A Synopsis Report

On

Design and Development of Document Intelligence & Routing System

Submitted to the Department of Computer Science and Engineering

In partial fulfilment of the requirements

For the degree of

Bachelor of Technology in Computer Science and Engineering

By

Nivedika Sharma
2200140100072

Pankhi Gupta
2200140100073

Sanmita Biswas
2200140100093

Saumya Gupta
2200140100096

Group No. 08

Guided By

Dr. Shahjahan Ali

Department of Computer Science *and Engineering*
Shri Ram Murti Smarak College of Engineering & Technology, Bareilly
Dr.A.P.J. Abdul Kalam Technical University, Lucknow
October, 2025

Acknowledgement

We are thankful to our college **Shri Ram Murti Smarak College of Engineering & Technology, Bareilly** for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Project Guide **Dr. Shahjahan Ali** and for his kind help and valuable advice during the development of project synopsis and for his guidance and suggestions.

We are deeply indebted to Head of the Computer Science and Engineering Department **Dr. Shahjahan Ali** and our Principal **Dr. Prabhakar Gupta**, for giving us this valuable opportunity to do this project.

We express our hearty thanks to Project InCharge **Dr Jyoti Agrawal** for her assistance without which it would have been difficult to finish this project synopsis and project review successfully.

We convey our deep sense of gratitude towards all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Computer Science and Engineering Department. We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Nivedika Sharma (2200140100072)

Pankhi Gupta (2200140100073)

Sanmita Biswas (2200140100093)

Saumya Gupta (2200140100096)

Abstract

Across industries, organizations are struggling with an ever-increasing volume of documents arriving from multiple channels—emails, collaboration platforms, messaging apps, and scanned hard copies—in formats that range from multilingual text and embedded tables to images and official directives. The absence of a centralized, intelligent system to manage this information flow leads to inefficiencies such as delays in accessing actionable insights, duplicated effort in summarization, fragmented knowledge across departments, and risks of non-compliance with regulatory requirements. These challenges translate into slower decision-making, higher operational costs, and loss of institutional knowledge over time.

This project introduces a web-based intelligent document management platform designed to act as a unified repository for all organizational information. The system automates document ingestion, organization, and classification, while eliminating redundancy through advanced techniques like fingerprinting and stapling. Beyond storage, it generates concise, contextualized summaries and ensures intelligent routing of information to stakeholders based on roles and responsibilities. This guarantees that employees access the most relevant information with full traceability to the source material.

By streamlining knowledge flow, the platform reduces manual duplication, safeguards institutional memory, and ensures compliance directives and operational updates are delivered to the right audience at the right time. The solution is industry-agnostic and scalable—whether for manufacturing, finance, healthcare, logistics, or public services—and establishes the foundation for adaptive, future-ready knowledge infrastructure. Ultimately, it empowers organizations to achieve faster decision-making, greater collaboration, and sustainable operational excellence in the face of document overload.

Table of Contents

List of Figures

1. Introduction of Project

In today's digital-first world, organizations across all sectors are generating and receiving an unprecedented volume of documents every day. These range from operational records, contracts, invoices, and compliance directives to technical reports, multimedia files, and communication logs. The inflow comes through diverse channels such as emails, enterprise applications, messaging platforms, and physical document scans, often in heterogeneous formats and languages.

While this abundance of information has the potential to empower decision-making, in practice it often overwhelms organizations. Traditional storage systems and manual processes are no longer sufficient to handle the complexity, scale, and diversity of modern documentation.

Common issues include duplication of effort, difficulty in locating relevant information, lack of cross-departmental visibility, delayed decision-making, and risk of non-compliance with regulatory standards. Over time, these challenges erode efficiency, increase operational costs, and threaten organizational resilience by making critical knowledge dependent on individual employees.

Moreover, industries today are under mounting pressure to adapt quickly to regulatory changes, digital transformation initiatives, and competitive market forces. In such an environment, delays in retrieving or processing essential documents can have cascading effects—impacting compliance readiness, operational safety, customer satisfaction, and financial performance. Organizations must therefore adopt proactive systems that not only reduce information latency but also enable faster, data-driven responses to evolving challenges.

