# Improving Document Digitization with Machine Learning-Based OCR

**P. Sri Charitha[1], P. Nithisha[2], P. Rahul[3], M.Ram Mohan[4], Haritha A[5]**

[1, 2, 3, 4, 5]Department of Information Technology, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, India

**Abstract**

**In today's digital era, the extraction of text from unstructured formats such as images, PDFs, and handwritten documents is critical for digitization and automation. Traditional methods often struggle with scalability, complex layouts and multi-language support. This project addresses these challenges by leveraging Machine Learning, Optical Character Recognition (OCR), AWS Textract model and microservices architecture to create a robust, scalable, and efficient text extraction system.**

**The proposed solution integrates advanced technologies such as Java Spring Boot for backend development, PostgreSQL for secure data storage, and containerized microservices for enhanced modularity and scalability. The system performs preprocessing to improve image quality, employs deep learning algorithms for accurate text recognition. Parallel processing and task queuing ensure high throughput and low latency for real-time and bulk operations.**

**By converting unstructured data into structured like JSON or CSV ,this system facilitates seamless integration into existing workflows. This study highlights the design, functionality, and benefits of this innovative approach to text extraction, driving efficiency in document management and automation.**

**Keywords: Text Extraction, Optical Character Recognition (OCR), Image Processing, AWS Textract, PostgreSQL**

## I. Introduction

In the digital environment, organizations and individuals alike manage an ever-increasing volume of documents in diverse formats, including images, scanned PDFs, and handwritten notes. Extracting useful text from these materials is crucial for digitization, data analysis, and process automation. However, traditional methods often fall short when dealing with large-scale operations, complex document layouts, and multi-language content. This underscores the necessity of a solution that combines efficiency, precision, and adaptability to meet the demands of modern data management.

The proposed system leverages state-of-the-art technologies, including Machine Learning, Optical Character Recognition (OCR), and microservices architecture, to create a scalable and modular framework for extextraction. Java Spring Boot serves as the foundation for back end development, providing aro bust and expandable platform. PostgreSQL ensures secure and efficient data storage, while containerized microservices enable seamless scaling and resilience. The system employs advanced preprocessing techniques to enhance input quality, deep learning models for accurate recognition , and supports structured content such as tables and forms. Multi-language compatibility and handwritten text recognition further enhance its versatility.

Designed for real-world application, the system incorporates parallel processing and task queuing to maintain high throughput and low latency, making it suitable for both real-time and bulk operations. By converting unstructured data into structured formats like JSON or CSV, it enables smooth integration with existing workflows and systems. Additionally, the solution prioritizes security and compliance, offering encrypted data handling and role-based access control.This innovative approach redefines text extraction, providing a reliable and scalable solution that meets the demands of modern document management and automation.

## II. Literature Survey

| **Authors** | **Title of the paper** | **Approach** | **Contributions** |
|---|---|---|---|
| Raisietal.(2014) | *A Survey on Various Approaches of Text Extraction in Images* | Explores both traditional and machine learning techniques for text extraction. | Provides a detailed comparison of various techniques, highlighting their strengths and limitations. |
| Staar,Dolfi,Auer,and Bekas(2018) | *Corpus Conversion Service: A Machine Learning Platform to Ingest Documents at Scale* | Describes a cloud-based system that uses machine learning for document ingestion and text extraction. | Introduces a scalable platform capable of handling large volumes of documents with high Precision and recall. |
| Sasirekhaand Chandra(2012) | *Enhanced Techniques for PDF Image Segmentation and Text Extraction* | Focuses on segmentation techniques and OCR to improve text extraction. | Tackles challenges like varying font styles and orientations, proposing effective improvements in segmentation. |
| Smithetal.(2020) | *Efficient and Accurate Text Extraction from Noisy Documents* | Combines image pre-processing with deep learning models for robust extraction. | Addresses the problem of extracting text from noisy or poor-quality documents, achieving superior accuracy. |

ShivaniSurana et al.(2022)

Surana and her team reviewed ML-based techniques for converting handwritten and printed text from images into digital formats. They used preprocessing, segmentation, and OCR methods like region-based and texture-based techniques for accurate text detection. Their research highlighted applications in domains such as healthcare and banking, improving data accessibility and reducing manual effort in text processing tasks

MSinthujaa et al.(2023)
Sinthujaa and her team proposed a hybrid model combining CNNs for spatial feature extraction and BiLSTMs for sequential context in text extraction tasks. This approach achieved notable accuracy improvements of 88.58% for handwritten text and 90.8% for printed text, outperforming traditional models.Their work addressed challenges like variability and complexity in text recognition, demonstrating the hybrid model's potential for OCR advancements and reliable text analytics.

