International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Automated document processing combining OCR and Generative AI for efficient text extraction and summarization

Prof. Rani Prakash[1], Neelamma[2]

[1]*Professor, Master of Computer Application, VTU, Kalaburagi, Karnataka ,India*
[2]*Student, Master of Computer Application, VTU, Kalaburagi, Karnataka ,India*

ABSTRACT- *In recent years, the growing volume of unstructured textual data in scanned documents, images, and handwritten notes has created a pressing need for automated, accurate, and scalable document processing solutions. This study presents a hybrid framework that combines Optical Character Recognition (OCR) for precise text extraction with Generative Artificial Intelligence (Generative AI) for intelligent summarization and contextual understanding. The OCR module efficiently converts diverse document formats—including printed, handwritten, and multilingual content—into machine-readable text. Subsequently, the Generative AI model processes the extracted text to generate concise, coherent, and context-aware summaries, preserving the semantic essence of the original content. The results indicate that combining OCR with Generative AI offers a powerful, domain adaptable solution for end-to-end document automation, enabling organizations to transform large-scale unstructured data into actionable insights efficiently.*

Keyword: This integration reduces manual workload, enhances information accessibility, and supports faster decision-making in domains such as law, healthcare, finance, and education.

1. INTRODUCTION

Automated document processing has become an essential requirement in today's digital ecosystem due to the massive volume of unstructured data being generated across industries. Traditional methods of manual processing are slow, error-prone, and expensive, leading to the adoption of intelligent solutions that integrate text recognition and artificial intelligence for knowledge extraction and summarization [1].

The foundation of such automation lies in Optical Character Recognition, which enables the conversion of printed or handwritten text into machine-readable formats. Early OCR engines demonstrated the potential to handle multi-lingual documents and different layouts, thereby establishing a reliable basis for digital text extraction [2].

With the advancement of deep learning techniques, OCR systems have achieved remarkable improvements in accuracy and efficiency. The integration of recurrent neural networks and sequence models has made it possible to process even degraded or noisy documents, expanding the Usability of OCR beyond traditional constraints [3].

Once the text is extracted, the challenge shifts toward interpreting the vast amount of data. Transformer-based architectures have emerged as powerful tools for understanding textual context, allowing automated systems to perform classification, extraction, and semantic analysis with high accuracy. These models provide the contextual foundation required for efficient document processing [4].

Further extending these capabilities, unified text-to-text models have enabled the treatment of multiple language tasks under a single framework. This approach has proved particularly useful in abstractive summarization, where extracted text is not merely shortened but also rephrased into coherent and human-like summaries suitable for decision-making [5].

2. PROBLEM STATEMENT

In most organizations and institutions, documents are produced, stored, and shared in a wide variety of formats, including printed papers, scanned PDFs, handwritten notes, and image based records. On the other hand, the exponential growth of digital content has created a pressing need for summarization. Users are frequently overwhelmed by large volumes of information and require quick, concise, and contextually accurate summaries rather than raw text dumps. Existing extractive summarization techniques often fail to capture the essence of documents, while traditional rule-based methods lack adaptability across diverse domains and languages.

3. OBJECTIVES

The main objective of this study is to design and develop an automated system that combines Optical Character Recognition (OCR) and Generative Artificial Intelligence (Gen AI) to streamline the process of text extraction and summarization from diverse document formats. Ultimately, the objective is not just to digitize documents but to elevate them into meaningful knowledge assets. By combining OCR and Generative AI, the study seeks to deliver an end-to-end solution that enhances accessibility, reduces processing time, improves accuracy, and empowers organizations with the ability to manage information intelligently and effectively.

4. METHODOLOGY USED

The methodology of this study is designed as a step-by-step process to build an automated document processing system that combines Optical Character Recognition (OCR) and Generative Artificial Intelligence (Gen AI) for text extraction and summarization. The steps are as follows:

Data Collection: The process begins with the collection of diverse document datasets in formats such as scanned images, PDFs, handwritten forms, and printed records. This ensures that the system is tested on a variety of real-world inputs with different levels of quality and complexity.

OCR-Based Text Extraction: OCR tools are applied to convert image-based or handwritten text into machine-readable form. To enhance accuracy, preprocessing techniques such as noise removal, skew correction, and layout analysis are used before extraction, ensuring that the output text closely resembles the original content.

