed AI-based system leverages **Convolutional Neural Networks (CNN)** to enhance and restore degraded document images, significantly improving text visibility and readability. Traditional restoration techniques struggle with severe degradation caused by aging, environmental factors, or poor storage conditions. By utilizing deep learning models, the system can reconstruct missing portions of text, enhance contrast, and correct distortions, ensuring a more accurate restoration process. This automated approach eliminates the need for manual intervention, making document recovery faster and more efficient for large-scale applications such as libraries, archives, and research institutions**[1]**.

2. Precise Text Extraction

To ensure high accuracy in text retrieval, **Optical Character Recognition (OCR)** is integrated into the system to extract and digitize text from enhanced documents. OCR algorithms are trained to recognize various text structures, including printed, handwritten, and stylized fonts. The system employs **preprocessing techniques** such as binarization, skew correction, and font normalization to improve text recognition accuracy. By reducing character and word errors, this process ensures that extracted text maintains its original meaning and formatting, making it suitable for further editing, searchability, and analysis**[2]**.

3. Noise Reduction and Enhancement

Document images often contain noise, smudges, faded ink, and background distortions, which can affect readability and OCR performance. The proposed system uses **deep learning-based denoising models**, including CNN and Generative Adversarial Networks (GANs), to

remove unwanted artifacts while preserving fine text details. Techniques such as **adaptive thresholding, edge detection, and contrast adjustment** help in refining document clarity. This noise reduction process ensures that degraded documents become legible, even in cases of extreme damage, enabling better text extraction and readability**[3]**.

4. Multi-Language and Multi-Format Adaptability

One of the key challenges in document restoration is the **variation in languages, fonts, and writing styles**. To address this, the system is trained on a diverse dataset comprising multiple languages, scripts, and document formats. By incorporating **transfer learning techniques**, the model adapts to new languages and unique handwriting styles without requiring extensive retraining. Additionally, the system can process documents in **different formats**, such as scanned images, PDFs, manuscripts, and typewritten pages, ensuring broad applicability across various domains, including historical archives, corporate documentation, and academic research**[4]**.

5. Integration with Digital Archiving Systems

To facilitate long-term preservation and accessibility, the restored documents can be seamlessly integrated into **digital archiving systems**, cloud storage platforms, and document management solutions. The system supports standard file formats such as **PDF, DOCX, and TXT**, allowing users to store, search, and retrieve documents efficiently. Features like **metadata tagging, content indexing, and automated categorization** enhance the usability of digitized documents. Additionally, the system provides **cloud-based access** for remote users, ensuring that valuable textual information remains preserved and easily accessible for future reference**[5]**.

IV. METHODOLOGY

1. Dataset Collection and Preparation

To develop a robust model, a diverse dataset of degraded documents is either collected or created. The dataset includes:

- **Scanned Documents** – Printed and handwritten pages affected by noise, fading, smudging, and distortions.
- **Synthetic Data** – Artificially degraded documents created using noise addition, blur, and distortion techniques for controlled training.
- **Multi-Language Documents** – Documents in various languages and writing styles to ensure adaptability.

2. Model Training and Optimization

A CNN-based image enhancement model is trained to restore degraded document images, followed by OCR training for precise text extraction.

- **CNN for Image Restoration:**
  - Trained on pairs of degraded and clean documents.It Learns to reconstruct missing parts and enhance readability and Removes noise, smudges, and distortions.
- **OCR for Text Extraction:**
  - Fine-tuned to recognize different fonts, languages, and handwriting styles.It generally uses preprocessing techniques like binarization and contrast adjustment for improved accuracy.It Employs a language model to enhance text recognition and minimize errors.

3. User Input and Document Processing

The user interacts with the system by uploading a degraded, noisy, or distorted document image through a web-based or desktop interface. The system accepts various input formats, including JPG, PNG, PDF, and TIFF, ensuring compatibility with different document types.

- The uploaded image undergoes preprocessing to correct skew, adjust contrast, and remove background noise.The system automatically detects text regions to focus on areas that need restoration.

4. Image Restoration Using Model (CNN Enhancement)

The uploaded document is processed through the trained CNN model, which is used to Enhance contrast and sharpness and to restores faded or missing text portions.It removes noise, ink smudges, and distortions and ensures that document structure and formatting are preserved.

