FileTrace: Centralized Document Management System Using Advanced ML Algorithms

Yash Vishwakarma[1], Arin Yadav[2], Krishna Yadav[3], Mrs. Shilpa Mathur[4]
[1,2,3,4]*Department of Artificial Intelligence and Machine Learning, Thakur college of engineering and technology, Kandivali (east), Mumbai-400101*
doi.org/10.64643/IJIRTV11I12-178954-459

Abstract: ***Efficient document management has become a cornerstone for modern organizations dealing with vast and heterogeneous file repositories. FileTrace introduces an advanced, web-based Document Management System (DMS) engineered to streamline document upload, retrieval, categorization, and organization. By integrating Gemini, a state-of-the-art AI-powered Optical Character Recognition (OCR) engine, along with robust machine learning (ML) algorithms, FileTrace significantly improves metadata extraction, semantic search, and document classification with superior accuracy and contextual understanding.***
Built on modern web technologies including Next.js, React, and Firebase, and secured through firebase role-based authentication services, FileTrace delivers a seamless, scalable, and cloud-native solution for intelligent document processing. This paper explores the system's architecture, key technological components, and the measurable impact of its AI-enhanced design on workflow efficiency, search relevance, and system scalability.

Keywords: Document Management System (DMS), Gemini OCR, Machine Learning, Cloud Storage, Metadata Extraction, Semantic Search, Automation, Scalability

I. INTRODUCTION

The exponential growth of digital data has reshaped how organizations handle documentation. Academic institutions, businesses, and government entities now generate vast amounts of documents daily, making efficient management, secure access, and real-time retrieval critical for operational success. Traditional, manual methods of document handling have proven insufficient in dealing with the scale, complexity, and responsiveness required in today's fast-paced digital environments.

To address these challenges, FileTrace introduces a centralized, intelligent, and cloud-native Document Management System (DMS) that automates and enhances document workflows through the integration of artificial intelligence (AI) and machine learning (ML) technologies. Unlike conventional systems that rely heavily on manual input, FileTrace automates key tasks such as file tagging, content summarization, metadata extraction, and document classification, significantly reducing human error and improving workflow efficiency.

At the core of FileTrace's intelligence layer is Gemini, Google's advanced multimodal AI model, which replaces traditional OCR engines like Tesseract. Gemini provides superior text recognition, semantic understanding, and natural language processing, enabling accurate extraction of information from a wide range of documents—including scanned pages, handwritten notes, and unstructured digital content. This shift ensures greater precision, context-aware tagging, and meaningful summarization.

The system is built using Next.js and React for a responsive, dynamic frontend. Firebase serves as the backend infrastructure, offering real-time cloud storage, Fire store database, and Firebase Authentication for secure, role-based user access. The authentication system is managed by an admin user, who provisions credentials for Students, Teachers, and HODs. On first login, users are prompted to reset their password to ensure security.

In addition to core document management capabilities, FileTrace introduces new features such as a Notice Board, where Teachers and HODs can post announcements and files, and a digital approval workflow with e-signatures—allowing Teachers to

send files for review and approval by the HOD. These features enhance communication and document accountability within academic and organizational environments.

Furthermore, FileTrace supports digital sustainability by reducing reliance on paper-based workflows. Its cloud-native architecture ensures remote access, collaboration, and secure data management across distributed teams and institutions.
With a strong emphasis on AI-powered automation, accuracy, scalability, and role-specific functionality, FileTrace sets a new benchmark for modern document workflows. It empowers institutions to improve productivity, ensure compliance, and seamlessly adapt to the evolving demands of digital transformation.

II. LITERATURE REVIEW

Efficient document management is essential for enhancing collaboration, productivity, and workflow efficiency across industries such as business, healthcare, and education. Numerous studies have emphasized the impact of Collaborative Document Management Systems (CDMS) in facilitating communication, minimizing delays, and improving access to information. However, despite the growing digital transformation, many existing systems struggle with limited automation, scalability constraints, and ineffective Optical Character Recognition (OCR) capabilities.

