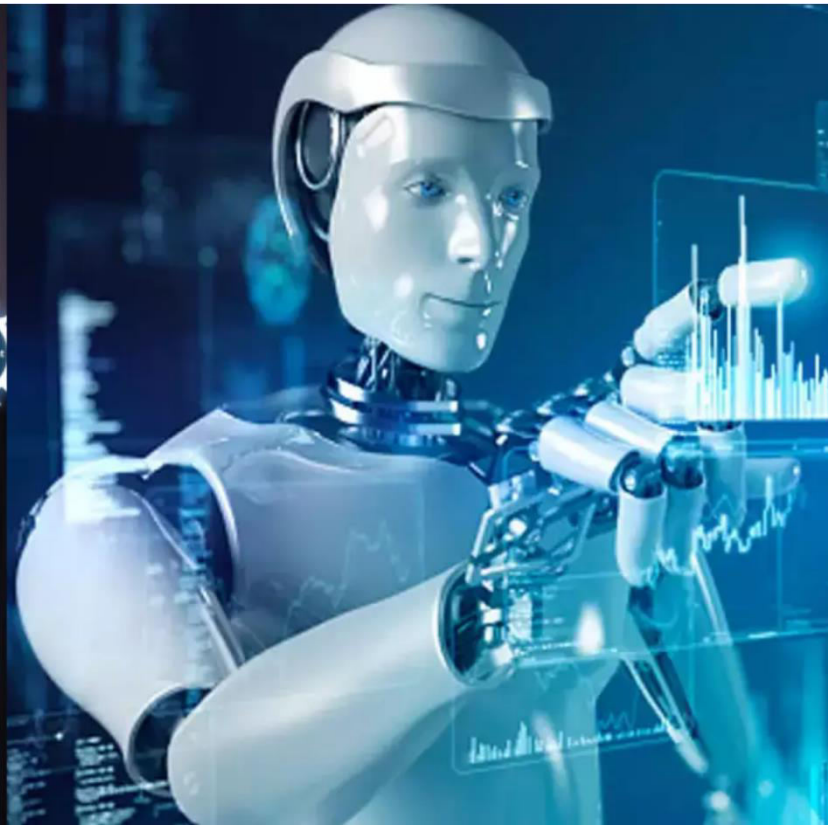
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 5, May 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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405298|

Transformo Docs: A Machine-Readable Document Management System

Dr. Nandini S, K M Rakshith, G Likhith Kumar Reddy

Associate professor, Dept. of ISE, S J C Institute of Technology, Chickballapur, Karnataka, India

UG Student, Dept. of ISE, S J C Institute of Technology, Chickballapur, Karnataka, India

UG Student, Dept. of ISE, S J C Institute of Technology, Chickballapur, Karnataka, India

ABSTRACT: Modern enterprises generate and handle vast volumes of unstructured data in the form of documents, including scanned files, handwritten notes, and multi-format attachments. Traditional document management systems lack the capability to efficiently parse, process, and retrieve such documents. This paper presents Transformo Docs, a machine-readable document management system designed to convert unstructured content into structured, searchable formats using Optical Character Recognition (OCR), metadata tagging, and rule-based extraction techniques. The system supports JSON, XML, and indexed PDFs for easy integration with enterprise applications. With self-hosting capabilities, role-based access control (RBAC), and strong compliance features, Transformo Docs is a scalable and secure alternative to cloud-based document solutions.

KEYWORDS: Information-intensive, Document processing, Document retrieval, Document management systems (DMS), Unstructured data, Scanned documents, Metadata tagging, Optical Character Recognition (OCR), Structured data formats, Format standardization, Scalability, Compliance, Integration, Data accessibility, Retrieval speed, Document security, Workflow efficiency.

I. INTRODUCTION

One of the primary objectives of a document management system in modern enterprises is to enable intelligent, automated processing of unstructured data without compromising accuracy, efficiency, or security. In contemporary business environments, documents are generated in various formats including PDFs, scanned images, and handwritten inputs. Traditional document management systems often lack the adaptability and intelligence needed to process such unstructured data, resulting in inefficiencies, data retrieval challenges, and operational delays. Transformo Docs addresses these challenges by introducing a machine-readable document management solution that utilizes Optical Character Recognition (OCR), Natural Language Processing (NLP), and metadata tagging to convert unstructured content into structured, searchable formats such as JSON and XML.

In particular, the system enhances information flow by implementing path-aware extraction and classification mechanisms that reduce manual effort while maintaining end-to-end document processing reliability. Using a combination of node selection schemes (NSS) and channel-like routing logic (conceptually analogous to CSS in networking), Transformo Docs determines optimal processing paths and data fields to be extracted and indexed. This reduces computational overhead, preserves energy (in terms of resource use), and ensures minimal latency during document searches and access. Furthermore, with integrated role-based access control (RBAC), encryption protocols, and compliance with regulatory frameworks like GDPR and HIPAA, the system ensures secure, trustworthy handling of sensitive data. Through intelligent integration APIs, it also enables seamless interoperability with enterprise systems like ERP and CRM, establishing itself as a secure, scalable, and efficient framework for structured document communication.

