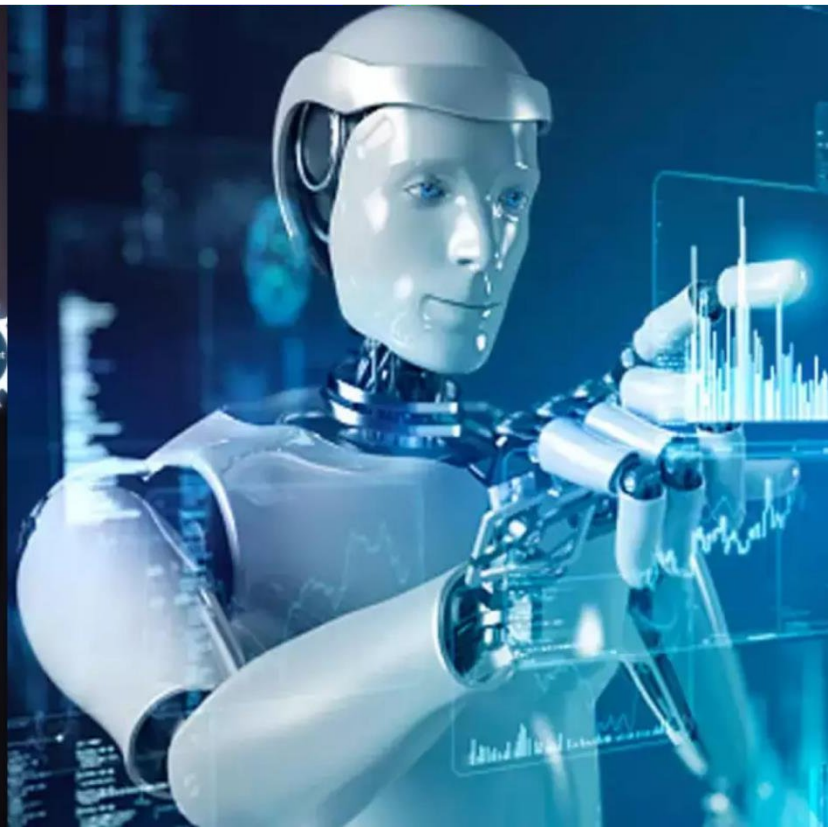
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 15, Issue 4, April 2026

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

To Implement a Web-Based Intelligent System for Knowledge Extraction from Books

Dickson David R[1], Arun Kumar C[2], Bhavan A[3], Navakrishnan T[4]

Assistant Professor, Department of Computer Science and Engineering, P.S.V College of Engineering and Technology, Krishnagiri Dt., Tamil Nadu, India[1]

UG Scholars, Department of Computer Science and Engineering, P.S.V College of Engineering and Technology , Krishnagiri Dt., Tamil Nadu, India[2,3,4]

ABSTRACT: This project, titled *"To Implement a Web-Based Intelligent System for Knowledge Extraction from Books,"* presents an advanced AI-powered platform designed to process and analyze PDF documents efficiently. The system enables users to upload documents and automatically perform text extraction, summarization, translation, and text-to-speech conversion through an intuitive web interface. It integrates Natural Language Processing (NLP) techniques to generate meaningful summaries and supports multilingual processing, including regional languages such as Tamil, using Optical Character Recognition (OCR) for scanned documents. The system also provides audio outputs for both original and translated content, enhancing accessibility and user interaction. By combining multiple intelligent features into a single platform, the proposed solution reduces manual effort, improves productivity, and offers a scalable approach for academic, research, and professional applications in document analysis and knowledge extraction.

KEYWORDS: Knowledge Extraction, Document Processing, PDF Extraction, Summarization, NLP, Translation, Text-to-Speech, Django Domain: AI

I. INTRODUCTION

The rapid increase in digital documents has created a need for intelligent systems capable of efficiently extracting meaningful information from large volumes of text. This project, titled *"To Implement a Web-Based Intelligent System for Knowledge Extraction from Books,"* presents an AI-based solution that allows users to upload PDF documents and automatically perform text extraction, summarization, translation, and text-to-speech conversion. By integrating Natural Language Processing (NLP) and Optical Character Recognition (OCR), the system supports multilingual content, including Tamil, and ensures accurate processing of both digital and scanned documents. The platform is designed to be user-friendly, efficient, and scalable, making it highly useful for academic, research, and professional applications.

II. LITERATURE REVIEW

''AI-Based Text Summarization Using Transformer Models" This study focuses on using transformer-based models such as BERT and BART for automatic text summarization. The system reduces large text into concise summaries while preserving meaning. It highlights how NLP techniques improve the efficiency of processing long documents.

"Optical Character Recognition for Multilingual Document Processing" This research explains the use of OCR techniques for extracting text from scanned and multilingual documents. It emphasizes the importance of handling regional languages and image-based PDFs, which are common in real-world applications.