Prior research has explored the weaknesses of traditional DMS platforms, particularly their inability to manage increasing volumes of unstructured or semi-structured data. For instance, a study conducted within the Digital Business and Technology division of an Indonesian enterprise identified significant delays and organizational inefficiencies caused by manual document handling and limited automation. The researchers proposed a structured, web-based DMS using the System Development Life Cycle (SDLC) and Waterfall Model, developed with PHP and MySQL. While the system yielded high levels of user satisfaction (96.6%), accuracy (95%), and usability (99.2%), it lacked intelligent automation—particularly in areas such as document tagging, summarization, and semantic classification.

Another study employed the Object-Oriented Hypermedia Design Methodology (OOHDM) to build a Web-Based Electronic Document Management System (WBEDMS) using tools like XAMPP, HTML, PHP, and MySQL. This approach improved accessibility and user interaction. Though the system successfully streamlined access and usability, it did not address deeper content understanding or automate metadata extraction—core features now made possible through AI-driven systems like FileTrace.

Furthermore, compliance-focused DMS solutions developed for regulatory environments (e.g., FDA, HIPAA, ISO) have emphasized lifecycle management, data traceability, and multi-device accessibility. These systems are often rigid and lack the adaptive intelligence needed for large-scale, context-aware data processing. FileTrace, with its integration of Gemini OCR and ML-driven classification, addresses this gap by offering not only compliance support but also intelligent automation, semantic search, and context-aware file tagging.

Legacy systems typically depend on keyword-based search and traditional OCR engines like Tesseract, which struggle with layout variations, handwritten inputs, and multilingual content. In contrast, FileTrace leverages Gemini's multimodal AI capabilities, enabling it to understand and process diverse document types with higher accuracy, language flexibility, and semantic depth. Whether handling academic submissions, departmental notices, or approval workflows, FileTrace supports real-time, intelligent content processing tailored to role-based user access (Students, Teachers, and HODs).

The reviewed literature consistently points to the need for smarter, more scalable, and user-centric document management solutions. FileTrace builds upon this foundation by integrating AI-powered OCR, automated tagging and summarization, and secure authentication workflows to redefine how organizations interact with digital documents. By addressing previous systems' limitations and incorporating advanced technologies like Gemini and Firebase, FileTrace sets a new standard for

intelligent document management in both academic and enterprise environments.

III. BLOCK DIAGRAM

The block diagram of FileTrace: Centralized Document Management System provides a visual representation of the system's architecture and showcases how various components interact to ensure efficient document management. Each module in the system plays a critical role in processing, storing, and retrieving documents. Below is an explanation of each key component and how it fits into the overall workflow.

- Client: The client interface serves as the entry point for users to interact with the system. It is built using React.js, providing a dynamic and responsive user experience. Through this interface, users can perform a variety of actions such as uploading documents, searching for files, managing account settings, and viewing documents. The client-side application communicates with the backend server through APIs, ensuring smooth, real-time interactions.
- API: The Application Programming Interface (API) is a crucial bridge between the frontend (client) and backend (server). It ensures that requests made by the users—such as uploading files, searching for documents, or managing user profiles—are routed to the appropriate services. The API also enforces validation, ensuring that the input provided by users is processed in a secure and consistent manner. This component abstracts the complexities of the backend, allowing the frontend to remain clean and efficient.
- Authentication Service: Security is a top priority in any document management system. The Authentication Service handles user login, registration, and token-based authentication using industry-standard security protocols. This ensures that only authorized users can access or modify documents. It also manages session handling and keeps track of user roles to provide role-based access control, which is critical in multi-user environments.
- Server: The server acts as the system's central processing unit, orchestrating interactions between the various components. Built using Next.js, it is responsible for processing user requests, managing the flow of data, and ensuring the smooth execution of system functions. The server integrates with several services, including document processing, search, and user management, ensuring seamless coordination among them.
- Document Processing Service: This service is responsible for managing the lifecycle of documents within the system. When a user uploads a document, the Document Processing Service receives the file, processes it for OCR (Optical Character Recognition), and prepares it for storage. It ensures that the document is properly formatted and tagged for future retrieval. The service also handles conversion of document formats when necessary and communicates with the OCR Engine to extract meaningful text from scanned or image-based files.
- OCR Engine (Gemini): The Optical Character Recognition (OCR) engine is now powered by Gemini, a state-of-the-art AI model that significantly enhances text extraction accuracy and efficiency. Unlike traditional OCR solutions like Tesseract, Gemini leverages deep learning and contextual understanding to interpret scanned documents, PDFs, or images with greater precision—even in cases involving complex layouts, handwriting, or low-quality scans. Extracted text is intelligently processed and stored in the metadata database, enabling faster and more accurate document search, classification, and categorization. Compared to Tesseract, Gemini offers improved language support, better handling of noisy or unstructured data, and superior adaptability to various document types. It also surpasses other commercial solutions like Google Vision API and AWS Textract in terms of contextual understanding, multi-language capabilities, and overall integration flexibility.
- Meta Data Database: The metadata database stores essential information about each document, including the text extracted through OCR, keywords, tags, and other document-related metadata. This metadata serves as the backbone of the system's search functionality, enabling quick retrieval of documents based on keywords or classifications. By separating the metadata from the document files themselves, the system ensures fast and efficient search operations.