II. PROBLEM STATEMENT

In the modern digital ecosystem, organizations increasingly generate a diverse range of documents from scanned files and handwritten forms to PDFs and images creating a vast repository of unstructured data. Traditional document management systems are not equipped to handle the complexity and volume of such formats effectively. As a result,

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405298|

critical challenges such as manual data entry, inefficient document retrieval, and disjointed workflows continue to hinder operational efficiency. These systems often lack the intelligence to extract meaningful content, classify information, or enable context-aware search, which leads to increased workload, slower decision-making, and higher chances of human error.

Additionally, concerns around data compliance and security further complicate the adoption of digital solutions. Manual processes expose organizations to risks such as data breaches, regulatory non-compliance, and workflow bottlenecks that negatively affect productivity. Moreover, the absence of intelligent document classification and structured storage mechanisms contributes to inconsistent data accessibility and poor integration with enterprise tools. There is a clear need for a system like Transformo Docs that can address these issues through intelligent automation, structured data conversion, and secure, standards-compliant management of documents. By doing so, organizations can achieve a seamless, efficient, and scalable document processing environment.

III. LITERATURE REVIEW

Recent research in document understanding and metadata extraction has increasingly adopted deep learning, graph-based models, and transformer architectures. In [1] a unified framework leveraging transformers is introduced to extract visual information from documents, enabling end-to-end processing across multiple modalities. Graph learning combined with layout-awareness, as proposed in [2] further enhances multimodal document understanding by preserving spatial relationships. PICK [3] employs improved graph convolutional networks to extract key fields, while DocFormer [4] processes layout, image, and text features jointly for improved document parsing. Visual cues such as font and layout are utilized in [5] to extract metadata from scanned theses and dissertations, demonstrating the importance of visual features in document analysis.Other contributions focus on classification, clustering, and archiving of documents. A deep learning approach to classify technical documents using CNNs and LSTMs is proposed in [6] showing strong performance on domain-specific corpora. An OCR-based system using PyTesseract [7] supports document archiving and indexing for efficient record management. In [8], NLP techniques are used to classify and analyse documents semantically, while keyword extraction drives the unsupervised document clustering method discussed in [9]. Metadata extraction from scientific PDFs is further improved in [10] using a hybrid model of vision and natural language understanding. Lastly, [11] presents an end-to-end system for processing scanned electronic theses using both visual and structural cues, contributing to more intelligent academic document management.