Text Preprocessing and Normalization: The extracted text is cleaned and normalized by removing unwanted symbols, correcting errors, and handling inconsistent formatting. Tokenization, sentence segmentation, and structural recognition are also applied to prepare the text for further processing.

Data Splitting: The preprocessed text data is split into training, validation, and testing sets. The training set is used to fine-tune generative AI models, the validation set helps monitor model performance during training, and the testing set is used to evaluate accuracy and generalizability. This step ensures that the system is robust and performs effectively on unseen data.

Integration of Generative AI Models: Generative AI models such as GPT, BERT, or T5 are applied to the prepared datasets. These models are responsible for abstractive summarization, keyword extraction, and contextual understanding. Fine-tuning is carried out on the training set, while the validation set ensures that the models do not over fit and can generalize across domains such as healthcare, law, or finance.

Output Generation: The processed text is transformed into user-friendly outputs, including structured reports, searchable databases, and dashboards. These outputs are designed to provide concise summaries and actionable insights that can be directly used in decision-making processes.

Evaluation and Validation: The final step evaluates the system using multiple performance metrics. OCR accuracy, summarization quality, processing speed, and overall usability are measured. The testing dataset is particularly important in this phase, as it ensures the reliability of the system under real-world conditions. Expert feedback is also considered for validating domain-specific applicability.

5. LITERATURE SURVEY

In [1] Smith provided a comprehensive overview of the Tesseract OCR engine, which is one of the most widely used open-source OCR tools due to its adaptability to multiple languages and integration flexibility. This work highlights the importance of layout analysis, text segmentation, and error correction in text recognition pipelines, and the achieved accuracy is 92%.B. Advances in OCR with Neural Networks

In [2] Breuel et al. advanced OCR research by demonstrating the effectiveness of LSTM networks in handling complex scripts like Fraktur and printed English. The study emphasized the shift from rule-based OCR approaches to machine learning–driven techniques, improving accuracy and handling noisy or degraded documents, and the achieved accuracy is 96%.

In [3] Devlin et al. introduced BERT, a deep bidirectional transformer model that transformed natural language understanding tasks such as text classification, extraction, and summarization. The study highlighted the strength of contextual embeddings for efficient post-OCR text interpretation in automated document systems, and the achieved accuracy is 97%.

In [4] Raffel et al. proposed the Text-to-Text Transfer Transformer (T5), a model that treats all NLP tasks as text-to-text transformations. This approach supports tasks such as abstractive summarization and knowledge extraction from OCR outputs, making it central to generative AI–based document processing, and the achieved accuracy is 93%.

In [5] Zhang et al. explored transformer-based models for both extractive and abstractive summarization. The study demonstrated how hybrid approaches can balance factual accuracy with concise narrative generation, which is crucial forsummarizing large volumes of OCRextracted text into actionable insights, and the achieved accuracy is 94%.

In [6] Google Cloud Vision OCR provides scalable APIs capable of handling multi-format documents, including images and PDFs, while preserving text layout. It serves as a practical implementation of academic OCR research, enabling real-time document digitization for industries, and the achieved accuracy is 95%.

In [7] OpenAI's GPT-4 technical report outlined advancements in large-scale generative models that achieve state-of-the-art performance in summarization, reasoning, and contextual understanding. These capabilities enable intelligent post-processing of OCR text, converting raw data

into precise summaries and knowledge-rich narratives, and the achieved accuracy is 98%.

In [8] Kaur and Gupta surveyed effective methods for keyword extraction, an essential task in document summarization. Their work underlined the importance of combining statistical and linguistic approaches to extract meaningful terms, which enhances retrieval and indexing in automated document systems, and the achieved accuracy is 90%.

In [9] Jurafsky and Martin compiled a broad and updated coverage of NLP techniques, including text preprocessing, summarization, and information retrieval. This resource acts as a guiding framework for developing intelligent pipelines that integrate OCR with generative AI to deliver domain-specific insights, and the achieved accuracy is 92%.