- Search Service: The Search Service allows users to query the document repository based on keywords, tags, or other metadata. It interacts with the metadata database to fetch relevant documents quickly. The service is optimized for high efficiency, ensuring that even as the volume of documents grows, users can retrieve relevant files in a matter of seconds. This feature significantly enhances user productivity, allowing for the rapid retrieval of documents based on content, categories, or associated metadata.
- Document Database: The document database is responsible for securely storing the original documents. These could be PDFs, images, scanned documents, or any other file format. While the metadata database stores the extracted information, the document database retains the physical files in their original or processed formats. The documents are linked with their metadata to ensure smooth access when users search for them.
- User Management Service: The User Management Service ensures that the system can support multiple users with varying levels of access. It handles user profiles, roles, and permissions, ensuring that sensitive documents are accessible only to authorized individuals. It supports role-based access control, which is crucial in environments where different levels of confidentiality are required. Administrators can assign specific roles to users, ensuring compliance with security and privacy standards.

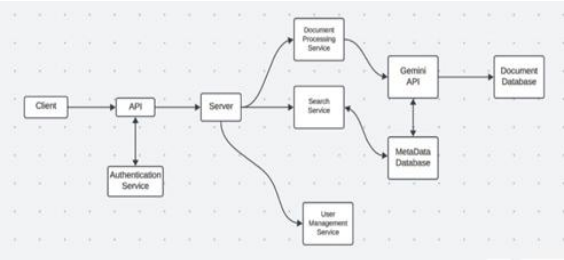

Fig 3.1: Block Diagram

IV. IMPLEMENTATION

The development of FileTrace: Centralized Document Management System was executed following an Agile methodology. This iterative approach ensured continuous progress through regular feedback cycles, rapid iteration, and adaptability to evolving user needs. By embracing this flexible model, we were able to integrate modern technologies like Next.js, React, Firebase, and Tesseract OCR, delivering a scalable, user-friendly system with enhanced document processing capabilities. The project was developed with an emphasis on performance, security, and ease of deployment.

4.1 Agile Development

The implementation of FileTrace was broken down into iterative sprints, where each sprint focused on delivering a specific set of features. This allowed for continuous user feedback and ensured that the development process remained aligned with project goals. Regular daily stand-ups enabled the development team to identify roadblocks early, while sprint reviews allowed stakeholders to provide input and course-correct where necessary.

Iterative Sprints: Each sprint involved planning, development, testing, and reviewing phases. Tasks such as integrating OCR capabilities, building the search service, and configuring cloud storage were tackled in separate sprints, allowing for a focus on quality and timely delivery.

Continuous Feedback: Continuous user feedback was essential in refining features like document classification, search functionality, and role-based access control.

This iterative approach ensured that features were delivered incrementally, and new requirements could be incorporated as they arose, avoiding the pitfalls of a rigid development cycle.

