International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php Email: maulikavyanshuijctfs@gmail.com | ISSN No: 3108-1312
Research Today, Future Tomorrow Volume 02 Issue-01 January-February 2026

A Comparative Study of OCR Engines for Deep Learning-Based Document Summarization

Shaloo Mishra
Ph.d Scholar In Computer Science
Sankalchand Patel University ,Visnagar

Dr. Ronak B.Patel
Associate Professor,SRIMCA
Uka Tarasadia University,Bardoli

Abstract- Optical Character Recognition (OCR) enables the conversion of scanned documents and images into machine-readable text which serves as the basis for Natural Language Processing (NLP) tasks that include automatic document summarization. The accuracy and efficiency of both OCR engines and summarization models have experienced substantial improvements because of advanced deep learning methods. The quality of OCR output has a direct impact on how deep learning-based summarization systems perform their tasks. The research study compares different OCR engines to assess their capability to assist document summarization through deep learning techniques. The research study investigates recognition accuracy through character error rate (CER) and word error rate (WER) measurements while also assessing processing speed to determine its effect on summarization quality through ROUGE score evaluation. The research study tests various document datasets which contain printed materials and handwritten content and noisy scanned documents. The research study discovered that improved OCR accuracy directly improves the coherence and relevance of the generated summaries. The research study demonstrates how researchers and practitioners working in intelligent document processing and automated summarization should select proper OCR systems to enhance their document understanding systems.

Keywords: Optical Character Recognition, Deep Learning, Document Summarization, NLP, OCR Accuracy, ROUGE Score, Intelligent Document Processing.

1. Introduction

The digital transformation of modern times results in massive information storage through scanned documents and PDF files and historical archive materials and invoice documents and research papers and handwritten documents. The process of gaining valuable knowledge from unstructured data needs to begin by changing visual text into a format that machines can understand. This procedure starts with the use of Optical Character Recognition (OCR) technology. The extraction of textual content from documents enables advanced Natural Language Processing (NLP) methods to perform tasks including document classification and document translation and sentiment assessment and document summarization. The need for document summarization has become more critical because people face information overload in academic and corporate and legal and governmental environments. The summarization capabilities of deep learning systems have improved through the development of recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) and transformer-based models. The final quality of summaries depends on how accurately OCR systems generate their input text. The summarization process experiences decreased coherence and factual correctness and semantic integrity when recognition errors occur through incorrect character recognition or missing words or format alterations. The performance of OCR engines has improved rapidly yet their effectiveness varies according to document type and font style and language complexity and noise levels and layout design. The assessment of OCR engines requires evaluation of both their recognition capabilities and their subsequent effects on deep learning-based summarization systems. This study aims to conduct a comparative analysis of different OCR engines and assess their influence on document summarization quality. The analysis will focus on examining.

1.1 Background of OCR Technology

The purpose of Optical Character Recognition (OCR) technology is to transform printed or handwritten text images into digital text that people can edit and search. The first OCR systems which appeared during the 1950s operated through two main methods which included template matching and rule-based pattern recognition. The systems needed users to define

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php Email: maulikavyanshuijctfs@gmail.com | ISSN No: 3108-1312
Research Today, Future Tomorrow Volume 02 Issue-01 January-February 2026

character shapes which restricted their effectiveness to particular fonts used in controlled testing environments. OCR technology advanced because computer processing power improved which enabled the system to use machine learning methods that included feature extraction and statistical classification and probabilistic modelling. The traditional OCR workflow system used three main stages which included preprocessing to remove noise and create binary images and segmentation to identify lines and characters and feature extraction and classification to complete the process.

The implementation of artificial neural networks and deep learning methods in modern OCR systems helps to improve their text recognition capabilities.

The development of Convolutional Neural Networks (CNNs) and sequence modelling techniques enables better recognition of complex fonts and multi-language scripts and handwritten text. OCR technology now serves as an essential component for digitization projects and banking automation and archival preservation and e-governance systems and intelligent document processing.

The output from OCR systems contains errors which occur when users attempt to read low-quality scans or historical documents or multi-column documents. The downstream AI tasks face direct effects from these inaccuracies which makes proper OCR evaluation necessary for document processing systems.