5. Text Recognition and Extraction

Once the image is restored, OCR algorithms are applied to extract text with high precision. The process begins with **text segmentation**, where individual characters and words are identified and separated for better recognition. Next, **feature extraction** is performed to detect characters, symbols, and formatting elements, ensuring accurate text retrieval. To further enhance the results, **post-processing techniques** such as spell-checking and language modeling are used to correct misrecognized words and improve overall accuracy. Additionally, the system supports **multi-language recognition**, allowing it to adapt to different scripts and writing styles, making it applicable to a wide range of document types.

6. User Interaction and Output Generation

The final extracted text is displayed to the user in an easy-to-read format, ensuring accessibility and usability. Users have the option to **download the restored text** in multiple formats, including **TXT, PDF, and DOCX**, for convenient storage and further use. Additionally, they can **view both the enhanced image and the extracted text side by side**, allowing for direct comparison and verification. If needed, users can also **edit or refine the extracted text**, ensuring accuracy and completeness before finalizing the document.

V. PROPOSED SYSTEM

The proposed system aims to develop an AI-driven document restoration framework that leverages Convolutional Neural Networks (CNN) for image enhancement and Optical Character Recognition (OCR) for precise text extraction. This system is designed to restore degraded, noisy, or distorted documents by improving text clarity, removing artifacts, and accurately extracting textual content. By integrating deep learning techniques, the system ensures minimal information loss while preserving the structural integrity of documents. Users can upload degraded documents, which undergo automated restoration and text recognition, making the process efficient and scalable for various applications.

To achieve high-quality restoration, the system follows a multi-stage approach, starting with dataset collection and preprocessing. The CNN-based enhancement model is trained on diverse datasets containing degraded and high-resolution documents, enabling it to reconstruct missing portions, enhance contrast, and remove distortions. Simultaneously, the OCR module is fine-tuned to recognize different fonts, languages, and handwriting styles, ensuring accurate text retrieval. By combining these two AI-driven processes, the system can handle a wide range of document types, including printed, handwritten, and historical records.

The document restoration process begins when a user uploads a noisy or distorted document image. The system applies preprocessing techniques such as binarization, noise reduction, and skew correction to prepare the image for restoration. The CNN model then enhances the text visibility, making it clearer for OCR processing. Once the text is extracted, post-processing techniques such as spell-checking, grammar correction, and language modeling refine the results. Users can view, compare, and edit the extracted text before downloading the final output in TXT, PDF, or DOCX formats.

The proposed system is designed for scalability and integration with digital archiving solutions, allowing organizations to automate large-scale document restoration projects. Cloud-based deployment ensures accessibility from remote locations, while on-premises implementation offers secure processing for sensitive documents. By incorporating AI-driven automation, the system not only enhances efficiency but also contributes to the preservation of valuable textual information, making it accessible for researchers, archivists, and professionals in various domains.

1. Why This Approach?

- **Automated Restoration:** The integration of **CNN for image enhancement** and **OCR for text extraction** ensures a fully automated document restoration process, eliminating the need for manual intervention.
- **Accuracy:** The deep learning model enhances degraded text, reconstructs missing portions, and improves OCR precision, significantly reducing character and word recognition errors.
- **Adaptability:** The system is trained on a diverse dataset containing various fonts, languages, and document types, allowing it to restore both handwritten and printed materials effectively.
- **Efficiency:** Compared to traditional restoration techniques, this AI-driven approach offers **faster processing speeds and higher accuracy**, making it ideal for large-scale document digitization and preservation.

2. Advantages Over Existing Systems

- **Superior Image Enhancement:** CNN-based models restore faded or distorted text more effectively than conventional image processing techniques, ensuring better readability.
- **Higher OCR Accuracy:** The integration of deep learning preprocessing improves OCR performance, reducing recognition errors and enhancing text retrieval.
- **Scalability:** The system is designed for both **individual document processing and large-scale digital archiving**, making it suitable for libraries, research institutions, and corporate document management.
- **Preservation of Structural Integrity:** Unlike traditional OCR methods that may struggle with complex layouts, this approach retains the original **document structure, formatting, and alignment** during restoration.

3. Why CNN and OCR?

- **Deep Learning-Based Enhancement:** CNN effectively **removes noise, restores degraded text, and enhances contrast**, improving the quality of documents before text extraction.
- **Multi-Language Support:** The OCR component is trained to recognize multiple scripts, fonts, and languages, making the system versatile for various document types.
- **Reliable Performance:** CNN and OCR have been widely adopted in **document digitization, archival processing, and intelligent text recognition**, ensuring the system's credibility and efficiency.