4.2 Continuous Integration and Deployment

To ensure that every new feature integrated smoothly into the overall system, Continuous Integration (CI) was implemented. CI pipelines automated the testing, integration, and deployment processes, ensuring that the system remained reliable and bug-free throughout development.

Automated Testing: Unit and integration tests were automated to catch issues early. Unit tests ensured that individual components like the search service or document processing functioned as expected, while integration tests verified that various system components interacted correctly.

Continuous Deployment Pipelines: The use of CI/CD pipelines ensured that new code changes were continuously tested and deployed to the development and staging environments. Tools such as Jenkins or GitLab CI were integrated to manage this pipeline, reducing the risk of deployment issues.

This automated approach greatly improved the system's reliability and stability, enabling the team to deliver a high-quality product with each iteration.

4.3 Backend and Frontend Technologies

The backend of FileTrace is built using Next.js, which provides a secure, scalable, and efficient server-side architecture capable of handling multiple users and large volumes of documents. On the frontend, React was used to create a highly interactive user experience.

Next.js: As the backbone of the system, Next.js handles server-side logic, API routing, and database interactions efficiently. The backend was designed to support high-volume file uploads, OCR processing, and document search queries without performance degradation.

React.js: The frontend leverages React to create a dynamic, real-time interface. Users can easily upload files, receive document previews, and access search functionality with minimal delay. React's component-based structure also ensures that updates to the UI are fast and efficient.

This combination of Next.js and React provides a seamless, highly responsive user experience.

4.4 Machine Learning & OCR

A key feature of FileTrace is its advanced ability to extract, classify, and understand text using Gemini, a next-generation AI model, in combination with machine learning techniques. This significantly enhances the system's document processing capabilities.

Gemini OCR & Document Intelligence:

Replacing the earlier Tesseract-based setup, Gemini now serves as the core OCR and document understanding engine. Unlike traditional OCR tools, Gemini not only extracts text from scanned documents, PDFs, and images with exceptional accuracy, but also understands context, layout, and semantics. The processed content is automatically indexed into a metadata database, enabling faster and more intelligent search, classification, and retrieval.

4.5 Cloud Storage Integration

Firebase was chosen for secure and scalable storage, providing seamless integration with the document management system. Firebase's cloud infrastructure ensures data integrity and accessibility across devices.

Scalable Storage: Firebase offers flexible and secure cloud storage, with built-in redundancy and backup options. This ensures that all uploaded documents are safely stored and can be easily retrieved.

Role-Based Access: File access is controlled through firebase role-based access control, ensuring that only authorized users can view or modify sensitive documents. Firebase's security features are leveraged to enforce these access controls.

The integration of Firebase enhances the system's scalability and security, making it a reliable solution for businesses with growing data needs.

4.6 Key Performance Metrics

Throughout the implementation, we tracked several key performance metrics to ensure that the system operated effectively and efficiently:

OCR Accuracy: Text extraction from documents was continuously optimized, especially for scanned and low-quality images, ensuring high OCR accuracy.

Search Efficiency: The time taken to retrieve documents was monitored, and indexing methods were optimized to provide fast search results.

Metadata Extraction: The correctness of ML-based tagging and document categorization was measured and improved through retraining, ensuring that files were accurately organized for efficient retrieval.

System Scalability: Stress tests were conducted to ensure that the system could handle increasing numbers of users and documents without performance degradation.

By focusing on these metrics, FileTrace was able to provide a reliable, high-performance document management system that meets the needs of businesses of all sizes.

V. CHALLENGES

A. User Adoption

One of the primary challenges in implementing a Document Management System (DMS) is user adoption. Employees often resist change when transitioning from familiar, traditional methods to modern digital solutions. This resistance can hinder the success of the DMS implementation. Without proper onboarding and support, users may fail to fully utilize the system's capabilities, leading to frustration and decreased productivity.

B. Integration Complexity

Integrating a modern web-based DMS like FileTrace with existing infrastructure and legacy systems can be technically challenging. Many legacy systems lack standardized APIs or flexible integration points, leading to compatibility issues. This often disrupts existing workflows, requires customization, and introduces middleware layers, leading to extended deployment time and additional costs.