"Web-Based Intelligent Document Processing Systems" This research presents a web-based system that integrates multiple AI functionalities such as text extraction, summarization, and translation. It demonstrates how combining these features into a single platform improves usability and productivity.

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

"**Neural Machine Translation for Language Conversion**" This work discusses the implementation of neural machine translation models to convert text between different languages. It shows how translation systems can improve accessibility and enable users to understand content in their preferred language.

III. METHODOLOGY

A. EXISTING SYSTEM

The existing systems for document processing mainly include traditional PDF readers, standalone summarization tools, translation applications, and text-to-speech systems. These systems operate independently and require users to switch between multiple platforms to perform different tasks such as reading, translating, and converting text to audio. Most of these systems lack intelligent automation and do not provide integrated solutions for extracting meaningful knowledge from large documents. As a result, users face difficulty in efficiently managing and understanding extensive textual data.

B. DISADVANTAGE

1. The system performance depends on internet connectivity for translation and text-to-speech services. Slow or unstable internet may affect processing speed and output generation.
2. The accuracy of text extraction may vary for low-quality or highly complex scanned PDF documents. OCR may produce errors when handling unclear text or special fonts.
3. The system may take more processing time for large PDF with many pages .This can lead to delays in generating summaries, translations, and audio outputs.

C. PROPOSED SYSTEM

The proposed system is a web-based intelligent platform designed to automate knowledge extraction from PDF documents using AI techniques. It allows users to upload documents and automatically performs text extraction, summarization, translation, and text-to-speech conversion within a single interface. The system also integrates Optical Character Recognition (OCR) to handle scanned and multilingual documents, including Tamil. By combining multiple functionalities into one system, it enhances efficiency, reduces manual effort, and provides a seamless user experience.

D. ADVANTAGES

1. Integrates multiple features into a single platform
2. Reduces time and effort in document processing
3. Supports multilingual translation including regional languages
4. Provides audio output for better accessibility
5. Handles both digital and scanned PDF documents
6. User-friendly and efficient interface

E. DESIGN OF THE SYSTEM

The system is designed using a modular and layered architecture to ensure efficient processing and scalability. It consists of a user interface layer for interaction, a document processing layer for handling uploads and storage, and an AI processing layer that includes text extraction, summarization, translation, and text-to-speech modules. The system also incorporates a storage layer to manage user data, documents, and generated outputs. The data flows sequentially from user input through processing modules to the output layer, ensuring smooth and accurate knowledge extraction.

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

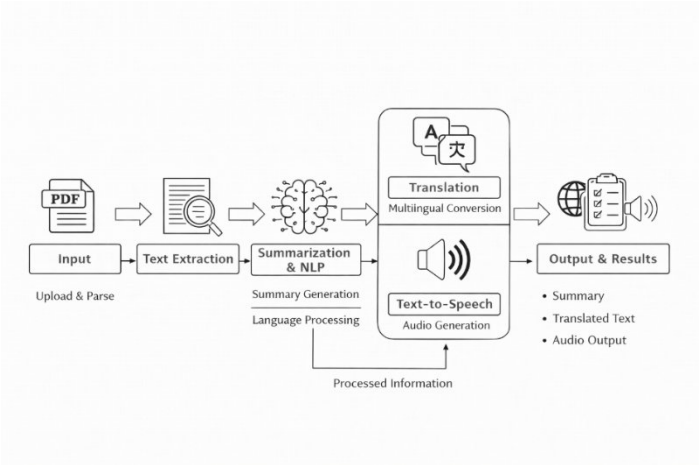


Fig.1 The system takes a PDF, extracts text, processes it using AI for summarization and translation, and finally provides both text and audio output to the user.

IV. IMPLEMENTATION

MODULE DESCRIPTION

1. Document Upload Module

This module allows users to upload PDF documents into the system through a web interface. It handles file validation, storage, and management, ensuring that the uploaded files are securely saved in the system for further processing.

2. Text Extraction Module

This module extracts textual content from the uploaded PDF documents. It uses tools like PyMuPDF for digital PDFs and Optical Character Recognition (OCR) for scanned or image-based PDFs, including multilingual content such as Tamil.

3. Text Preprocessing Module

This module cleans and prepares the extracted text for further processing. It removes unwanted characters, normalizes the text, and structures it in a format suitable for AI-based analysis and processing.

4. Summarization Module

This module uses Natural Language Processing (NLP) techniques to generate a concise summary of the extracted text. It reduces large content into meaningful and shorter information while preserving the key ideas.

5. Translation Module

This module converts the summarized text into different languages based on user selection. It supports multilingual processing, enabling accessibility for users from different language backgrounds.

6. Text-to-Speech (TTS) Module

This module converts textual content into audio format using text-to-speech technology. It generates audio outputs for both original and translated text, improving accessibility and user experience.