IV. DESIGN

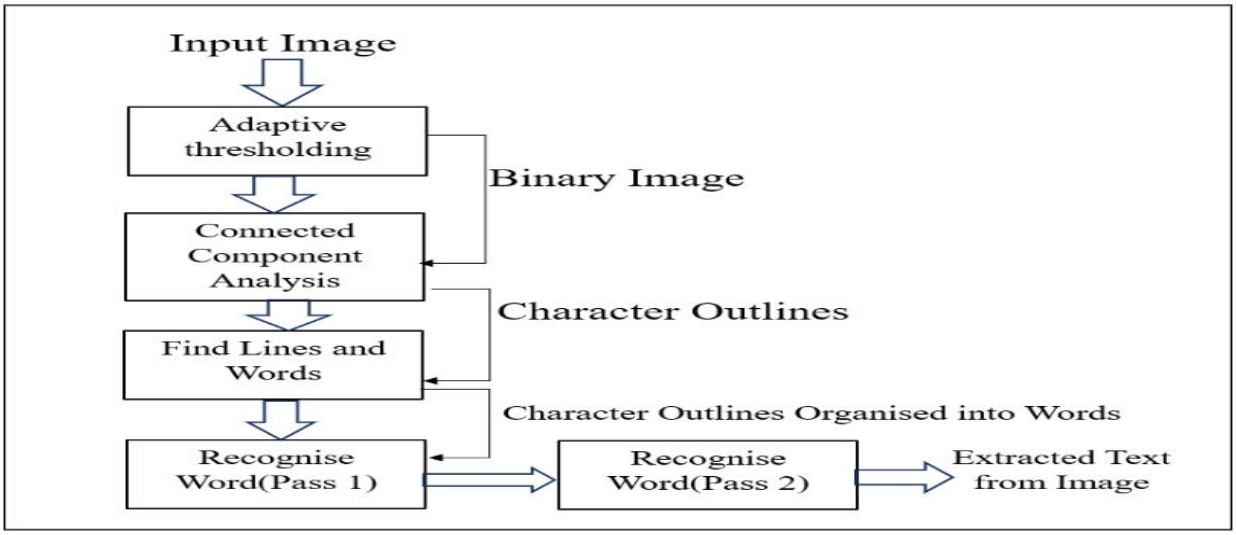


Fig. 1 Data Flow Diagram For processing extraction of text from an image

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405298|

The OCR pipeline transforms visual data from input images into structured, machine-readable text through a series of image processing and recognition stages. The process begins with an input image containing textual content. The first step is adaptive thresholding, a preprocessing technique that converts the grayscale image into a binary image. This transformation enhances contrast by distinguishing foreground (text) pixels from the background, allowing better segmentation in later stages, as shown in Fig. 1.

The binary image is then processed using Connected Component Analysis (CCA). As illustrated in Fig. 1, CCA groups neighbouring pixels into distinct regions, each corresponding to potential character components. This step isolates the character outlines, which are critical for subsequent segmentation and recognition tasks.

Once character outlines are identified, the system proceeds to text line and word detection, organizing character components into coherent lines and words, as depicted in Fig. 1.

The text recognition component operates in two successive passes:

1. Pass 1 – Recognize Word: An initial recognition is performed to estimate the textual content.
2. Pass 2 – Recognition Refinement: Contextual and statistical methods are used to improve the recognition output (see Fig. 1).

Finally, the recognized and refined text is compiled as the extracted text from the image, completing the OCR pipeline illustrated in Fig. 1.

V. IMPLEMENTATION

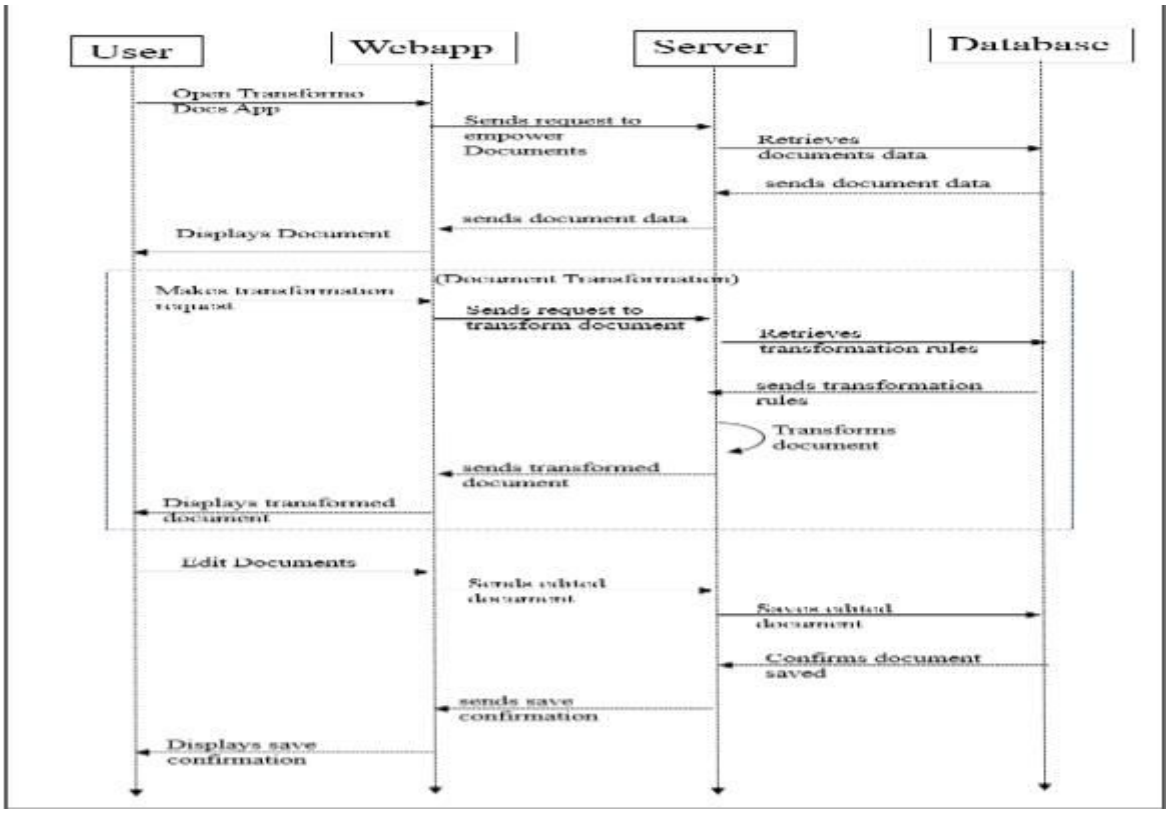


Fig. 2 Sequence diagram for representation of interaction between different components

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405298|

The implementation of Transformo Docs follows a modular development approach using Python with Flask as the backend framework and Tailwind CSS with JavaScript on the frontend. The backend includes modules for document configuration, validation, parsing (including OCR via Tesseract), and storage. As seen in Fig. 2, the system handles user interaction through a well-defined request-response cycle involving the web application, server logic, and database. This interaction is further illustrated through a sequence diagram, which outlines the step-by-step flow of operations from initial document retrieval to transformation and user-driven editing.