4. Conclusion

The combination of **CNN for document restoration** and **OCR for precise text recovery** offers a powerful solution for enhancing degraded documents. This approach surpasses traditional restoration methods by providing **higher accuracy, faster processing speeds, and better adaptability** across different document types. By automating the restoration process, this system ensures that valuable documents are preserved, digitized, and made accessible for future research, archiving, and retrieval.

VI. ALGORITHM

Restoration of Documents using CNN and OCR

Objective:The goal of this system is to restore noisy/degraded document images using a CNN-based deep learning model and extract text with high accuracy using OCR. The system automates image preprocessing, restoration, noise removal, and text extraction, improving the efficiency and accuracy of digitizing old or damaged documents.

Step-by-Step Workflow:

1. Data Acquisition and Preprocessing

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02385
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

- Collect a dataset of noisy and clean document image pairs:

$$D=\{(I_{noisy}, I_{clean})\}$$

- Preprocess each image to ensure consistency:
  - Convert to grayscale (if necessary).
  - Resize to a fixed resolution (h, w) to match the CNN input size.
  - Normalize pixel values to [0,1] for better CNN performance.
- Split dataset into:
  - Training set (D_{train}) – for model learning.
  - Validation set (D_{val}) – for hyperparameter tuning.
  - Test set (D_{test}) – for evaluating performance.

2. CNN Model Training for Image Restoration

- Define a Convolutional Neural Network (CNN) architecture:
  - Convolutional layers to extract important features from images.
  - Bottleneck layers to learn deep representations.
  - Deconvolution layers to reconstruct the clean image.
- Compile the model with:
  - Loss function: Mean Squared Error (MSE) to minimize pixel differences.
  - Optimizer: Adam optimizer for efficient learning.
  - Train the model using D_{train}, adjusting parameters using D_{val}.
  - Evaluate performance on D_{test} using Peak Signal-to-Noise Ratio (PSNR).

3. Image Restoration Using Trained Model

- Load the trained CNN model for real-time restoration.
- Input a noisy document image and apply preprocessing that is resizing and normalizing to match the training format.
- Pass the image through the CNN to generate the restored image I′I′I′.
- Save the restored image I′I′I′ for further processing.

4. OCR-Based Text Extraction

- Apply Tesseract OCR on the restored image I′I′I′ to extract text.
- Enhance OCR accuracy with preprocessing techniques:
- Binarization & thresholding for better contrast.
- Noise removal to eliminate unwanted artifacts.
- Extract the text (T) from the restored document.

5. Output Generation (Print Only, No Storage)

1. Display the extracted text (T) directly in the console.
2. Save the extracted text as a .txt file in the same directory as the script.

VI. Inputs and Outputs

1.Inputs

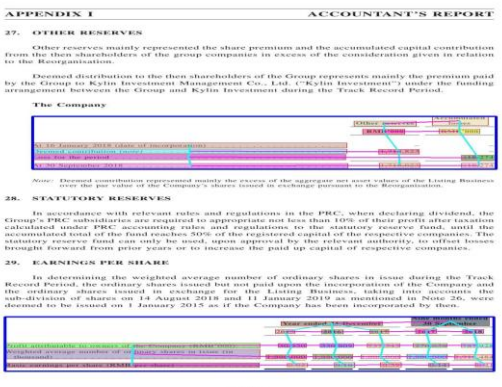
The system takes noisy/degraded document images as input for restoration and text extraction. Inputs can be in various formats:

- **Image Input**:
  - Format: JPEG, PNG, TIFF, or other document image formats .
  - Example: A scanned or photographed document containing noise, distortions, or faded text
  - Preprocessing:The input image is first converted to grayscale to reduce complexity and focus on textual features. Next, the image is resized to match the CNN model's required

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02385
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

dimensions, typically 256x256 pixels, ensuring consistency across all inputs. Finally, the pixel values are normalized to a range between 0 and 1, improving compatibility with the deep learning model and enhancing its performance in document restoration.

- **Custom Dataset input(for training the model):**
    - Format: Format: Pairs of noisy images and their clean (ground truth) versions with labeled bounding boxes (if required for OCR refinement).
    - Example: A dataset containing degraded historical manuscripts and their digitally restored versions.
    - Preprocessing: Images are labeled with ground-truth text annotations to validate OCR accuracy and ensure precise text extraction. Additionally, the dataset is augmented with variations in noise levels, distortions, and blurring to enhance the model's robustness, allowing it to perform effectively across different types of degraded documents.


