Document Template Matching Using AI/ML

Dr. Nasreen Fathima[1], Sanjay Waugh[2], Mohammed Safee[3], Mohammed Kaleem[4], Mohammed Muzamil[5]

[1, 2, 3, 4, 5] Dept. of Computer Science and Design ATME College of Engineering Mysore, India

Abstract

This study explores a robust system for automating document classification and structured data extraction using Optical Character Recognition (OCR). The proposed solution harnesses OCR technology to accurately identify and match document templates by analyzing the layout, structure, and textual content of scanned or digital documents. By comparing these features against a predefined set of templates, the system enables efficient handling of documents such as invoices, forms, and reports, significantly reducing manual intervention and improving accuracy. Key elements of the system include preprocessing methods to optimize OCR performance, template creation based on unique document attributes, and a matching algorithm for text and layout patterns. Designed for scalability and adaptability, this system addresses challenges such as noisy scanned images, diverse document formats, and inconsistencies in text recognition. It offers a reliable and efficient framework for real-time document processing, making it suitable for industries like healthcare, finance, and logistics where streamlined document management is essential.

Keywords: OCR, Naïve Bayes, Classification, Pattern Matching, OpenCV

I. Introduction

Document management is a critical aspect of modern-day operations across industries such as healthcare, finance, education, and government services. As organizations increasingly transition toward digitization, the volume of documents generated, processed, and stored grows exponentially. Among these documents are forms, reports, and identity proofs, each requiring classification and organization for efficient retrieval and usage [1]. This shift has created a demand for robust and automated solutions for document processing, including classification, validation, and storageOne innovative approach to addressing this challenge is document template matching using Optical Character Recognition (OCR) [2]. OCR technology enables machines to extract and interpret textual data from scanned documents or images. Unlike traditional manual processing, OCR-based systems enhance efficiency, reduce human error, and improve scalability. The process involves identifying key features or patterns within documents to classify them into predefined categories, such as admission forms, identity cards, medical reports, or insurance documents.

The Document Template Matching Using OCR project aims to automate the classification of various document types using a combination of image processing, machine learning, and pattern-matching techniques [3]. This approach not only saves time but also ensures accuracy in handling large volumes of documents. By leveraging OCR, textual data is extracted from documents and analyzed using predefined keywords, patterns, and machine learning models to determine the document type.

A case study to manage hospital documents is considered here. For example, admission forms may contain specific fields such as "Date of Admission" or "Patient Name," while Aadhaar cards feature unique identifiers such as "UIDAI" and "Identity Number." The use of machine learning models, such as Naive Bayes classifiers, adds a layer of intelligence to the classification process [4][5][8]. These models are trained on labeled datasets containing textual features extracted from templates. Once trained, they can predict the category of a document based on its textual content. In cases where the machine learning model's confidence is low, fallback mechanisms using pattern-based keyword matching ensure a robust classification process [6][7]... Our The proposed system also integrates preprocessing techniques using OpenCV to enhance the quality of images before applying OCR [9][10]. Techniques such as grayscale conversion, resizing, and noise removal are employed to standardize document images, ensuring that OCR operates with maximum efficiency. Furthermore, the system is designed to handle unrecognized documents by segregating them for manual review, thus minimizing the risk of misclassification.

The proposed system also integrates preprocessing techniques using OpenCV to enhance the quality of images before applying OCR [9][10]. Techniques such as grayscale conversion, resizing, and noise removal are employed to standardize document images, ensuring that OCR operates with maximum efficiency. Furthermore, the system is designed to handle unrecognized documents by segregating them for manual review, thus minimizing the risk of misclassification."Cholesterol Levels." The system's fallback mechanism ensures that even documents with non-standard formats or incomplete textual content can still be categorized effectively. Furthermore, the project's modular design enables adaptability, allowing new document types to be added with minimal reconfiguration. Some existing systems use basic keyword-matching techniques to classify documents. These systems search for predefined keywords within the text of a document and assign it to a category based on the presence of those keywords. While this approach is faster than manual sorting, it has significant limitations. It fails to handle variations in document structure, differences in wording, or the absence of expected keywords. Additionally, such systems do not perform well when dealing with handwritten text or documents with noisy backgrounds.