7. Storage Module

This module manages the storage of uploaded documents, extracted text, summaries, translations, and audio files. It uses a database and media storage system to efficiently organize and retrieve data.

8. User Interface Module

This module provides an interactive web interface for users to upload documents, view summaries, listen to audio outputs, and access translated content. It ensures a smooth and user-friendly experience.

V. RESULT

The developed system was successfully tested with multiple PDF documents of varying sizes and formats, including both digital and scanned files. The system effectively extracted text using PyMuPDF and Optical Character Recognition (OCR) for multilingual content such as Tamil. The summarization module generated meaningful and concise summaries that captured the key information from large documents. The translation module accurately

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

converted the summarized text into different languages based on user input, and the text-to-speech module successfully generated audio outputs for both original and translated content. The system demonstrated stable performance for documents with multiple pages by implementing optimized text handling techniques. The results indicate that the proposed system significantly reduces the time and effort required for manual reading and understanding of large documents. The integration of multiple features into a single platform provided a seamless and efficient user experience.

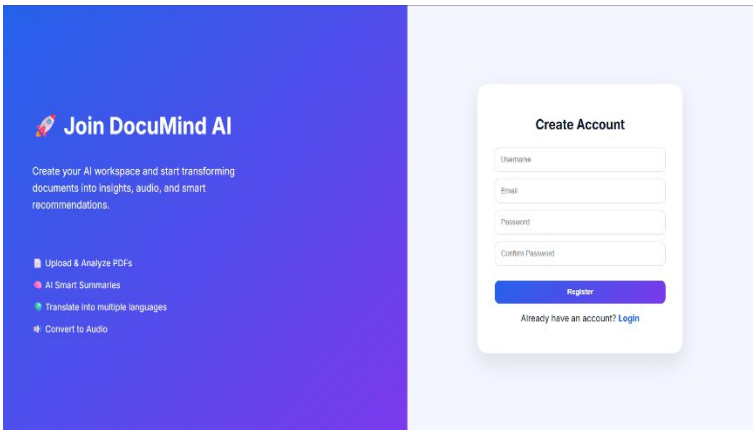


Fig 2: Register

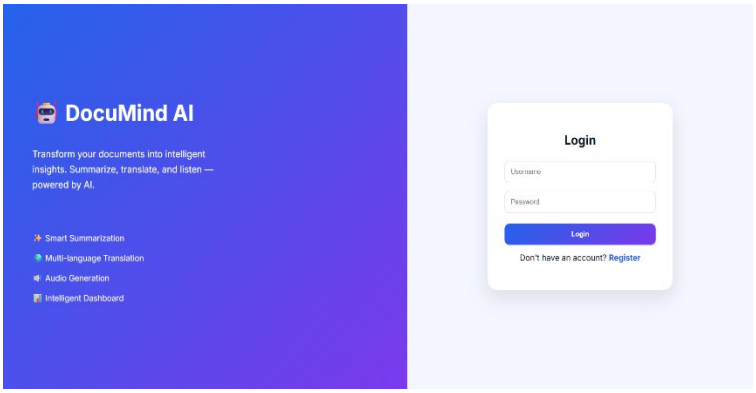


Fig 3: Login

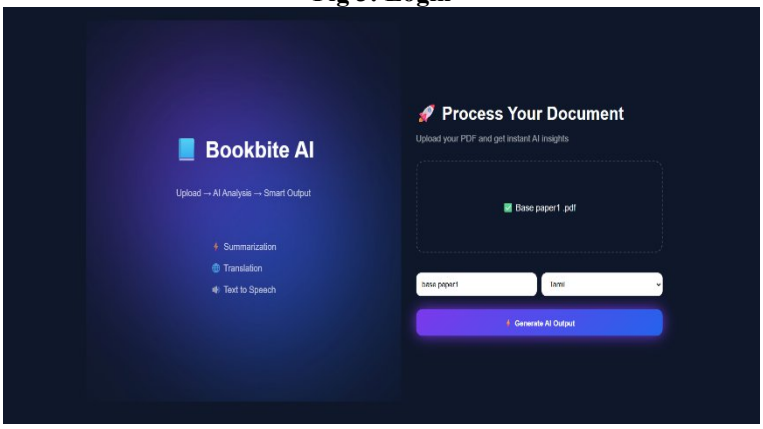


Fig 4: Document Uploade

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

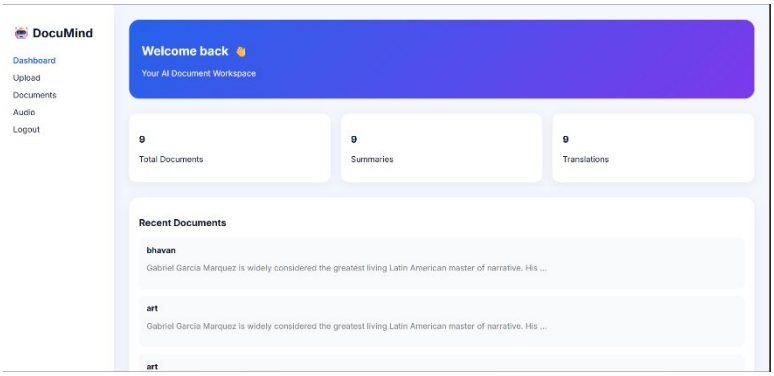
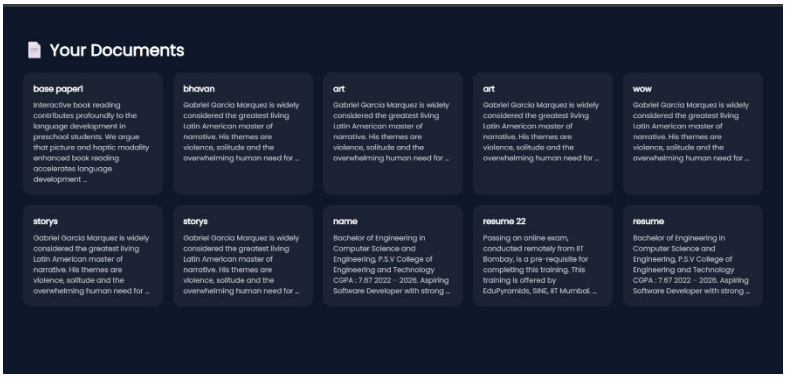


Fig 5: Dashboard



The system showed good accuracy in text extraction and summarization while maintaining usability across different document types. Overall, the implementation proved to be effective in achieving the objective of intelligent knowledge extraction, making it highly suitable for academic, research, and professional applications.

VI. CONCLUSION

The project titled *"To Implement a Web-Based Intelligent System for Knowledge Extraction from Books"* successfully demonstrates the application of Artificial Intelligence in automating document processing tasks. The system effectively integrates multiple functionalities such as text extraction, summarization, translation, and text-to-speech conversion into a single platform. It simplifies the process of understanding large volumes of textual data by converting them into concise summaries and accessible audio outputs. The inclusion of OCR enables the system to handle scanned and multilingual documents, including regional languages like Tamil, thereby increasing its usability and scope. Furthermore, the system provides a user-friendly interface that enhances interaction and accessibility for different types of users, including students, researchers, and professionals. By reducing manual effort and improving processing efficiency, the proposed system offers a practical and scalable solution for modern document analysis. Overall, the project highlights the potential of AI-driven technologies in transforming traditional reading and information extraction methods into intelligent and automated processes.