Anuradha Thorat et al.(2023)
Thorat and her team emphasized the importance of preprocessing techniques like noise reduction and binarization in OCR systems for handling degraded or distorted text. Their method utilized CNNs fortext detection and CRNNs for recognizing spatial and sequential data, improving recognition rates .There search also explored the social impact of converting text to audio formats for visually impaired users, enhancing accessibility and functionality.

ChaitanyaU et al. (2023)
This paper suggests a hybrid approach to the extraction of digital text from images by combining OCR and MSER algorithms .It highlights the need for high accuracy intext recognition from images .MSER detects stable regions in the image for text localization, while OCR processes the detected regions with CNN layers for character detection and extraction. This method results in enhanced performance than when OCR is used in isolation, especially when images contain alphabets , numbers or both .However, the model faces challenges with skewed oriented text .The model can be used in document digitization, license plate detection, and automated data entry. Future enhancements may include multilingual text recognition, video-based text extraction, and intelligent error correction.

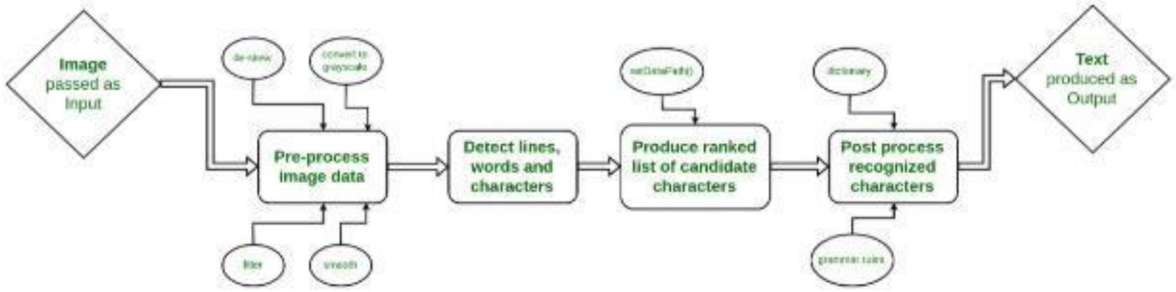
Hemalakshmi etal .(2017)
Hemalakshmi and her colleagues developed a multilingual OCR system using Tesseract for text extraction and Bing Translator API for translation. The system achieved over 90% accuracy in recognizing text across diverse languages and formats .Their scalable and portable approach addressed challenges like varying fonts and degraded input quality, offering effective solutions for multilingual text recognition and translation tasks.

HongLiang et al.(2017)
The paper reviews the methods of extracting text features and emphasizes the role of deep learning in this field.

Traditional methods like filtering, fusion, mapping, and clustering rely on handcrafted features, whereas deep learning models such as CNNs, RNNs, autoencoders, and DBNs automatically learn feature representations from rawdata. These models excel in complex tasks like text classification, sentiment analysis , and information retrieval by leveraging hierarchical feature learning. Deep learning overcomes traditional approaches to handling big data and unstructured data hence text mining is both efficient and accurate. Future applications will involve perfecting feature extraction for varied uses, including integration with other developments like generative adversarial networks.

## III. PROPOSED SYSTEM

The proposed text extraction system is an ext-generation solution designed to meet the growing demand for efficient, accurate and scalable text processing. Built on Java Spring Boot with a micro services architecture, it ensures robust, modular, and adaptable performance across various environments. Leveraging machine learning-driven OCR, it excels in recognizing complex layouts , multi-languagetext,andhandwrittencontentwithexceptionalaccuracy.The system's preprocessing techniques was a enhance in a recognition rates by optimizing input quality. PostgreSQL ensures secure, scalable, and compliant data management, while its modular design allows the workflows customization for diverse need. Supports structured data extraction, multi-language processing, and handwriting recognition, it integrates seamlessly into existing workflows. With features like encrypted data handling, role-based access control, and industry-standard compliance, it is a cutting-edge solution for document digitization. Its efficiency, scalability, and adaptability make it an indispensable tool for organizations seeking automation and operational excellence.