An intelligent, unified document management platform offers a pathway to overcoming these hurdles. By leveraging advanced technologies such as natural language processing (NLP), optical character recognition (OCR), and machine learning, such a system can automatically ingest, classify, and summarize documents, while ensuring redundancy is eliminated and traceability is maintained.

This shift from passive storage to active knowledge management ensures that stakeholders—whether in operations, compliance, administration, or leadership—receive timely and relevant insights that directly support decision-making. In doing so, organizations can transform

document overload from a liability into a strategic asset, unlocking long-term scalability, collaboration, and resilience.

2. Literature Review

[1] Madhavi et al. (2025): This paper proposed the development of an OCR-driven pipeline integrating text extraction, abstractive summarization, translation, and classification for low-resource Indian languages. The system achieved strong results across OCR, summarization, and multilingual translation tasks.

[2] Godbole et al. (2024): This paper drives contrasts between LLM based and the traditional ways of summarizing documents. It proposes a workflow from model selection to optimization; and uses case studies in legal, medical, and business domains to show how these models significantly improve efficiency and accuracy over traditional methods.

[3] Sunitha et al. (2019): The authors proposed an abstractive summarization system for Malayalam using clause identification and semantic triples(subject–object–predicate). The work addressed challenges of agglutinative morphology and showed promising results compared to extractive methods.

[4] Majid et al. (2019): This paper proposes a strategic model for the railway industry, arguing that traditional, internal-only Knowledge Management is insufficient for innovation. It suggests integrating "Open Innovation" by using online communities as a bridge to connect with external knowledge bases.

[5] Liu & Chen (2017): The authors in this paper developed the DAM, a model for big data, which normalizes multi-source data and applies object/association mapping and a "walk" mechanism. The model enables efficient, scalable, and intuitive multidimensional data analysis and visualization.

[6] Ho & Kim (2014): This paper proposed a fingerprinting-based method using sim-hash and trie-tree structures to detect near-duplicate documents and spam in social networks. Their approach achieved ~98% accuracy, effectively addressing redundancy and malicious data detection in online platforms.

[7] Fan & Huang (2012): This paper introduced a fusion of shingling, I-Match, and Simhash algorithms for near-duplicate document detection. Their multi-fingerprint Simhash method improved precision–recall balance and enhanced digital library search efficiency

[8] Fesehaye, Malik, & Nahrstedt (2010): This paper proposed a scalable distributed file system for cloud computing that improves upon HDFS by using multiple name nodes and a lightweight front-end server. Their protocol enhances bandwidth utilization, reduces transfer latency, and removes single points of failure.

3. Motivation

The motivation for selecting this problem arises from the universal challenge organizations face in managing the ever-growing flood of documents and unstructured information. In nearly every industry, the inability to efficiently capture, organize, and utilize documents leads to delays in communication, duplication of work, regulatory risks, and ultimately reduced productivity. This issue is not confined to one sector but is a global concern for enterprises of all sizes, making it a highly relevant and impactful area of study.

The project presents an opportunity to design a system that addresses these pain points using advanced technologies such as natural language processing, optical character recognition, and machine learning. Unlike traditional document storage systems, the proposed platform emphasizes intelligence—automated classification, redundancy removal, contextualized summarization, and role-based information routing. By doing so, it directly contributes to faster decision-making, improved compliance, and enhanced collaboration, which are priorities for modern organizations.

Furthermore, the scalability of the problem makes it an ideal project choice. The system can begin by handling hundreds of documents daily and gradually expand to manage tens of thousands, making it adaptable to real-world industry needs. This scalability, combined with the solution's cross-industry applicability, ensures that the project does not remain limited to a theoretical exercise but holds strong potential for practical deployment.