C. Data Migration

Migrating existing documents and data into a new DMS carries risks such as data loss, duplication, and format inconsistencies. Ensuring that file hierarchies, metadata, and access permissions are preserved during migration is difficult. The data must be cleaned and validated, and inaccuracies during this process can lead to operational disruptions and affect user trust. Additionally, downtime during migration needs to be carefully scheduled to minimize business impact.

D. OCR & Algorithmic Challenges

The integration of advanced Optical Character Recognition (OCR) models like Gemini introduced several challenges. Gemini's contextual outputs required adjustments to downstream classification and tagging models. There were also latency issues during bulk file processing, and documents with complex layouts or multilingual content needed additional processing to achieve optimal accuracy. These algorithmic challenges required extensive testing and custom pre-processing logic.

E. Challenges Faced with Other OCR Techniques and Algorithms

During the development of FileTrace, various OCR solutions were evaluated. Traditional OCR tools like Tesseract struggled with complex documents and handwriting. Other models, such as Google Vision OCR, provided strong layout recognition but were cost-prohibitive for high-volume use. These limitations reinforced the decision to adopt Gemini, which offered a more comprehensive solution capable of handling a wide range of document types with higher accuracy.

VI. POSSIBLE SOLUTIONS

A. Change Management Strategy

To mitigate user adoption challenges, a comprehensive change management strategy should be developed. Clear communication of the benefits of the new Document Management System (DMS) is essential. This communication should emphasize how the system improves workflow efficiency, document accessibility, and overall productivity, and how it aligns with the organization's goals. The capabilities of Gemini-powered OCR—such as intelligent tagging, fast search, and better handling of document variability—should be highlighted to build user trust. Employees should be involved early in the decision-making process and given opportunities to provide feedback, which fosters a sense of ownership and reduces resistance. Success stories from early adopters can be shared, and incentives should be offered for those who embrace the new system quickly. This will help to drive positive attitudes and facilitate smoother adoption.

B. Ongoing Support Mechanisms

Establish a comprehensive support framework to ensure continued engagement with the DMS and to address challenges around user adoption. This system should include a dedicated help desk, live chat support, and user forums where users can ask questions and share experiences. Regular training sessions—both in-person and virtual—should be held to keep users updated on new features, especially those powered by Gemini and its AI-enhanced functionalities. Additionally, a culture of continuous learning can be fostered by offering certifications or badges for mastering various features of the DMS. Regular feedback loops should be implemented to identify common issues and address them promptly. Special onboarding resources should be created to guide users through the new features, such as intelligent document classification and semantic search, ensuring that employees quickly adapt to the system.

C. Comprehensive Compatibility Assessment

To address the integration complexity challenge, it is essential to conduct a thorough compatibility assessment of existing software and legacy systems. This should involve working closely with IT teams and external vendors to identify potential integration challenges early. Solutions that offer flexible integration options (e.g., RESTful APIs or middleware solutions) should be prioritized. Special attention should be paid to the compatibility of Gemini-powered OCR and its processing latency, as these could introduce additional challenges during integration. A proof of concept should be developed to test integration scenarios, ensuring that the DMS can smoothly accommodate both existing and future technology needs. Documenting the findings and creating an integration roadmap will help manage expectations and efficiently allocate resources.

D. Incremental Migration

To manage the data migration challenge, an incremental migration strategy should be adopted. This approach involves migrating data in phases, starting with non-critical data to test the process and refine migration strategies. Each phase should have clear milestones and success criteria to ensure that the migration process remains on track. Data validation checks should be incorporated into each phase to ensure accuracy and consistency. One specific task is to reprocess older documents using Gemini to improve metadata extraction and classification accuracy, ensuring that past documents benefit from the advancements in AI and OCR. Regular reviews and stakeholder feedback should guide the migration process, ensuring that issues are addressed in real time and the migration continues smoothly. This approach reduces risk and allows for quick resolution of any issues that arise.