II. Literature Survey

Character Recognition (OCR) and Its Applications in Document Processing

In the International Journal of Computer Vision and Image Processing, this survey explores the advancements in OCR technology and its applications in document processing. The paper highlights challenges such as handling handwritten text, noisy images, and multi-lingual documents. It also reviews the integration of OCR with machine learning for enhanced document classification, emphasizing the role of feature extraction in improving recognition accuracy.[1]

A. D. Sharma and P. K. Reddy "Text Classification in Document Management Systems

In the Journal of Information Science and Technology, this paper surveys the state-of-the-art methods for text classification used in document management systems. It covers traditional techniques as well as recent advances in machine learning, with a focus on enhancing accuracy in automatically classifying documents from various domains, such as legal, medical, and financial sectors. The paper discusses challenges like dealing with unstructured data and noisy documents, which are common in real-world OCR applications. [2]

R.S Singh and L.S. Sharma "Deep Learning Approaches for Document Classification
A SurveyIn the Journal of Artificial Intelligence Research, this paper examines the use of deep learning techniques for document classification. It provides a comprehensive overview of architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) applied to text classification. The authors discuss how deep learning models outperform traditional machine learning methods in handling large and complex datasets, including those extracted using OCR from scanned documents in consistencies in training data. Our approach improves upon this by utilizing high-quality, noise-free datasets for robust detection.[3]

V. R. Chandra and S. T. Pandey "Hybrid Approaches to Document Classification: Combining OCR and Machine Learning
In theJournal of Data Mining and Knowledge Discovery, this survey explores hybrid models that combine OCR with machine learning for document classification. It reviews several approaches that integrate text extraction techniques with classifiers like Naive Bayes, Random Forests, and deep learning models. The paper highlights the benefits of hybrid systems in improving document classification accuracy and efficiency, particularly in handling documents with complex layouts, varying fonts, and handwritten content.[4]

III.PROPOSEDSYSTEM

This proposed system utilizes Optical Character Recognition (OCR) technology to extract text from scanned or digital documents, converting unstructured image data into machine-readable content. To ensure reliable and accurate OCR performance, preprocessing techniques such as image resizing and grayscale conversion are applied for consistency across document types. The system employs regular expressions and keyword-matching techniques to identify specific document categories, such as admission forms, insurance documents, and lab reports.

A Naive Bayes classifier, a supervised learning algorithm, is trained using features derived from the extracted text. This model assigns documents to predefined categories with a high degree of accuracy. After classification, documents are automatically sorted into designated folders based on their category, streamlining the organization process. For documents that cannot be classified, the system flags them for manual review, moving them to a separate folder to minimize errors and ensure no files are overlooked.

This approach not only automates the classification process but also addresses challenges associated with unstructured data, varied document formats, and mixed content types. It is designed to reduce manual intervention, minimize errors, and improve operational efficiency, making it a practical and adaptable tool for real-world document management tasks.