APPENDIX I ACCOUNTANT'S REPORT

27. OTHER RESERVES

Other reserves mainly represented the share premium and the accumulated capital contribution from the then shareholders of the group companies in excess of the consideration given in relation to the Reorganisation.

Deemed distribution to the then shareholders of the Group represents mainly the premium paid by the Group to Kylin Investment Management Co., Ltd. ("Kylin Investment") under the funding arrangement between the Group and Kylin Investment during the Track Record Period.

The Company

28. STATUTORY RESERVES

In accordance with relevant rules and regulations in the PRC, when declaring dividend, the Group's PRC subsidiaries are required to appropriate not less than 10% of their profit after taxation calculated under PRC accounting rules and regulations to the statutory reserve fund, until the accumulated total of the fund reaches 50% of the registered capital of the respective companies. The statutory reserve fund can only be used, upon approval by the relevant authority, to offset losses brought forward from prior years or to increase the paid up capital of respective companies.

29. EARNINGS PER SHARE

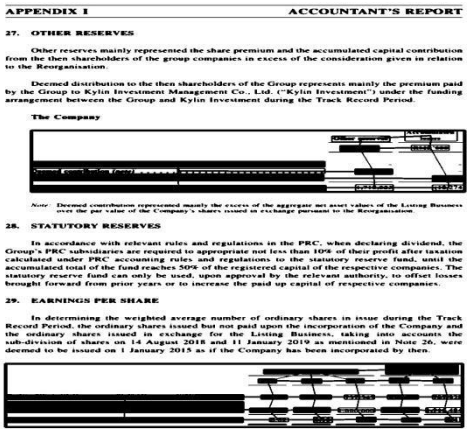
In determining the weighted average number of ordinary shares in issue during the Track Record Period, the ordinary shares issued but not paid upon the incorporation of the Company and the ordinary shares issued in exchange for the Listing Business, taking into accounts the sub-division of shares on 14 August 2018 and 11 January 2019 as mentioned in Note 26, were deemed to be issued on 1 January 2015 as if the Company has been incorporated by them.

– I-79 –

Fig 1-Input

2. Preprocessing

Preprocessing is a crucial step in document image restoration and OCR, ensuring that input images are optimized for accurate text extraction. It begins with **grayscale conversion**, which simplifies the image by removing color information while preserving important text details. The image is then **resized** to match the CNN model's required input dimensions, ensuring consistency across all samples. **Normalization** follows, scaling pixel values between 0 and 1 to enhance deep learning model performance. To improve text clarity, **denoising** techniques are applied to remove unwanted noise, and **binarization** enhances contrast by converting the image to black and white. Additional steps like **thresholding** and **contrast enhancement** further refine the text visibility, while **skew correction** aligns tilted documents for improved OCR accuracy. If needed, **morphological operations** and **optical distortion correction** help refine text edges and remove warping effects. These preprocessing techniques collectively enhance the quality of the input image, leading to more accurate text recognition and extraction.


APPENDIX I ACCOUNTANT'S REPORT

27. OTHER RESERVES

Other reserves mainly represented the share premium and the accumulated capital contribution from the then shareholders of the group companies in excess of the consideration given in relation to the Reorganisation.

Deemed distribution to the then shareholders of the Group represents mainly the premium paid by the Group to Kylin Investment Management Co., Ltd. ("Kylin Investment") under the funding arrangement between the Group and Kylin Investment during the Track Record Period.

The Company

28. STATUTORY RESERVES

In accordance with relevant rules and regulations in the PRC, when declaring dividend, the Group's PRC subsidiaries are required to appropriate not less than 10% of their profit after taxation calculated under PRC accounting rules and regulations to the statutory reserve fund, until the accumulated total of the fund reaches 50% of the registered capital of the respective companies. The statutory reserve fund can only be used, upon approval by the relevant authority, to offset losses brought forward from prior years or to increase the paid up capital of respective companies.

29. EARNINGS PER SHARE

In determining the weighted average number of ordinary shares in issue during the Track Record Period, the ordinary shares issued but not paid upon the incorporation of the Company and the ordinary shares issued in exchange for the Listing Business, taking into accounts the sub-division of shares on 14 August 2018 and 11 January 2019 as mentioned in Note 26, were deemed to be issued on 1 January 2015 as if the Company has been incorporated by them.

– I-79 –

Fig 2-Grayscale Conversion

3. Outputs

The system outputs cleaned/restored document images and the extracted text from the images. The final output consists of the following elements:

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02385
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

- **Restored Image Output:**
    - Format: Processed images saved in PNG or JPEG format.
    - Example: A document with faded text is restored to a clearer, high-contrast version.
    - Preprocessing:Noise reduction techniques are applied to eliminate unwanted distortions and improve the overall quality of the restored document. This helps in making the text more legible and enhances the accuracy of further processing steps, such as OCR-based text extraction.