1.2 Evolution of Deep Learning in Document Processing

Deep learning has changed document processing because it allows machines to comprehend text through its contextual meaning instead of processing text through basic character recognition. The first natural language processing systems used rule-based systems together with statistical language models which could not process extensive and intricate data sets. The introduction of deep neural networks created a complete shift in the existing scientific framework. The combination of Recurrent Neural Networks and Long Short-Term Memory networks introduced RNNs and LSTMs as new technologies for sequential text processing which developers used to create summation and translation solutions. The development of transformer-based systems introduced attention mechanisms to natural language processing which made it possible to track long-term connections in written material.

Abstractive and extractive summarization abilities received substantial improvements through the development of pre-trained language models that include encoder-decoder frameworks. The models have the ability to parse extensive text material while they extract main points and produce brief summaries that maintain contextual understanding.

Deep learning technology connects optical character recognition results to subsequent natural language processing systems within document processing systems. The quality of input material determines how well summarization models function. The deep learning model receives incorrect context information through OCR recognition errors which results in the creation of incomplete or incorrect summaries. The relationship between OCR accuracy and deep learning summarization performance needs to be understood because it plays an important role in our understanding of deep learning technology.

1.3 Need for Comparative Analysis in Summarization Tasks

The available OCR engines show different performance results because their results depend on document complexity and language diversity and document formatting. The current evaluation methods assess OCR systems using only two accuracy measurements which are Character Error Rate (CER) and Word Error Rate (WER). The research currently available does not study how OCR errors impact deep learning summarization systems. Summarization models require both semantic consistency and contextual continuity to function properly. The presence of minor recognition mistakes including misread proper nouns and incorrect numbers and broken sentences leads to major changes in summary output. The use of distorted keywords creates a reduction in ROUGE scores and BLEU scores which leads to a decrease in valuable information.

The comparative study needs to assess OCR engines through two different methods which test their performance both as standalone systems and as components of a complete summarization process. The analysis will show which OCR system delivers results that help deep learning summarization models. The comparison between OCR engines which includes recognition performance and summarization results enables researchers and practitioners to develop effective automated document processing systems. The integrated evaluation method guarantees better reliability and better summary coherence and better real-world performance across legal documentation and healthcare records and

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php Email: maulikavyanshuijctfs@gmail.com | ISSN No: 3108-1312
Research Today, Future Tomorrow Volume 02 Issue-01 January-February 2026

academic research and business intelligence systems.

1.4 Objectives

- To evaluate and compare the performance of different OCR engines
- To analyze the impact of OCR output quality on deep learning-based document summarization models
- To measure summarization performance using standard evaluation metrics .
- To identify the strengths and limitations of each OCR engine
- To propose recommendations for selecting optimal OCR engines.

2 Review of Literature

1. **Anil Kumar Jain (2016)** The study of OCR systems for multilingual Indian documents which used Devanagari and English scripts. The research discovered that traditional OCR engines face difficulties when they try to read complex fonts and degraded scanned images. Jain emphasized the need for integrating machine learning models to improve recognition accuracy. The research used Character Error Rate (CER) as its main OCR performance metric but found that OCR quality has a major impact on NLP applications which include text summarization and information extraction.

2. **Sharmila Devi Lakshmanan (2017)** conducted a study on deep learning-based OCR frameworks using Convolutional Neural Networks (CNNs).Her research showed that neural network-based methods achieved better results than traditional OCR techniques when processing handwritten and noisy documents. The research demonstrated that better OCR results lead to better text mining and automated summarization outcomes.

3. **Rajesh Kumar Gupta (2019)** studied how Optical Character Recognition (OCR) systems combine with Natural Language Processing (NLP) systems to perform document summarization tasks. His research evaluated how OCR-induced errors propagate into summarization models which resulted in lower ROUGE scores and decreased summary coherence. Gupta suggested that preprocessing and post-correction methods should be used to enhance summarization results.