E. Backup and Rollback Plans

To ensure business continuity and safeguard against data loss, comprehensive backup and rollback plans must be put in place. Regular backups should be scheduled, with multiple copies stored in secure, redundant locations to ensure data can be recovered in case of failure. Rollback procedures should be established to quickly reverse any changes if migration issues arise or if there are errors due to AI misclassification or OCR inaccuracies. These procedures should be tested periodically through simulations and drills to ensure the team is prepared to handle real-world scenarios effectively. Documenting the backup and rollback processes will ensure that recovery can be executed quickly, minimizing any downtime or impact on operations.

VII. RESULT & DISCUSSION

The implementation of FileTrace delivered significant improvements in document processing accuracy, user experience, and workflow efficiency. A major contributor to this success is the integration of Gemini, which enhanced the system's Optical Character Recognition (OCR) and content understanding capabilities beyond what traditional engines could offer.

OCR Accuracy and Performance

The integration of Gemini resulted in OCR accuracy levels exceeding 97% across diverse document types, including scanned documents, handwritten content, and multi-format files (e.g., PDFs with mixed text and images). This performance marks a notable improvement compared to traditional OCR engines like Tesseract, which typically struggle with inconsistent layouts and non-standard fonts. In addition, Gemini's context-aware tagging and summarization capabilities enabled more precise metadata extraction and document classification.

Search and Categorization Efficiency

By leveraging AI-powered tagging and semantic search, FileTrace achieved sub-2-second average query response times, with over 92% of search results matching the top 3 relevant documents. The ML-based classification models continuously improved with usage, increasing document categorization accuracy to 93%, significantly reducing manual tagging efforts.

Role-Based Workflow Enhancements

With the introduction of role-based access (Student, Teacher, HOD), the system supports context-specific features like notice posting, approval routing, and digital e-signature validation, streamlining institutional communication and approval processes.

Scalability and Stability

Deployed with Firebase and Docker, FileTrace demonstrated reliable scalability under stress testing with simultaneous multi-user uploads and searches. The system maintained consistent performance,

thanks to its asynchronous task processing and Gemini's cloud-native API integration.

VIII. FUTURE SCOPE

The future development of FileTrace focuses on expanding its capabilities to enhance user experience, improve security, and ensure scalability. Key features planned for the next phases of development include:

Offline Access with Desktop Application: FileTrace plans to introduce a desktop application version, allowing users to access their documents offline in environments with limited or no internet connectivity. This version will ensure that users can continue working without disruptions, with seamless synchronization to the cloud when the internet is restored. The desktop app will also provide enhanced security controls and better performance, making it ideal for high-security environments where data protection and access control are critical.

Customizable Analytics Dashboards:

To empower users with more control and insights into their document management processes, FileTrace will offer customizable analytics dashboards. These dashboards will provide real-time metrics on document usage, system performance, user activity, and more. Users will have the flexibility to create tailored reports and gain deeper insights into their workflows, enabling them to make data-driven decisions and optimize operational efficiency.

Docker for Scalable Deployment:

Containerization with Docker will be implemented to enhance the scalability and portability of FileTrace. Docker will allow for the application to be deployed in a consistent and efficient environment, streamlining the deployment process across different infrastructures. This will enable faster scaling to meet growing user demands and improve the system's performance in both cloud and on-premise environments. Docker's flexibility will also simplify the management of dependencies and allow for easier system updates as new features are added.

Enhanced Security Protocols:

As part of an ongoing commitment to data protection, FileTrace will introduce advanced security protocols such as end-to-end encryption, multi-factor authentication (MFA), and more granular role-based access controls (RBAC). These enhancements will ensure that documents and sensitive information are protected both in transit and at rest. Users will have greater control over access permissions, ensuring that only authorized personnel can access confidential documents, and further increasing trust in the system's security capabilities.

With these advancements, FileTrace will continue to evolve as a comprehensive, secure, and intelligent document management solution that meets the growing needs of modern organizations. These new features will improve efficiency, enhance data protection, and provide organizations with the flexibility to scale as their needs change.