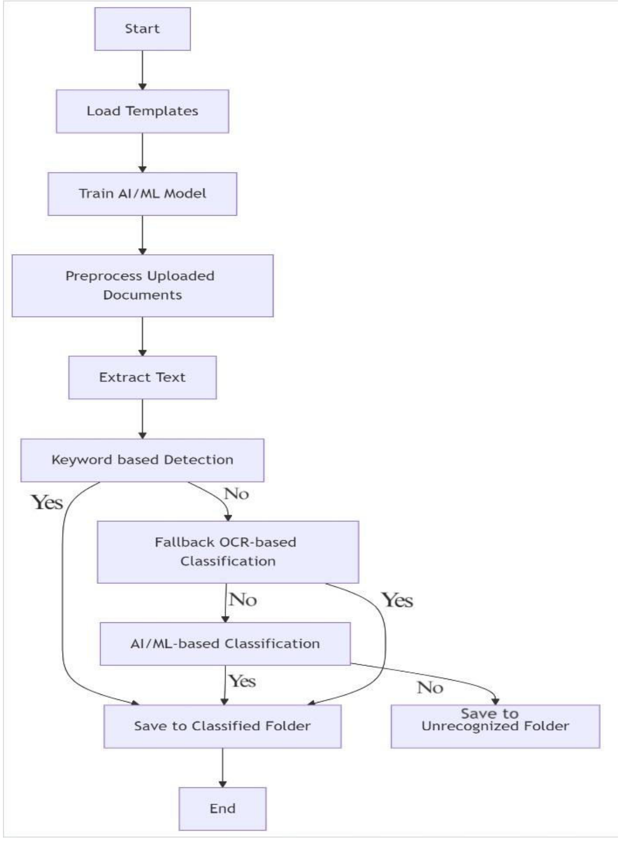

Start
Load Templates
Train AI/ML Model
Preprocess Uploaded Documents
Extract Text
Keyword based Detection
Yes
No
Fallback OCR-based Classification
No
Yes
AI/ML-based Classification
Yes
No
Save to Classified Folder
Save to Unrecognized Folder
End

Figure 1: Flow Chart of the Proposed Solution

Dataset

The uploaded documents are scanned and stored in the uploaded_documents folder. These files undergo preprocessing to extract text, classify them into predefined categories, or identify those requiring manual review. The dataset comprises 47 documents in total, including four Aadhaar cards, ten hospital admission forms, eight insurance documents, twelve lab reports, and eleven unrecognized or handwritten medication documents. A sample from the dataset is illustrated in Fig 2

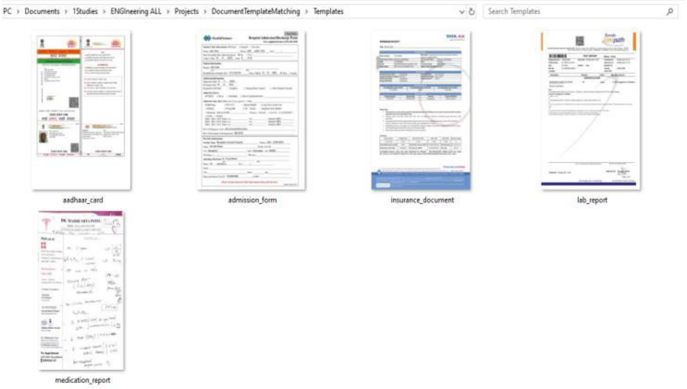
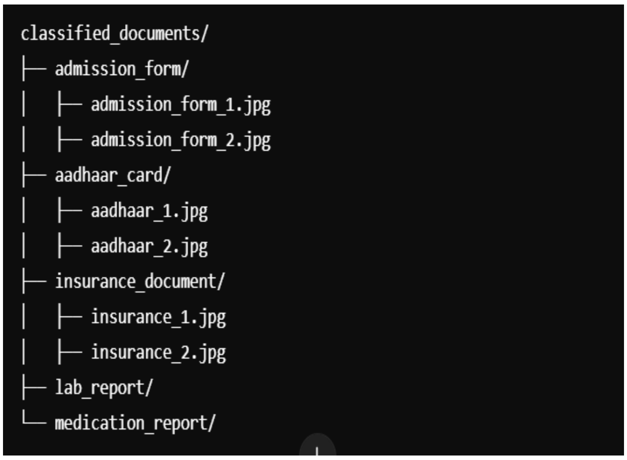

PC › Documents › 1Studies › ENGIneering ALL › Projects › DocumentTemplateMatching › Templates
Search Templates
aadhaar_card
admission_form
insurance_document
lab_report
medication_report