Lastly, the problem is highly relevant in the context of emerging trends such as digital transformation, knowledge automation, and AI-driven enterprise systems. By addressing document overload, the project contributes to building future-ready organizations capable of sustaining efficiency and resilience in a data-driven economy. This alignment with current and future industry demands makes the problem statement both timely and compelling as a project.

4. Problem Statement

Organizations across industries are increasingly struggling to manage the massive volumes of documents generated daily from diverse sources such as emails, collaboration tools, messaging applications, and scanned hard copies. These documents exist in various formats—including multilingual text, embedded tables, images, and official directives—and the absence of a

centralized, intelligent system to manage them results in significant inefficiencies. Information overload delays access to critical insights, duplication of effort arises from repeated manual summarization and storage, fragmented awareness across departments impedes collaboration and informed decision-making, regulatory or compliance documents may be overlooked, creating organizational risks, and valuable institutional knowledge is often lost when key personnel transition, hampering operational continuity.

Objectives:

- To develop an intelligent document management system that centralizes and automates the organization, retrieval, and summarization of documents.
- To minimize manual effort and duplication by leveraging AI-based summarization and classification.
- To enhance collaboration and accessibility of information across departments through a unified platform.
- To ensure compliance and secure handling of sensitive or regulatory documents.
- To preserve institutional knowledge and improve continuity by maintaining a well-structured and easily retrievable document repository.

5. Methodology

The proposed system for document management follows a modular pipeline architecture, ensuring that heterogeneous inputs are systematically ingested, processed, routed, stored, and

delivered to end-users in an actionable form. The methodology can be outlined in the following stages:

1. Data Ingestion Layer

- Sources: Documents are collected from multiple channels including emails, SharePoint repositories, Maximo exports, WhatsApp PDFs, cloud links, and scanned hard copies.
- Preprocessing:
    - Optical Character Recognition (OCR) using Tesseract to convert scanned images and PDFs into machine-readable text.
    - Language detection and multi-language support using Google Translate API and language models.
- Objective: Normalize all formats into a consistent, text-based intermediate representation for downstream processing.

2. Processing Layer

- Summarization: Employ Hugging Face Transformers with models such as Pegasus and GPT-based summarizers to generate concise, role-relevant digests of lengthy documents.
- Keyword Extraction & Translation: Key terms and phrases are automatically extracted to improve searchability and cross-language accessibility.
- AI-driven Contextualization: Natural Language Processing (NLP) pipelines classify and tag documents according to themes such as engineering, procurement, HR, compliance, and safety.

3. Routing & Personalization

- Role-specific distribution: A rules-based engine ensures that each stakeholder—engineers, HR managers, finance officers, or executives—receives only the relevant digests.
- Contextual alerts: Time-sensitive or compliance-related updates trigger real-time notifications via Firebase or Twilio SMS/WhatsApp API, ensuring urgent matters are not overlooked.

4. Repository & Knowledge Hub

- Centralized storage: All documents and their summaries are stored in a knowledge repository with version control.
- Search functionality: Integration with Elasticsearch enables fast retrieval of both summaries and original documents.
- Redundancy management: Stapling and fingerprinting techniques prevent duplicate storage of similar or identical files.

5. Outputs & User Interfaces

- Dashboards: Interactive dashboards built using React and Power BI provide real-time visibility into documents, updates, and alerts.
- Traceability: Every summary links back to its original document, preserving trust and compliance.
- User access: Role-based authentication ensures secure access, minimizing information overload while guaranteeing accountability.