### Architecture

Microservices architecture provides a modular approach to building software systems by breaking down a monolithic application into a collection of smaller, independently deployable services. Each service is designed to handle a specific business functionality, promoting reusability and ease of management. In the context of text extraction systems, microservices enable parallel processing and seamless integration of components like file preprocessing, OCR, and text postprocessing. These services communicate through lightweight protocols such as REST APIs, the system achieves adaptability to varying workloads. This architecture also facilitates independent development and deployment of services, reducing downtime and improving maintainability. Enhanced security measures like token-based authentication and encrypted communication further ensure data protection across the ecosystem.

### *Software Requirement*

The proposed text extraction system is built on a robust foundation of software and hardware components to ensure optimal performance, scalability, and reliability. On the software side, the system utilizes Java as the primary programming language, employing the Spring Boot framework to enable efficient backend services and modular application development.

For data management, PostgreSQL is used, featuring JSONB support for the storage and management of unstructured data, ensuring flexibility and performance.

The hardware requirements are equally essential to ensure the system's ability to handle intensive workloads. A processor with at least 8 cores and a clock speed of 4.0 GHz or higher is necessary to manage concurrent processes efficiently.

Together, these requirements form a highly scalable and efficient ecosystem capable of processing large volumes of data with high accuracy and speed. This architecture ensures the system meets the demands of modern enterprises, delivering reliable and scalable solutions for text extraction and document automation.

## IV. **METHODOLOGY**

In this document, we present a detailed methodology that outlines the key steps involved in building and operating an efficient Text Extraction System. This methodology ensures the extraction process is stream lined, secure, and tailored to meet diverse user requirements. This system addresses the increasing demand for extracting structured and unstructured data from diverse document formats, transforming it into usable, editable, and analyzable formats. By integrating feature are user authentication, file preprocessing, and compliance with regulatory standards, the system ensures a seamless and reliable experience.

Step1:User Registration and Authentication

The first step in the Text Extraction System is managing user accounts securely. Users register by providing their name, email, and password through REST APIs. To ensure account validity, email or phone verification is implemented. Once registered, user credentials and dataaresecurely stored in a PostgreSQL database. Authentication is handled through JSON Web Tokens (JWT), enabling secure token-based sessions

Step2:Upload File

After logging in, users can upload documents through the system's interface. The system supports various filetypes, including images(JPG,PNG), PDFs(singleandmulti-page), and scanned documents. During this step, the uploaded files are validated to ensure they meet the required format , size ,and resolution criteria.This guarantees that the input is suitable for further processing.

Step3:Manage  User Credentials

Once the files are uploaded, the system processes them without storing the actual files. Instead, only the user credentials, such as their username, email, and any relevant metadata about the user (e.g., upload timestamp or user activity logs), are securely stored in a PostgreSQL database. This approach ensures privacy and minimizes storage requirements, while still allowing the system to associate processing actions with specific users for accountability and future reference.

Step4:Preprocess Files

Before text extraction begins ,the uploaded files under go preprocessing to enhance their quality and prepare them for OCR. This step includes operations such as deskewing to straighten misaligned images, removing noise, and improving resolution for better readability.The Multi-page PDFs are split into individual pages to enable efficient processing.

Step5:PerformTextExtraction

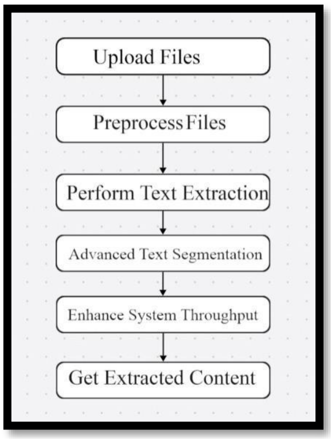
The system uses OCR (Optical Character Recognition) tools like Tesseract and AWS Textract models to extract text from the preprocessed files.It is equipped with the capability to recognize multiple languages ,including those written in right- to-left scripts such as Arabic and Hebrew. Moreover, the system can identify handwritten text and convert it into editable formats, enabling comprehensive text extraction from diverse document types.