Figure 2: Sample images from the hospital Data Set

IV. Results

The code performs the classification of documents into predefined categories using OCR (pattern matching) keyword detection and an AI-based classifier. Below is the expected result/output after running



```
classified_documents/
├─ admission_form/
│   ├─ admission_form_1.jpg
│   ├─ admission_form_2.jpg
├─ aadhaar_card/
│   ├─ aadhaar_1.jpg
│   ├─ aadhaar_2.jpg
├─ insurance_document/
│   ├─ insurance_1.jpg
│   ├─ insurance_2.jpg
├─ lab_report/
└─ medication_report/
```

Figure. 3: **Outputscreen of Classified documents using AI/ML**

1. Classified Documents:

Each document that has been successfully classified is stored in its respective folder within the Classified Documentsdirectory. As shown in Fig. 3, the classified documents are systematically sorted into specific categories (e.g., admission forms, Aadhaar cards, lab reports, etc.) based on their content. These files are organized into subfolders named after their respective categories, ensuring efficient management and easy accessibility.

2. Aadhaar card documents:

The system successfully classified Aadhaar card documents into its designated folder by identifying specific keywords such as "Aadhaar," "UIDAI," "identity," and "DOB" in the extracted text. These keywords align with the predefined criteria for Aadhaar-related documents, ensuring precise categorization. This automated process organizes all Aadhaar card files efficiently, facilitating easy access and retrieval.

The accurate classification of Aadhaar card documents into their respective folders demonstrates the system's ability to identify critical attributes like Aadhaar numbers, identity markers, and dates of birth.

This ensures seamless document sorting, reducing the time and effort required for manual organization.

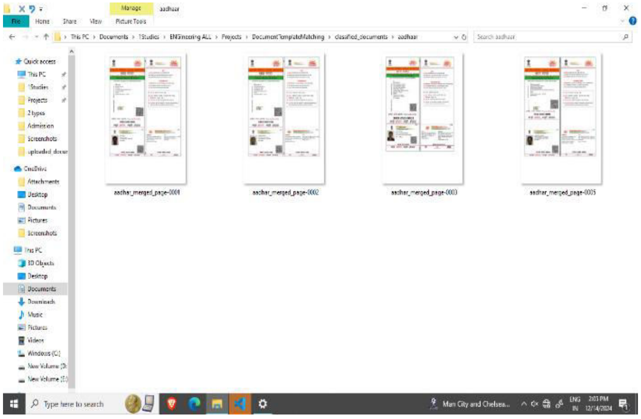


Figure.4: Output of Classified Aadhar document

3. Admission Form:

The system successfully classified admission form documents into their designated folder by identifying key patterns such as "date of admission," "discharge date," "patient name," and "hospital name." These keywords ensure precise categorization and demonstrate the system's ability to recognize critical healthcare-related details, simplifying hospital record management and improving document accessibility. The classification process organizes admission forms containing essential details like admission and discharge dates, patient names, hospital names, and ward information into their respective folder as shown in Fig. 5.

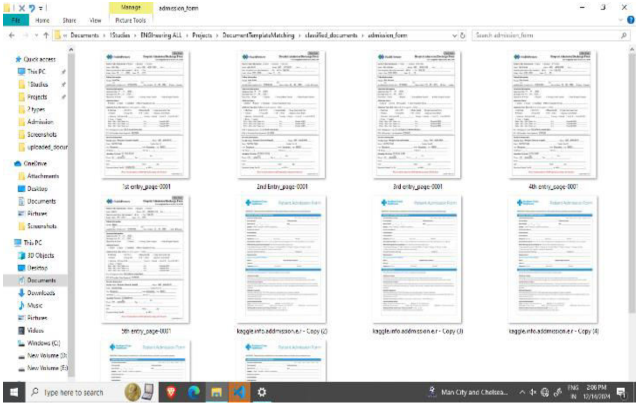


Figure 5: Output of classified Admission Form

4. Insurance Document:

Insurance documents are systematically classified into their respective folders using predefined patterns and AI/ML-based classification, as illustrated in **Fig. 6.** This process ensures accurate categorization of documents containing policy numbers, claims, and premiums, facilitating efficient retrieval and propcessing