VII. FUTURE WORK

The proposed system can be further enhanced by incorporating advanced AI models, such as transformer-based architectures, to improve the accuracy and contextual understanding of text summarization. The system can also be extended to support a wider range of regional and international languages, thereby increasing its usability across diverse user groups. Improvements in the text-to-speech module using more natural and expressive voice synthesis techniques can significantly enhance user experience. Additionally, optimizing the system for real-time processing will reduce latency and enable faster handling of large documents. Future development may also include integration with

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 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504214|

cloud platforms for better accessibility, implementation of a recommendation system for suggesting related content, and support for multiple file formats such as images and Word documents. Enhancing security features and developing a mobile application version of the system can further improve scalability, usability, and data protection, making the system more robust and adaptable for real-world applications.

REFERENCES

[1] A. Vaswani *et al.*, "Attention Is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.
[3] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, 2020.
[4] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," in *Proc. ACL*, 2020.
[5] J. Zhang *et al.*, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," in *Proc. ICML*, 2020.
[6] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," in *Proc. EMNLP*, 2019.
[7] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP*, 2020.
[8] K. Papineni *et al.*, "BLEU: A Method for Automatic Evaluation of Machine Translation," *Computational Linguistics*, 2002 (widely used evaluation method).
[9] P. Koehn, "Neural Machine Translation," *Cambridge University Press*, 2020.
[10] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 3rd ed., *Prentice Hall*, 2021.
[11] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proc. ICDAR*, 2007.
[12] A. Graves, "Supervised Sequence Labelling with Recurrent Neural Networks," *Springer*, 2012.
[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, 1997.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016.
[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proc. NeurIPS*, 2014.
[16] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners," *OpenAI Report*, 2019.
[17] J. Brown *et al.*, "Language Models are Few-Shot Learners," in *Proc. NeurIPS*, 2020.
[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, 2015.
[19] T. Mikolov *et al.*, "Efficient Estimation of Word Representations in Vector Space," in *Proc. ICLR*, 2013.
[20] F. Chollet, "Deep Learning with Python," *Manning Publications*, 2018.

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