6. Continuous Feedback & Improvement

- Monitoring: System logs and analytics track usage, summarization accuracy, and alert effectiveness.
- Learning loop: User feedback on summaries and document relevance is fed back into the NLP models to refine future output

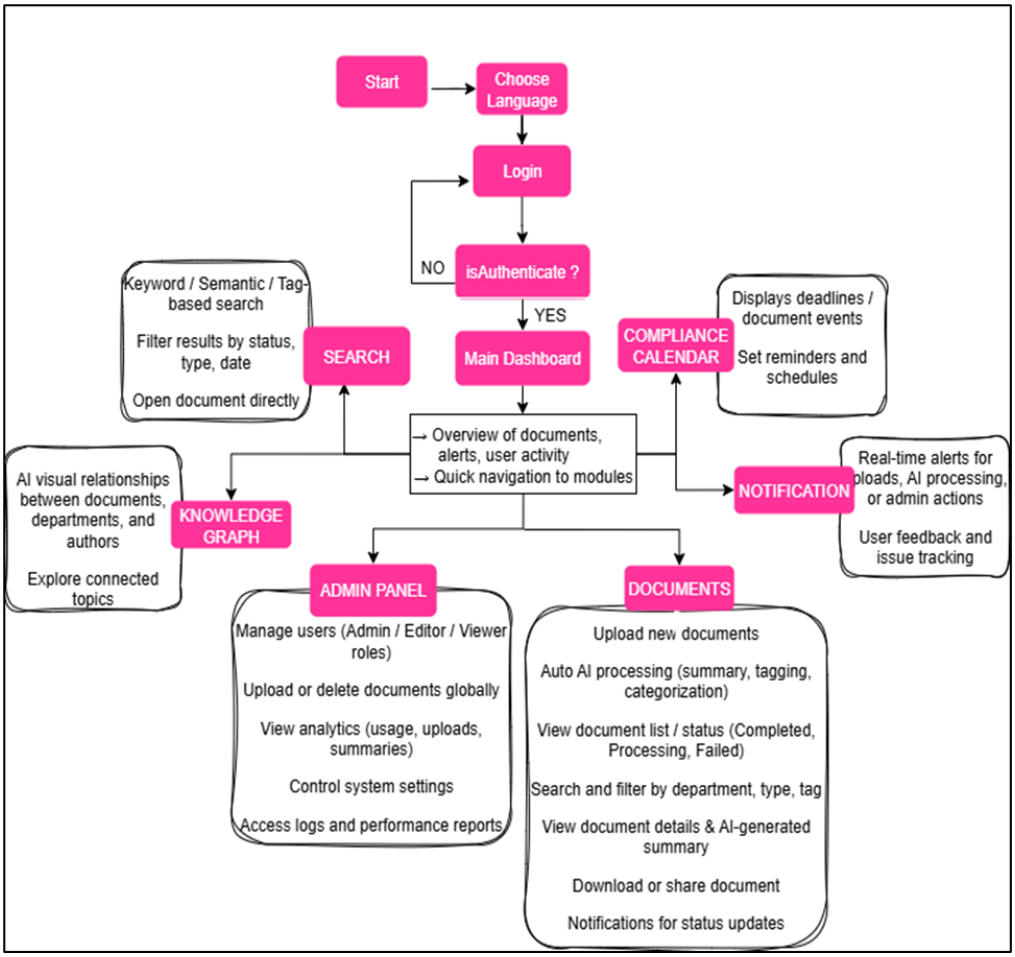

Figure 9.1 System Components

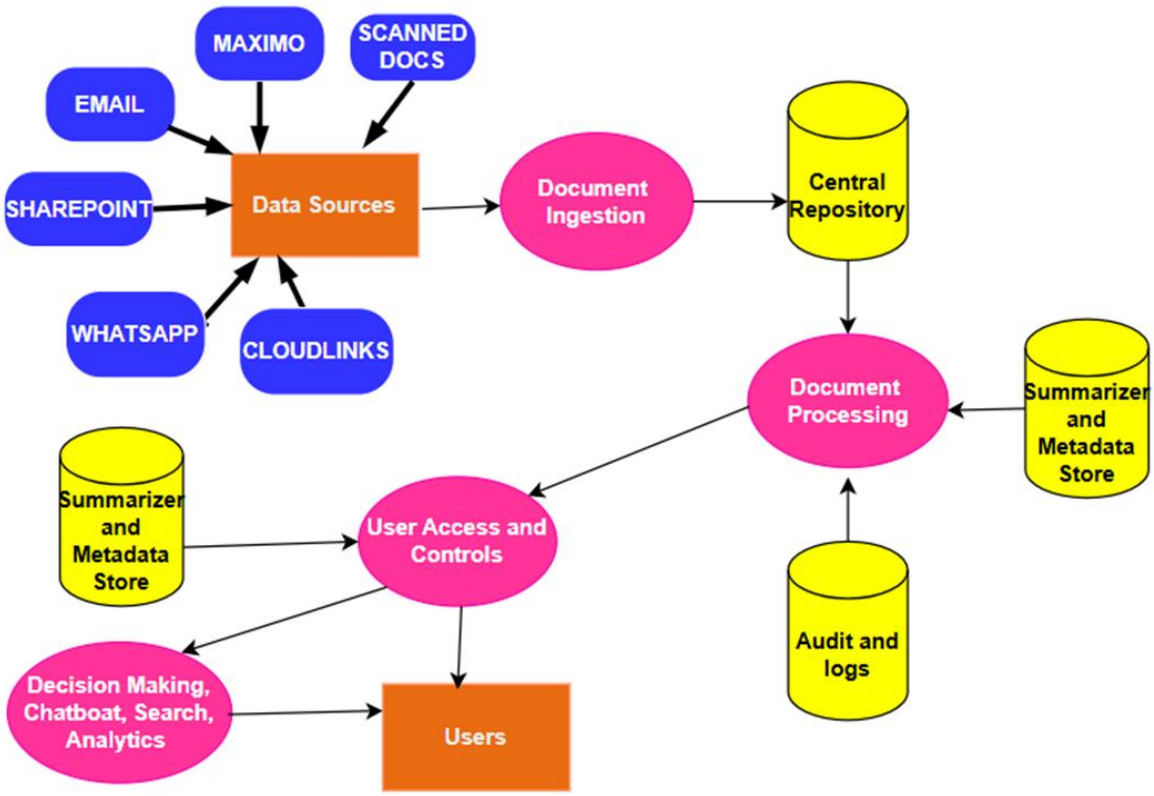

Figure 9.2 Data Flow Diagram

6. Technology Used

- OCR: Tesseract

  To extract text from scanned documents, PDFs, and images for further processing.
- NLP: Hugging Face Transformers

  To analyse, classify, and understand the extracted textual content.
- Summarization: Pegasus, GPT models

  To generate concise, actionable summaries of lengthy documents for quick decision-making.
- Search: Elasticsearch

  To enable fast, intelligent retrieval of documents and information based on keywords or context.
- Frontend: React, Power BI

  To build an interactive, user-friendly dashboard for visualizing summaries, alerts, and analytics.
- Notifications: Firebase, Twilio:

  To send real-time alerts and updates to stakeholders via email, SMS, or app notifications.
- APIs: WhatsApp Business API, Microsoft Graph API, Google Translate API