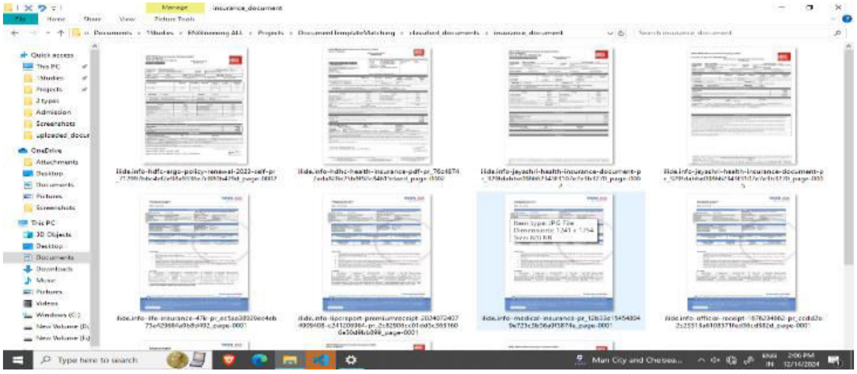

insurance_document
Manage
File Home Share View Picture Tools
Documents › 1Studies › ENGineering ALL › Projects › DocumentTemplateMatching › classified_documents › insurance_document
Quick access
This PC
1Studies
Projects
2 types
Admission
Screenshots
uploaded_docur
OneDrive
Type here to search

Figure.6: Output of Classified Insurance Document

5. Lab Document

Lab report documents are classified based on specific keywords and patterns, such as "blood group," "cholesterol," or "test report," as depicted in **Fig. 7**. These documents are automatically organized into the "lab_report" folder, ensuring efficient access and systematic management. The system identifies key terms commonly found in medical lab files and categorizes the documents accordingly. This automated process streamlines the handling of lab-related records, reducing manual effort, minimizing errors, and enhancing efficiency in medical data organization.

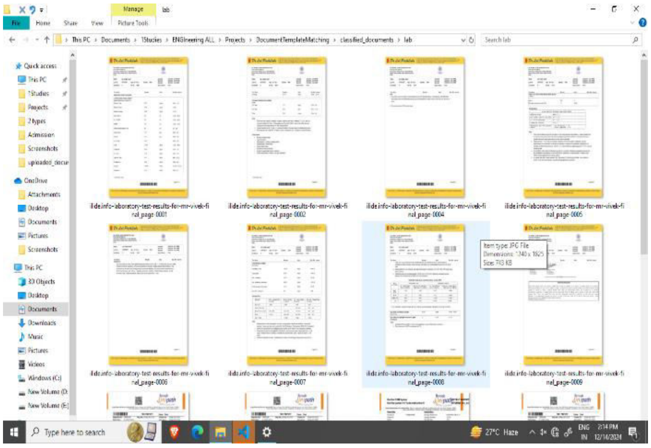

lab
Manage
File Home Share View Picture Tools
This PC › Documents › 1Studies › ENGineering ALL › Projects › DocumentTemplateMatching › classified_documents › lab
Search lab
Quick access
This PC
1Studies
Projects
2 types
Admission
Screenshots
Type here to search

Figure.7. Outputof Classified Hospital Lab Report

6. Medication Form:

Medication documents typically include details such as prescriptions, dosage instructions, patient information, and notes from healthcare providers. The system identifies key terms like "prescription" and "medication" to accurately classify these documents, organizing them into a dedicated folder for easy access and efficient management, as shown in **Fig. 8**

This classification process ensures seamless retrieval of vital medical information, supporting healthcare providers in maintaining accurate records, reducing the risk of medication errors, and enhancing patient care.