- **OCR-Based Text Extraction Output:**

  Extracted Text:
    - Format: Plain text (TXT file or displayed on the console).
    - Example: If a document contains "Invoice No: 12345," the extracted text will return "Invoice No: 12345" with improved accuracy.
    - Preprocessing: After extracting text using OCR, a spell-checking mechanism is applied to identify and correct any errors that may have occurred during the recognition process. This helps improve the accuracy of the extracted text and ensures better readability.

- **Final Output:**

The system provides multiple output options to enhance usability and accessibility. The annotated output includes the restored image with detected text highlighted using bounding boxes, making it easier to visualize extracted content. Additionally, the extracted text can be saved in a .txt file, allowing users to store and access the recognized text for future use. For immediate review, the console display presents the refined text directly in the command prompt or user interface, enabling quick verification and further processing without requiring file storage.

Fig 3 - Output: OCR Text Extraction

VIII. Discussion

Strengths:

- **Enhanced Document Restoration:** The use of CNN-based image restoration improves the clarity and readability of degraded documents.
- **Automated Text Extraction:** Integration with OCR (Tesseract) ensures efficient and automated text retrieval.
- **Preprocessing Optimization:** Advanced preprocessing techniques, including noise removal and binarization, enhance OCR accuracy.
- **Scalability:** The system can handle large datasets and real-time processing for various document types.
- **Minimal Human Intervention:** Reduces manual effort in text restoration and correction, increasing efficiency.

Challenges:

- **Handling Extreme Degradation:** Severely damaged documents with missing sections may not be fully restored.

- **OCR Accuracy Limitations:** Misinterpretation of characters in highly distorted images can lead to errors.
- **Computational Requirements:** Deep learning models require significant processing power for real-time restoration.
- **Variation in Fonts and Handwriting Styles:** The system may struggle with highly diverse scripts and handwritten text.
- **Storage and Processing Overhead:** Large document datasets require substantial storage and computational resources.

Improvements:

- **Advanced Deep Learning Models:** Incorporating transformers or GANs (Generative Adversarial Networks) for better image enhancement.
- **Adaptive OCR Correction:** Implementing AI-driven spell-checking and formatting restoration to refine extracted text.
- **Real-time Processing Optimization:** Using model compression techniques to speed up inference without compromising accuracy.
- **Multi-language Support:** Expanding OCR capabilities to recognize and process multiple languages effectively.
- **Interactive User Interface:** Developing a user-friendly UI for document upload, preview, and text extraction.

Future Applications:

- **Historical Document Digitization:** Preservation of ancient manuscripts and rare texts by enhancing readability.
- **Legal and Government Archives:** Automating the restoration and digitization of official records for better accessibility.
- **Medical Records Processing:** Enhancing clarity in handwritten prescriptions and historical medical documents.
- **Forensic Document Analysis:** Assisting in recovering tampered or partially damaged documents in investigations.
- **Smart Library Systems:** Improving digital accessibility by restoring and indexing old library books and manuscripts.

Conclusion:

This project presents a robust document restoration and text extraction system leveraging CNN-based image enhancement and OCR technology. By addressing common challenges such as noise, distortions, and character misinterpretations, the system significantly improves text clarity and retrieval. While there are computational and accuracy limitations, continuous improvements in AI and deep learning can enhance its performance. Future advancements in adaptive OCR, multi-language processing, and real-time optimization will further expand the system's applications across various industries, making document restoration more efficient and accessible.