  To integrate multiple data sources, enable multilingual support, and connect communication channels seamlessly.

7. Requirement and Specifications

7.1 Hardware Requirements

1. **Server:** 4-core CPU, 16 GB RAM, 1 TB storage
2. **Client PCs:** i3/i5 CPU, 8 GB RAM, 256 GB storage
3. **Backup:** External HDD or cloud storage
4. **Optional:** Scanners and printers

7.2 Software Requirements

1. **Programming Environment:** Python (3.9+), Node.js (16+), JavaScript
2. **Backend Framework:** FastAPI (Python) or Express.js (Node.js)
3. **Frontend Framework:** React.js (with Material-UI)
4. **Database:** SQLite or PostgreSQL (local instance for prototype)
5. **File Storage:** Local disk or small cloud bucket (e.g., AWS S3 free tier, Google Drive)
6. **OCR:** Tesseract OCR (open source, locally installable)
7. **NLP/AI:** OpenAI GPT-4 API (summaries), spaCy (local entity extraction)
8. **Search**: Basic PostgreSQL full-text search or ElasticSearch (single-node setup)
9. **Authentication:** Simple JWT or OAuth 2.0 (for demo, can use local user/password auth)
10. **Visualization:** Chart.js (for basic dashboard graphs)
11. **Version Control:** Git (GitHub or GitLab repo)
12. **Code Editor:** VS Code, PyCharm, WebStorm, or similar