Step6:Advanced Text Segmentation

Beyond simple text extraction, the system identifies and extracts structured elements like tables, forms, and other complex layouts with in the document .This step ensures that tabular data ,forms ,and other organized information are accurately extracted and presented in an editable format, maintaining the structure of the original content.

Step7:User Review and Interaction

After the system processes the uploaded files, users can review the extracted information or results. The system provides an intuitive interface for users to verify and interact with the data, ensuring accuracy and emphasizes user involvement in validating and refining the system's output.



*Technologies for TextExtraction*

OpticalCharacterRecognition(OCR):

OCR means Optical Character Recognition. It allows text from images, scanned documents, and handwritten notes to be converted to an editable, searchable digital format, enabling data retrieval, editing, and analysis. Currently, deep learning models such as CNN and RNN have been used to improve the accuracy of OCRs and are, therefore, able to overcome the challenges posed by varied fonts and distorting of images. Tesseract a Google-developed open-source OCR engine that could be used for multilingual text, skewed characters, and non-standard fonts is underscored as a typical application of OCR whereby machine learning techniques and preprocessing methods such as binarization and edge detection are also used. Thus OCR has become an inherent asset for use in document management, translation services, and for improving accessibility.

Tesseract:

Tesseract is an OCR engine that converts text in images into usable formats such as plain text or searchable PDFs through a multi-step process. It begins with preprocessing techniques like binarization, gray-scaling, noise removal, and skew correction to enhance image clarity. Text detection isolates the text regions, while character segmentation breaks the text into individual characters. Then, text recognition using complex or handwritten texts uses CNNs and LSTM networks. Finalizing steps in the process are also post-processing including spell correction and confidence scoring that fine-tunes the recognized text.

AWS Textract:

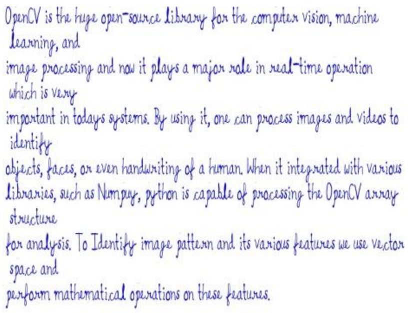
Text Detection (OCR): Textract utilizes deep learning models, including CNNs, YOLO/EAST, and LSTM, for accurate detection and recognition of both printed and handwritten text .Key-Value Pair Extraction: This makes use of transformers such as BERT and GNNs to extract structured data from forms, like names and address .Table Detection: Textract applies image segmentation models like Mask R-CNN and GNNs to detect tables and preserve the structure of these when extracting the data. Handwriting Recognition: A hybrid of CNNs and LSTMs is used to recognize handwriting. Improving sequence alignment is done with a CTC loss function.PDF Processing for Multi-Page Documents: With Textract for multi-page documents, they use Sage Maker-trained models and reinforcement learning to process documents.

$$Q_A = \frac{\sum_{i=1}^{128} X_i}{128} \sqrt{\frac{1}{128} \sum_{i=1}^{128} (X_i - \mu)^2} \in [0.2, 0.4]$$

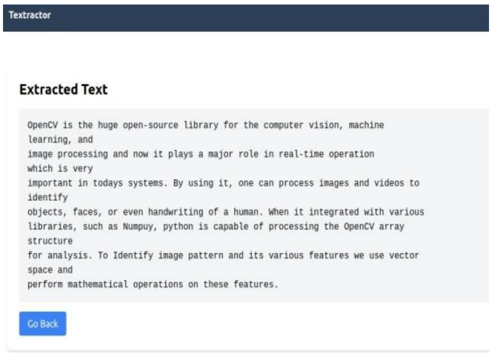
$$Q_A = \frac{\sum_{i=1}^{128} \frac{\left[x\left(\frac{N}{2}\right) + x\left(\frac{N}{2}+1\right)\right]}{2}}{128} \sqrt{\frac{1}{128} \sum_{i=1}^{128} (X_i - \mu)^2} \in [0, 0.2] \cap [0.4, +\infty]$$

The equations define a method for extracting features from word embeddings by focusing on the deviation of each embedding value from the median of the embedding matrix. Equation (1) calculates the summation of these deviations across all dimensions of the embedding, while Equation (2) applies a threshold to refine the feature extraction process. This approach helps normalize and emphasize significant variations in the embedding values, ensuring that the resulting features capture meaningful patterns for downstream tasks.