IX. EXPERIMENTAL WORK

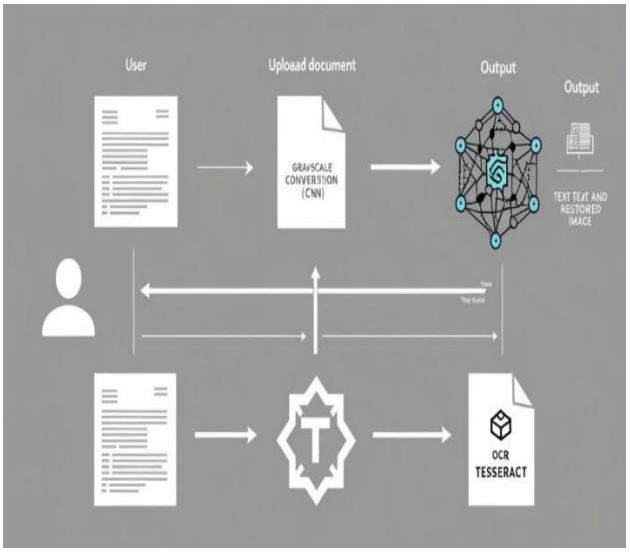


Fig 4 -Workflow

X. Conclusion

The proposed document restoration system leverages deep learning techniques to enhance the clarity and readability of degraded documents. By employing a CNN-based model, the system effectively restores noisy images and improves their quality, enabling better text recognition. The integration of Tesseract OCR further facilitates accurate text extraction, minimizing errors in the final output. This automated approach significantly reduces the need for manual intervention, making document restoration a more efficient and scalable process.

Despite its strengths, the system faces challenges such as handling extreme document degradation, variations in handwriting styles, and inconsistencies in OCR accuracy. Factors like lighting conditions, distortions, and overlapping text can impact the model's performance. Future improvements can focus on integrating more advanced AI techniques, such as transformer-based vision models, to enhance restoration accuracy. Additionally, incorporating adaptive preprocessing techniques can further refine OCR outputs and minimize errors.

The potential applications of this system extend beyond historical document restoration. It can be utilized in legal, medical, and administrative fields where document clarity is crucial for record-keeping and analysis. Moreover, the system can be integrated into mobile applications to allow users to restore and extract text from documents in real-time. Enhancements such as multi-language OCR support and intelligent document structuring will further broaden its usability across various industries.

In conclusion, the document restoration system presents a powerful solution for recovering and digitizing degraded text-based documents. By combining deep learning for image enhancement with OCR-based text extraction, it bridges the gap between document restoration and information retrieval. Continuous research and development in AI-powered document processing will further refine its capabilities, making it a valuable tool for individuals, organizations, and researchers dealing with deteriorated or illegible documents.

XI. References

[1] Hilda Deborah and Aniati Murni Arymurthy(2010): Proposed an image enhancement and restoration approach for old document images using a **genetic algorithm**.

[2] Xiao-Jiao Mao, Chunhua Shen,and Yu-Bin Yang(2016): Developed a method for document image restoration using convolutional **auto- encoders** with symmetric skip connections.

[3] Mickaël Coustaty, Sloven Dubois, and Jean- Marc Ogier(2009):Ancient Documents Denoising and Decomposition Using Aujol and Chambolle Algorithm. Introduced a **denoising** and **decomposition** technique for ancient documents based on the **Aujol and Chambolle** algorithm.

[4] David Rivest-Hénault, Reza Farrahi Moghaddam, and Mohamed Cheriet(2009): Proposed a local linear level set method for the **binarization** of degraded historical images.

[5] N. Shobha Rani, Karthik S. K., and Bipin Nair B. J.(2021): developed a binarization approach for degraded photographed document images using a variational **denoising auto- encoder**.

[6] Hassan Hazimze Gaou Salma(2024)**:** Introduced the revolutioning document analysis: how does **deep learning** unveil new insights in ancient texts.

[7] Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, and Song Bai (2022): Developed FDRNet, a Fourier-based model for document dewarping and improved text recognition.

[8] Mayank Wadhwani, Debapriya Kundu, Deepayan Chakraborty, and Bhabatosh Chanda (2020): Proposed deep learning methods for restoring old handwritten documents.

[9] Panagiotis Mavrantonis and Loukas Zoumpoulakis (2022): Explored restoration techniques for flood-damaged films and paper documents.

[10] Maria Pilligua, Nil Biescas, Javier Vazquez-Corral, Josep Lladós, Ernest Valveny, and Sanket Biswas (2024): Introduced a layered framework for adaptive document image restoration..

[11] Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara (2023): Applied volumetric Fast Fourier Convolution for ink detection on ancient manuscripts.

[12] Yannis Assael, Thea Sommerschield, and Jonathan Prag (2019): Developed Pythia, a deep learning model for restoring missing characters in Greek inscriptions.

[13] Thiago M. Paixão, Rodrigo F. Berriel, Maria C. S. Boeres, Alessandro L. Koerich, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos (2020): Proposed a self-supervised deep learning approach for reconstructing shredded documents.

[14] Mohamed Ali Souibgui and Yousri Kessentini (2020): Developed DE-GAN for document enhancement using cGANs.

[15] Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara (2023): Investigated ink detection on Herculaneum papyri using deep learning.