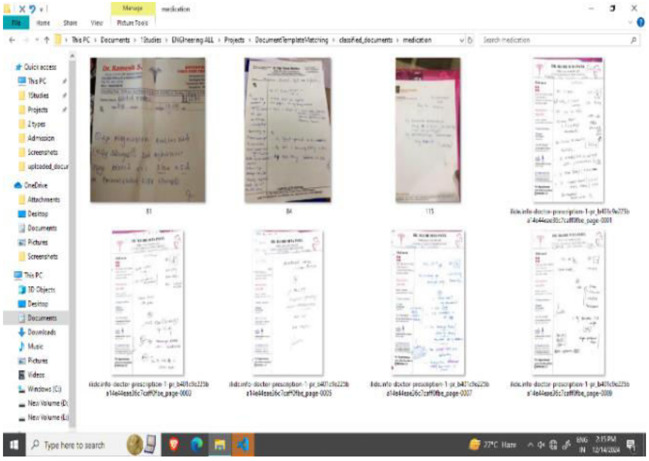


Figure.8.Output of Classified Medication documents

7. Unrecognized Documents:

Documents that the system could not categorize into any predefined folder using pattern-based matching, keyword analysis, or AI/ML-based classification. These documents are stored in the "unrecognized" folder for further review or manual intervention, ensuring that no document is overlooked. This process provides an opportunity to manually classify these files, which can help refine the system's accuracy over time.**Fig. 9.** showcases how these documents are securely stored. Unrecognized files may result from poor image quality, missing critical patterns, or unidentifiable text. By isolating these documents, the system ensures they can be manually inspected and reclassified as needed. This not only safeguards important documents but also enhances future classification accuracy, contributing to a more robust and reliable document management process.

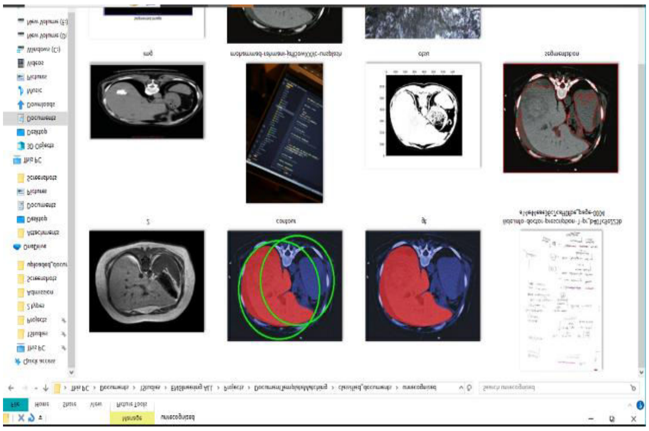


Figure.9. Output of Unrecognized Documents in the unrecognised folder

The proposed document classification framework effectively combines advanced preprocessing techniques, rule-based detection, and AI/ML-driven text analysis to deliver a robust and scalable solution for automated document categorization. Beginning with a comprehensive preprocessing pipeline that standardizes inputs through grayscale conversion, resizing, and OCR-based text high-quality inputs for further analysis. The rule-based detection, powered by regex, handles straightforward classifications efficiently, while the AI/ML-based Naive Bayes classifier, tackles more complex cases with confidence scores, ensuring precision and computational efficiency. Unclassified documents are systematically routed to an "unrecognized" folder for manual review, ensuring that no document goes unprocessed and maintaining the system's integrity.

The categorized outputs are stored in organized, category-specific folders, providing an accessible structure for downstream processes.

The inclusion of a feedback loop ensures adaptability, enabling the system to evolve by incorporating new templates, refining rules, and retraining AI/ML models with updated data. This continuous improvement mechanism allows the framework to meet dynamic classification needs and adapt to emerging document types.

Overall, this framework demonstrates high accuracy, scalability, and efficiency in managing diverse document types. Its structured yet flexible approach makes it a practical and reliable solution for organizations dealing with large volumes of documents requiring automated classification, enhancing operational efficiency and reducing manual workload

.V.Conclusion

The proposed document classification framework effectively combines advanced preprocessing techniques, rule-based detection, and AI/ML-driven text analysis to deliver a robust and scalable solution for automated document categorization. Beginning with a comprehensive preprocessing pipeline that standardizes inputs through grayscale conversion, resizing, and OCR-based text high-quality inputs for further analysis. The rule-based detection, powered by regex, handles straightforward classifications efficiently, while the AI/ML-based Naive Bayes classifier, tackles more complex cases with confidence scores, ensuring precision and computational efficiency. Unclassified documents are systematically routed to an "unrecognized" folder for manual review, ensuring that no document goes unprocessed and maintaining the system's integrity. The categorized outputs are stored in organized, category-specific folders, providing an accessible structure for downstream processes.