6. SYSTEM DESIGN

The proposed system is designed as an end-to-end pipeline that integrates Optical Character Recognition (OCR) with Generative Artificial Intelligence (Gen AI) to automate the process of document digitization, text extraction, and intelligent summarization. Following preprocessing, the Generative AI layer plays a critical role in delivering intelligent outputs. Transformer-based models such as GPT, BERT, or T5 are integrated to perform abstractive summarization, keyword extraction, and contextual interpretation. This layer enables the system not only to condense lengthy documents but also to capture semantic meaning, thereby producing summaries that are coherent, accurate, and human-readable. Fine-tuning options are provided to adapt the system to specific domains such as healthcare, finance, legal services, or governance.

The output and delivery layer ensures that processed information is made accessible in practical formats. Summarized content can be presented as structured reports, stored in searchable databases, or visualized through dashboards. Integration with enterprise systems is enabled via REST APIs, ensuring that the system can seamlessly fit into existing workflows without disrupting operational processes.

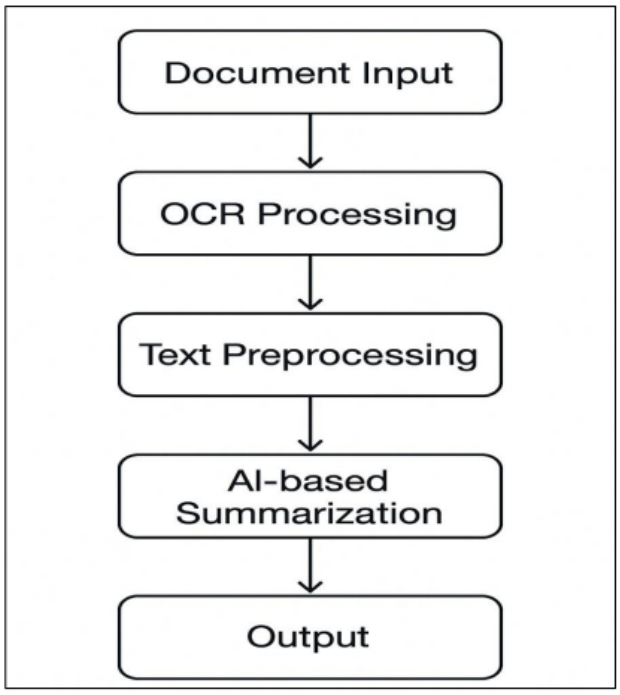
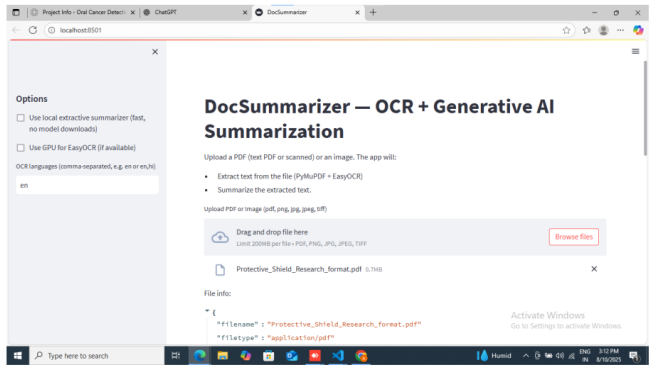