Results:

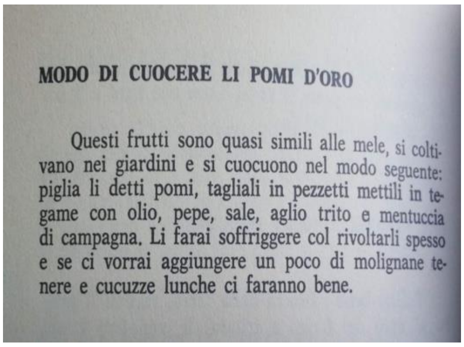


OpenCV is the huge open-source library for the computer vision, machine
learning, and
image processing and now it plays a major role in real-time operation
which is very
important in todays systems. By using it, one can process images and videos to
identify
objects, faces, or even handwriting of a human. When it integrated with various
libraries, such as Numpy, python is capable of processing the OpenCV array
structure
for analysis. To Identify image pattern and its various features we use vector
space and
perform mathematical operations on these features.



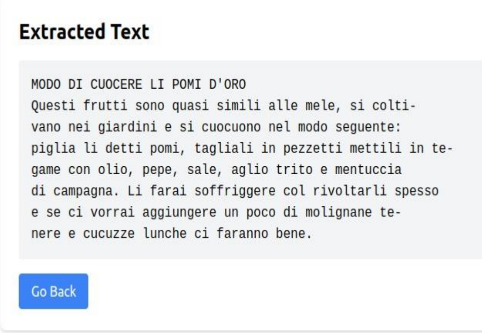
**Extracted Text**

```
OpenCV is the huge open-source library for the computer vision, machine
learning, and
image processing and now it plays a major role in real-time operation
which is very
important in todays systems. By using it, one can process images and videos to
identify
objects, faces, or even handwriting of a human. When it integrated with various
libraries, such as Numpuy, python is capable of processing the OpenCV array
structure
for analysis. To Identify image pattern and its various features we use vector
space and
perform mathematical operations on these features.
```



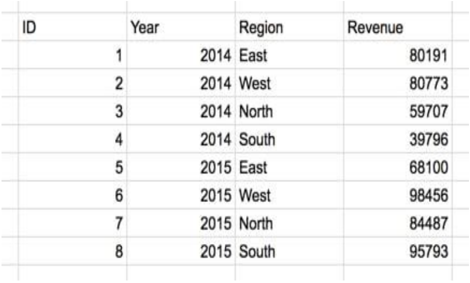
MODO DI CUOCERE LI POMI D'ORO

Questi frutti sono quasi simili alle mele, si coltivano nei giardini e si cuocuono nel modo seguente: piglia li detti pomi, tagliali in pezzetti mettili in tegame con olio, pepe, sale, aglio trito e mentuccia di campagna. Li farai soffriggere col rivoltarli spesso e se ci vorrai aggiungere un poco di molignane tenere e cucuzze lunche ci faranno bene.

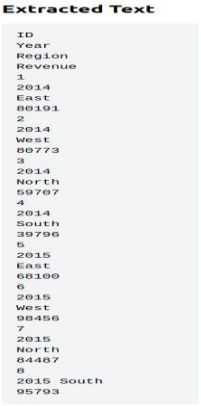


**Extracted Text**

```
MODO DI CUOCERE LI POMI D'ORO
Questi frutti sono quasi simili alle mele, si colti-
vano nei giardini e si cuocuono nel modo seguente:
piglia li detti pomi, tagliali in pezzetti mettili in te-
game con olio, pepe, sale, aglio trito e mentuccia
di campagna. Li farai soffriggere col rivoltarli spesso
e se ci vorrai aggiungere un poco di molignane te-
nere e cucuzze lunche ci faranno bene.
```