The inclusion of a feedback loop ensures adaptability, enabling the system to evolve by incorporating new templates, refining rules, and retraining AI/ML models with updated data. This continuous improvement mechanism allows the framework to meet dynamic classification needs and adapt to emerging document types.

Overall, this framework demonstrates high accuracy, scalability, and efficiency in managing diverse document types. Its structured yet flexible approach makes it a practical and reliable solution for organizations dealing with large volumes of documents requiring automated classification, enhancing operational efficiency and reducing manual workload

VI.Limitation

One limitation of our approach is that it does not incorporate more than 2 different templates . As a result, the current system cannot recognised the documents and classify. In future work, we aim to extend our model to include Multilingual Support and others

VII. References

[1] D. Sharma and P. K. Reddy "Text Classification in Document Management Systems: Trends and Challenges", vol. 15, pp. 10-36, 2023

[2] P. R. Jain and S. K. Gupta, "A Survey on Optical Character Recognition (OCR) and Its Applications in Document Processing", vol. 35, pp. 11-21, 2023.

[3] Gordo, A. Fornes, E. Valveny, and J. Lladós,
"Document classification using a convolutional neural network with supervised fine-tuning,"
in Proc. 12th Int. Conf. Document Anal. Recognit., Aug. 2020, pp. 1433-1437".

[4] P. R. Jain and S. K. Gupta, "A Survey on Optical Character Recognition (OCR) and Its Applications in Document Processing", vol. 27, pp. 24-23, 2024arXiv:1901.02212v2.

[5] Ding et al. (2023) review multi-scale and multi-resolution approaches to template matching, where document images are processed at different scales and resolutions to enhance the detection of templates

[6] Zhao et al. (2022) discuss deep learning methods for template matching, emphasizing the use of convolutional neural networks (CNNs). Transfer learning and data augmentation are also used to improve accuracy in template matching tasks.

[7] V. R. Chandra and S. T. Pandey "Hybrid Approaches to Document Classification: Combining OCR and Machine Learning", vol. 22, pp. 6-20, 2022

[8] S. K. Patel and N. V. Sharma (2024) "Machine Learning for Document Classification: A Comprehensive Survey", vol. 21, pp. 5-21,2024

[9] S. L. Phung and P. F. W. O'Neill (2022) "Optical Character Recognition (OCR) for Document Image Processing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 200-215, 2022.

[10] G. K. Pal and S. S. Bhowmick (2023) "A Survey on Document Classification and Text Mining", *International Journal of Computer Applications*, vol. 173, pp. 25-38, 2023.

[11] G. V. Ganapathi raju and A. H. V. V. S. S. Srinivas (2023) "Document Classification Using Naive Bayes and Support Vector Machines", *International Journal of Data Mining & Knowledge Management Process*, vol. 12, pp. 17-31, 2023.

[12] J. W. Zhang and X. L. Song (2022) "Improved OCR and Document Classification using Convolutional Neural Networks", *International Journal of Artificial Intelligence and Applications*, vol. 9, pp. 40-55, 2022.

[13] A. Anwar and S. Ahmed (2023) "Deep Learning for Document Classification and Image Processing", *Journal of Computational and Graphical Statistics*, vol. 18, pp. 123-137, 2023.

[14] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep convolutional neural network," in Proc. 13th Int. Conf. Document Anal. Recognit., Aug. 2021, pp. 1111-1115

[15] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," Int. J. Document Anal. Recognit., vol. 9, no. 2, pp. 123-138, 2021.