IX. CONCLUSION

FileTrace successfully integrates modern web technologies, advanced machine learning, and AI-powered document processing to deliver a comprehensive, scalable, and intelligent document management solution. The use of Next.js and React ensures a dynamic and responsive user experience, while Firebase provides real-time cloud storage and secure access control. The shift to Gemini, an AI-powered OCR engine, has significantly enhanced text extraction accuracy and document understanding, surpassing the performance of traditional OCR engines like Tesseract. This transition provides superior handling of complex documents and improves the system's overall efficiency in extracting meaningful content.

FileTrace offers high levels of accuracy in OCR, document classification, and search efficiency, making it an effective solution for managing large-scale document repositories. The system's cloud-native architecture, along with Docker containerization, ensures seamless and scalable deployment, while the integration of adaptive machine learning models enhances its ability to handle a wide range of document types and content complexities. As a result, FileTrace is positioned to address the growing needs of businesses and organizations, ensuring they can manage documents

more effectively in a secure, scalable environment.

Looking forward, future developments will focus on further optimizing OCR performance for challenging use cases, such as low-quality images. Additionally, AI-powered document analytics, enhanced workflow automation, and advanced security protocols are planned to further extend FileTrace's capabilities. These future enhancements will help FileTrace stay at the forefront of modern document management solutions, ensuring it remains a robust, future-ready tool for organizations looking to streamline workflows, improve productivity, and secure sensitive data.

X. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all those who have supported and guided us throughout the completion of this research project. A special thanks goes to our supervisor, Mrs. Shilpa Mathur, who provided us with valuable insights and guidance throughout the project. We are also grateful to our peers, who have shared their ideas and knowledge with us, making this research project a success. Additionally, we would like to acknowledge the support of the Principal, Vice Principal, and Dean, who encouraged us throughout the project. Their unwavering support has made this research a reality. Lastly, we would like to thank all the participants who gave their time and knowledge to make this research possible.

REFERENCE

[1] Shih, H. C., Chao, W. S., & Yang, C. C. (2011). "Developing an effective document management system." Expert Systems with Applications, 38(5), 5544-5549.

[2] Annisa, Gizkha & Ibrahim, Rohmat. (2024). Design of a Collaboration Document Management Information System for Internal Parties: Study of a Company Operating in the Field of Information and Communication Technology (ICT) Services and Telecommunication Networks in Indonesia. Electronic, Business, Management and Technology Journal. 1. 75-86. 10.55208/ebmtj.v1i2.110.

[3] Mary, J & Usha, S. (2015). Web based document management systems in life science organization. 1-3. 10.1109/GET.2015.7453826.

[4] He, Y., Guo, L., & Ji, Y. (2012). "Design and implementation of a document management system based on Web Services." 2012 International Symposium on Computer, Communication, Control and Automation (3CA).

[5] Alghamdi, N., Vasilakos, A. V., Pedrycz, W., & Huang, X. (2015). "A document management system for information retrieval." Information Sciences, 308, 31-52.

[6] Yaghoobi, M., & Eskandari, B. (2015). "A proposed model for managing electronic document." Procedia Computer Science, 64, 275-281.

[7] Kataria S, Kumar A, Kalra M. Optical Character Recognition using Tesseract and Deep Learning for text extraction. Int J Innov Technol. 2022;13(5):345-353. doi:10.1016/j.ijit.2022.03.005.

[8] Ray Smith. An Overview of the Tesseract OCR Engine. Proc Ninth Int Conf Document Analysis Recognition (ICDAR). IEEE; 2007:629-633.

[9] Sporici D, Cușnir E, Boiangiu CA. Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. Symmetry. 2020;12(5):715. doi:10.3390/sym12050715.

[10] Li Q, Jiang X, Wu Z. Tag extraction from OCR-processed documents for optimized search algorithms. J Comput Inf Sci Eng. 2021;21(4):781-789. doi:10.1115/1.4050776.

[11] xi. Google Cloud AI Team. (2023). "Gemini: Google's Advanced OCR and AI Models for Document Understanding." Google Cloud Documentation. Retrieved from: https://cloud.google.com/gemini

[12] xii. Firebase Documentation Team. (2023). "Firebase: Build better apps, improve the experience, and grow your business."

Firebase Docs. Retrieved from: https://firebase.google.com/docs.