8. Expected Outcomes

The implementation of the Document Management System (DMS) will centralize the storage and management of an organisation's diverse documents, including engineering reports, regulatory directives, HR policies, and vendor communications. By providing a single platform for document access, the system will reduce the time and effort required by managers and staff to locate and process critical information, thereby minimizing information latency.

With intelligent text extraction, NLP-based analysis, and automated summarization, stakeholders will receive concise, actionable insights from lengthy documents. This will prevent duplicated work across departments, enhance cross-team awareness, and ensure that key updates and regulatory instructions are promptly addressed, reducing compliance risks and operational delays.

The DMS will also safeguard institutional knowledge by preserving historical records and making them easily retrievable, even when employees transfer or retire. Secure access controls and traceability will ensure that sensitive information is protected while maintaining accountability for document usage.

Finally, by integrating real-time notifications, intelligent search, and user-friendly dashboards, the system will support faster, informed decision-making, improved inter-departmental coordination, and overall operational efficiency. Ultimately, the DMS will help deliver safer, more reliable, and passenger-centric metro services while reducing manual workload and operational costs.

9. Timelines

The implementation of the Document Management System (DMS) is planned in phased stages to ensure systematic development and timely delivery. The project will begin with requirement analysis and system design, followed by document ingestion and preprocessing. Subsequently, core modules for text extraction, NLP-based analysis, automated summarization, and

intelligent search will be developed. Frontend dashboards, notifications, and API integrations will be implemented in parallel. The final phases will include testing, deployment, and user training. Each stage is scheduled with clear milestones to ensure smooth progress, iterative improvements, and alignment with an organisation's operational needs

- Month 1: Requirement doc + architecture design
- Month 2: Data pipeline + database setup
- Month 3: Redundancy removal + basic platform
- Month 4: Full summarization + routing features
- Month 5: Pilot test + refinements
- Month 6: Deployment + training + final report