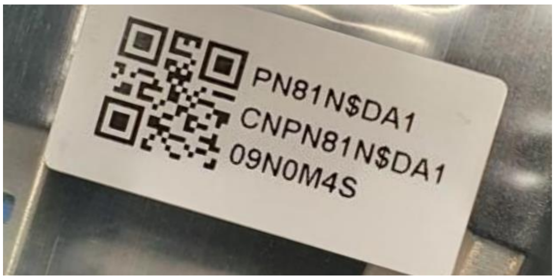


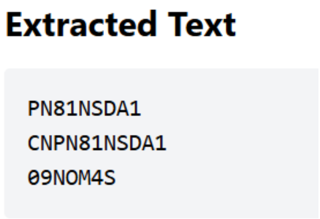
| ID | Year | Region | Revenue |
|---|---|---|---|
| 1 | 2014 | East | 80191 |
| 2 | 2014 | West | 80773 |
| 3 | 2014 | North | 59707 |
| 4 | 2014 | South | 39796 |
| 5 | 2015 | East | 68100 |
| 6 | 2015 | West | 98456 |
| 7 | 2015 | North | 84487 |
| 8 | 2015 | South | 95793 |



**Extracted Text**

```
ID
Year
Region
Revenue
1
2014
East
80191
2
2014
West
80773
3
2014
North
59707
4
2014
South
39796
5
2015
East
68100
6
2015
West
98456
7
2015
North
84487
8
2015 South
95793
```





Extracted Text

```
PN81NSDA1
CNPN81NSDA1
09NOM4S
```

## V. Conclusion

The proposed advanced text extraction system is designed to address the critical challenges for digitizing ,analyzing, and automating unstructured data within images, scanned PDFs, and handwritten documents. The system aims to avoid many limitations in existing solutions, including complex layouts, multiple languages, and high-volume workloads. Utilizing newer technologies such as machine learning-based OCR, containerized microservices, and PostgreSQL with JSONB support can leverage high accuracy, modularity, and adaptability in the overall system.

With Java Spring Boot for backend services, scalability and robust architecture will ensure that deployment and scalability will happen smoothly .Furthermore ,the modular design allows users to customize work flows according to particular needs in preprocessing, text extraction, and postprocessing steps.

This flexibility will then ensure that the system finds application in any sector-be it healthcare, education, legal, or finance-by leveraging digitization and automation as essential components. Thus, by addressing both functional and non-functional requirements in depth, the proposed solution is establishing a new benchmark for text extraction systems. Looking ahead, the system holds great promise for future upgrades. Integration of advanced AI models for enhanced accuracy, multi-language support, and augmented reality (AR) capabilities could further take its usefulness to a different level altogether. By filling the seamless gap between unstructured and structured data, this solution has the potential to unlock new efficiencies for organizations, marking it as a transformative tool for the digital age. This work will open the door to new technologies for processing documents, making the digital ecosystem more inclusive and accessible.

## References

[1] HTTPS://IEEEXPLORE.IEEE.ORG/DOCUMENT/9752274

[2] https://www.sciencedirect.com/science/article/pii/S1877050924007518?ref=pdf_download&fr=RR2&rr=8fa93c45fe8f3a3c

[3] https://ieeexplore.ieee.org/document/9752274

[4] https://www.ijraset.com/research-paper/digital-image-text-recognition-using-machine-learning-algorithms

[5] https://www.technoarete.org/common_abstract/pdf/IJERCSE/v4/i4/Ext_14086.pdf

[6] https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-017-0993-1

[7] https://www.postgresql.org/docs/

[8] A. P. S. Saurabh Dome, "Optical character recognition using tesseract and classifi- cation," in 2021 IEEE International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2021

[9] S.L.YuandongLuan,"Researchontextclassificationbasedoncnnandlstm,"in2019IEEEInternational

Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2019

[10] D.G.AhlamAlnefaie,I.R.MonowarH.Bhuyan,andM.P.PrashantGupta,"End-to-endanalysisfortext detection and recognition in natural scene images." IEEE.

[11] Y. M. R. Srinandan Komanduri and M. M. Bala, "Novel approach for image text recognition and translation," in 2019 IEEE Third International Conference on Com- puting Methodologies and Communication (ICCMC). IEEE, 2019.

[12] D. Pratik Madhukar Manwatkar and R.Singh, "A technical review on text recognition from images," in 2015 IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO). IEEE, 2015.