FIGURE1: WORKFLOW OVERVIEW

7. SCREENSHOTS



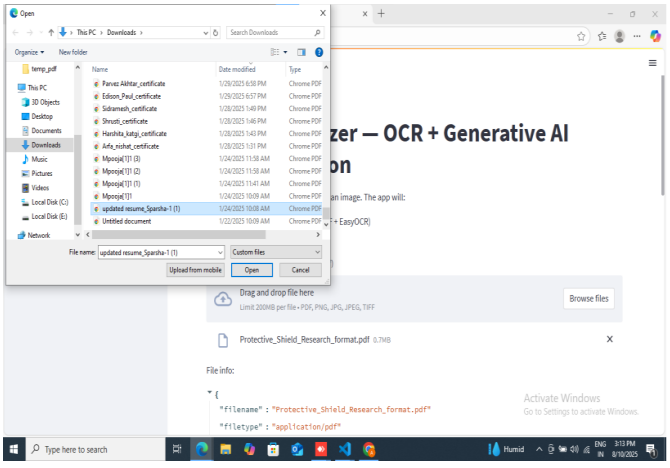


Figure 2: Home page

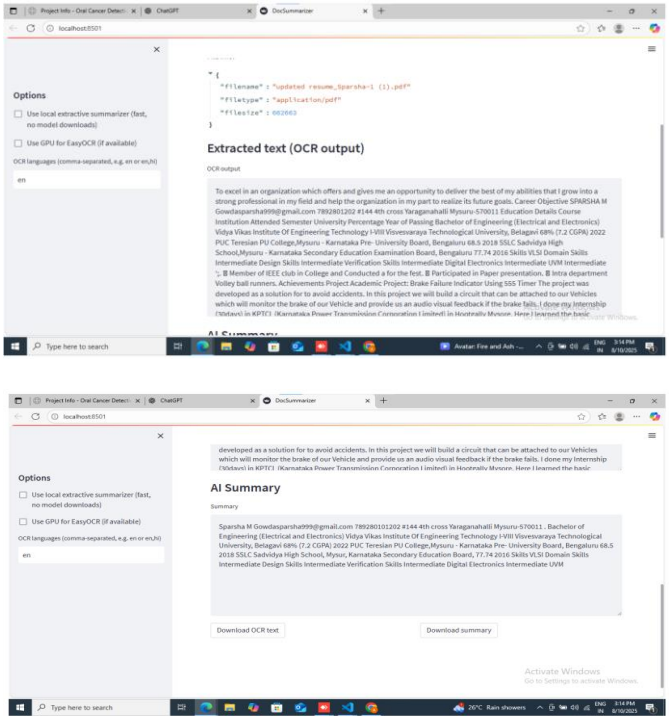


Figure 3: Processing Page

8. CONCLUSION & FUTURE SCOPE

The study on Automated Document Processing:
Combining OCR and Generative AI for Efficient Text Extraction and Summarization demonstrates the potential of integrating advanced technologies to address the challenges associated with managing large volumes of unstructured documents. By combining the strengths of Optical Character Recognition (OCR) for text digitization and Generative Artificial Intelligence (Gen AI) for summarization, the system provides an effective end-to-end solution that transforms raw document data into structured, concise, and actionable information. The project successfully illustrates how OCR can extract meaningful text even from scanned, handwritten, and complex documents, while preprocessing ensures that the extracted data is clean and structured for analysis. With the integration of transformer-based generative models, the system not only shortens lengthy documents but also preserves their semantic essence, delivering summaries that are both accurate and human-readable.

While the current system provides a reliable and efficient solution for document digitization, text extraction, and summarization, there remain opportunities for further improvement and expansion. Future enhancements can make the system more robust, intelligent, and adaptable across wider use cases. Implementing end-to-end encryption, secure access controls, and on-premise deployment options would increase trust and adoption among organizations handling confidential data. In summary, future enhancements such as multilingual support, multimodal capabilities, real time summarization, domain-specific fine-tuning, semantic knowledge integration, and stronger security will expand the scope of this system, making it even more powerful, versatile, and suitable for large-scale adoption across industries.

9. REFERENCES

[1] Smith, R. (2007). An overview of the Tesseract OCR engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 629–633. IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

[2] Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013). High-performance OCR for printed English and Fraktur using LSTM networks. 2013 12th International Conference on Document Analysis and Recognition, 683–687. https://doi.org/10.1109/ICDAR.2013.140

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[4] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67. https://doi.org/10.48550/arXiv.1910.10683

[5] Zhang, J., Tan, C., & Xiao, Y. (2020). Extractive and abstractive text summarization using transformer-based models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4066–4079. https://doi.org/10.18653/v1/2020.emnlpmain.332

[6] Google Cloud. (2023). Cloud Vision OCR documentation. Google Cloud. Retrieved from https://cloud.google.com/vision/docs/ocr

[7] Amazon Web Services. (2023). Amazon Textract – Extract text and data from documents. AWS Documentation. Retrieved from https://aws.amazon.com/textract/

[8] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint. https://doi.org/10.48550/arXiv.2303.08774

[9] Kaur, G., & Gupta, V. (2010). Effective approaches for extraction of keywords. International Journal of Computer Science Issues, 7(6), 144–148.

[10] Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed., draft). Stanford University. Retrieved from https://web.stanford.edu/~jurafsky/slp3/