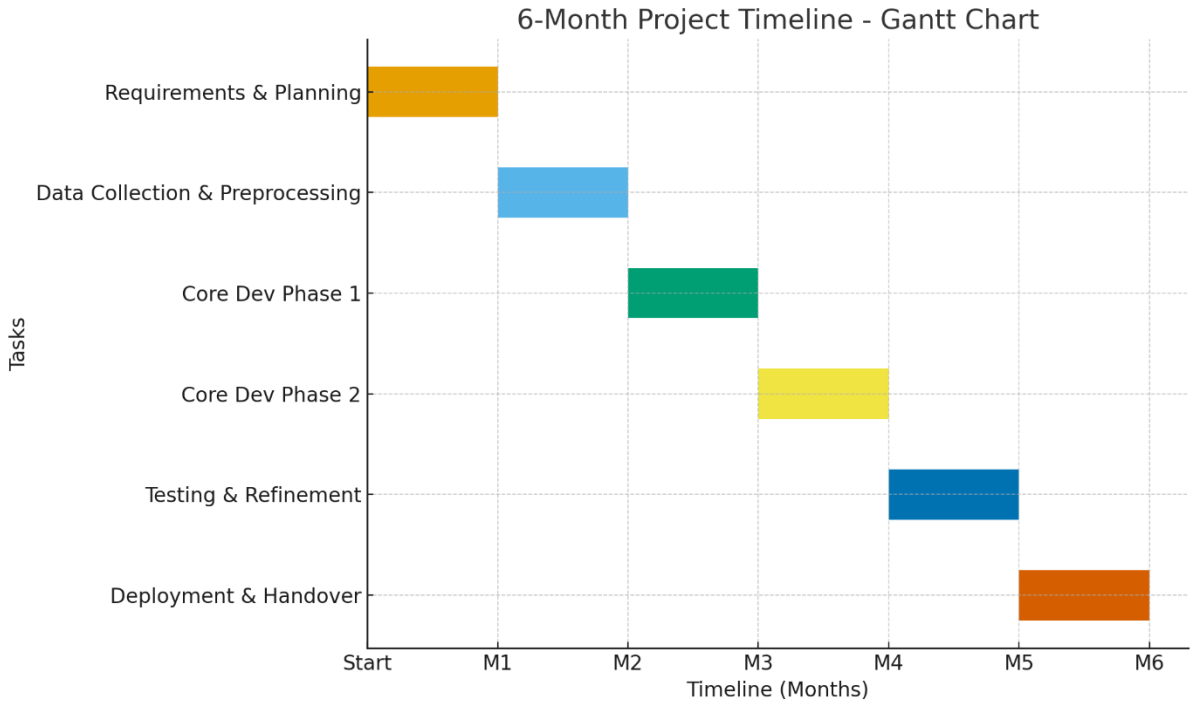

Figure 9.3 Gantt Chart

References

[1] Madhavi, H., Cherian, J., Khamkar, Y., & Bhagat, D. (2025) 'Low-resource language processing: An OCR-driven summarization and translation pipeline' Dr. Vishwanath Karad MIT World Peace University. DOI: https://doi.org/10.48550/arXiv.2505.1117

[2] Godbole, A., George, J., & Shandilya, S. (2024) 'Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications' DOI: 10.48550/arXiv.2409.18454

[3] Sunitha, C., Jaya, A., & Ganesh, A. (2019) 'Automatic summarization of Malayalam documents using clause identification method' International Journal of Electrical and Computer Engineering (IJECE), 9(6), 4929–4938. DOI: 10.11591/ijece.v9i6.pp4929-4938

[4] Babaei, M., Radfar, R., & Eshlaghy, A. (2019) 'Knowledge Management in Railway Industry: A Conceptual Model Based on Open Innovation and online Communities' Islamic Azad University, Tehran, Iran. DOI:10.22068/IJRARE.6.1.63

[5] Liu, P., & Chen, L. (2017) 'A multi-source data aggregation and multidimensional analysis model for big data' ITM Web of Conferences, 12, 05009. DOI: 10.1051/ itmconf/201 5009

[6] Ho, P.-T., & Kim, S.-R. (2014) 'Fingerprint-based near-duplicate document detection with applications to SNS spam detection' International Journal of Distributed Sensor Networks, 2014, Article ID 612970. DOI: http://dx.doi.org/10.1155/2014/612970

[7] Fan, J., & Huang, T. (2012) 'A fusion of algorithms in near duplicate document detection' In L. Cao et al. (Eds.), PAKDD 2011 Workshops: Advances in Knowledge Discovery and Data Mining (pp. 234–242). Springer. DOI: 10.1007/978-3-642-28320-8_20

[8] Fesehaye, D., Malik, R., & Nahrstedt, K. (2010) 'A scalable distributed file system for cloud computing' Proceedings of the International Conference on Cloud Computing. Retrieved from https://www.researchgate.net/publication/228946298_A_Scalable_Distributed_File_System_for_Cloud_Computing

Design and Development of Document Intelligence & Routing System

by

Group no. 8

Signature……..

Name: Nivedika Sharma

Signature……

Name: Pankhi Gupta

Roll no: 2200140100072

Signature……..

Name: Sanmita Biswas

Roll no: 2200140100093

Roll no: 2200140100073

Signature……

Name: Saumya Gupta

Roll no: 2200140100096

Dr. Shahjahan Ali	**Dr. Jyoti Agarwal**	**Dr. Shahjahan Ali**
HOD (CSE)	**PROJECT INCHARGE**	**SUPERVISOR/GUIDE**

Department of Computer Science and Engineering

Shri Ram Murti Smarak College of Engineering & Technology, Bareilly

Dr. APJ Abdul Kalam Technical University, Lucknow

October, 2025