When a user accesses the application, the frontend initiates a data fetch request, prompting the server to retrieve the relevant documents from the database. Upon transformation requests, the server applies pre-defined rules stored and managed in the backend and returns a modified version of the document. Real-time editing features are also supported, where user edits are validated and saved securely in the database, with confirmation relayed back through the application interface. This structured communication ensures a seamless and interactive user experience while maintaining data integrity and operational efficiency.

The system also incorporates asynchronous request handling to improve responsiveness and scalability, especially under concurrent usage. Each component communicates through RESTful APIs, enabling clean separation of concerns and simplifying maintenance. Logging and exception handling mechanisms are implemented throughout the backend for easier debugging and traceability. Furthermore, role-based authentication, session management, and secure data storage practices ensure that user interactions remain protected and compliant with best practices in web application security.

VI. CONCLUSION

TransformoDocs represents a significant step forward in intelligent document management, offering a streamlined, automated solution for processing, storing, and retrieving documents with minimal human intervention. Its modular architecture and intuitive user interface enable efficient handling of various document formats while reducing errors and operational overhead. With robust features such as real-time processing, role-based access control, document validation, and secure storage, the system ensures reliability, scalability, and compliance across diverse industries including legal, healthcare, finance, and education.

The application's backend is built to support high-volume document processing, enabling seamless performance in demanding environments. Advanced functionalities like version control, structured workflows, and metadata tagging provide users with powerful tools to manage documents systematically and securely. The integration of search and retrieval mechanisms allows for immediate access to critical data, further improving efficiency and user experience.
Designed with future scalability in mind, TransformoDocs is well-positioned to evolve alongside emerging technological needs. The platform’s potential integration with AI-driven capabilities such as intelligent classification, contextual search, and real-time collaboration—underscores its role as a forward-looking solution in the digital transformation landscape. As organizations continue to move toward paperless operations, TransformoDocs offers a secure, efficient, and adaptable framework that empowers users to manage documents intelligently and with confidence.

REFERENCES

[1] M. Xu, J. Li, M. Yang, C. Wang, and Y. Wu, “A Unified Document Analysis Framework for Visual Information Extraction Using Transformers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 7, pp. 1428–1439, Jul. 2024.
[2] S. Huang, H. Liu, and X. Chen, “Layout-Aware Graph Learning for Multimodal Document Understanding,” IEEE Transactions on Multimedia, vol. 25, no. 4, pp. 1123–1135, Apr. 2023.
[3] W. Zhang et al., “PICK: Processing Key Information Extraction from Documents Using Improved Graph Learning-Convolutional Networks,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 14440–14449, Jun. 2022.
[4] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “DocFormer: End-to-End Transformer for Document Understanding,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 5553–5562, Oct. 2021.

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 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405298|

[5] M. H. Choudhury, H. R. Jayanetti, J. Wu, W. A. Ingram, and E. A. Fox, "Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations," Journal of Information Science and Engineering, vol. 37, no. 1, pp. 1–16, Jan. 2023.

[6] R. Kumar and A. Singh, "Deep Learning for Technical Document Classification," International Journal of Computer Applications, vol. 184, no. 18, pp. 10–16, Jan. 2022.

[7] J. Jayoma, E. S. Moyon, and E. M. Morales, "OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines," in Proc. IEEE 12th Int. Conf. Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), pp. 1–6, Dec. 2020.

[8] P. S. Rawal and A. K. Gupta, "Smart Document Classifier and Analyzer Using NLP," International Journal of Computer Applications, vol. 178, no. 10, pp. 25–29, Jan. 2021.

[9] K. Patel and D. Shah, "Document Clustering Using Keyword Extraction," International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 7, pp. 1214–1218, Jul. 2021.

[10] T. Nguyen, H. Doan, and L. Ngo, "Vision and Natural Language for Metadata Extraction from Scientific PDF Documents," in Proc. 22nd ACM/IEEE Joint Conf. Digital Libraries (JCDL), pp. 1–4, Jun. 2022.

[11] M. L. Bautista and A. D. Reyes, "Processing Key Information Extraction from Documents Using Improved Graph Learning-Convolutional Networks (PICK)," Journal of Information and Knowledge Management, vol. 22, no. 1, pp. 1–12, Jan. 2023.

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