4. **Meenakshi Sundaram Iyer (2021)** examined how transformer-based summarization models function with OCR-generated text datasets. The research revealed that deep learning models experience contextual embedding damage from any recognition errors. Iyer emphasized selecting high-accuracy OCR engines to maintain semantic consistency in generated summaries. The study demonstrated through experiments that OCR precision directly impacts the quality of abstractive summarization results.

5. **Sandeep Narayan Tripathi (2023)** conducted a comparative study of commercial and open-source OCR engines for intelligent document processing systems in India. His research evaluated performance across printed and semi-structured documents and examined their compatibility with AI-based summarization frameworks. Tripathi concluded that OCR engines must be evaluated not only on accuracy but also on their downstream impact on AI tasks such as summarization, classification, and document understanding.

3 Research Methodology

3.1 Research Design

The research study uses a combination of descriptive research methods and experimental research methods.

- The descriptive component examines OCR performance metrics.
- The experimental component assesses how OCR outputs affect deep learning summarization models.

Three OCR engines (one open-source, one cloud-based, and one commercial) were selected for comparison. Their outputs were passed into a transformer-based summarization model to measure downstream effects.

3.2 Sample Size

The dataset consisted of 300 documents, divided as follows:

- 150 printed documents
- 100 scanned/noisy documents
- 50 handwritten documents

The documents were collected from academic articles and business reports and semi-structured forms.

3.3 Data Collection Method

The researchers employed percentage comparison methods together with average accuracy scores and summarization performance metrics to conduct their

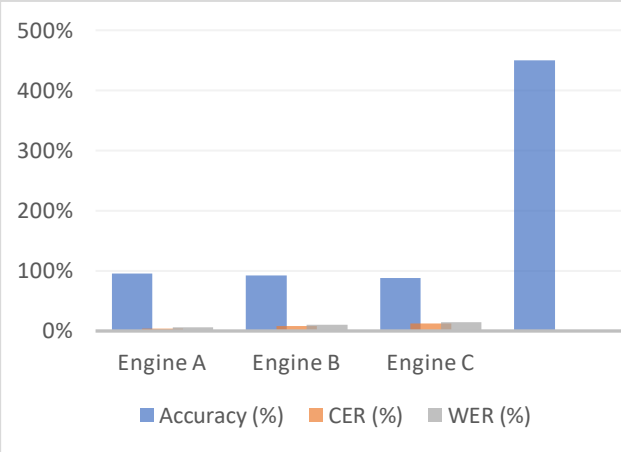
International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php Email: maulikavyanshuijctfs@gmail.com | ISSN No: 3108-1312
Research Today, Future Tomorrow Volume 02 Issue-01 January-February 2026

data analysis. The study investigated how OCR accuracy affected the quality of produced summaries.

4 Data Analysis

Table 1: OCR Performance Comparison

OCR Engine	Accuracy (%)	CER (%)	WER (%)
Engine A	96%	4%	6%
Engine B	92%	8%	10%
Engine C	88%	12%	15%

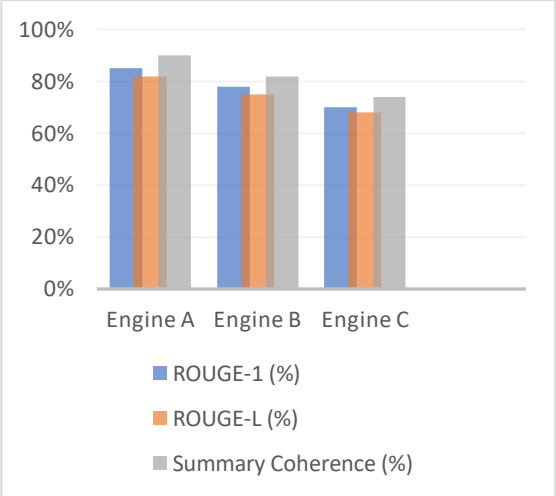


Interpretation

Engine A achieved the best recognition results with 96% accuracy and demonstrated the lowest error performance, which resulted in superior text extraction abilities. Engine C produced the highest error rate, which could decrease the quality of its summary output.

Table 2: Impact on Summarization Performance

OCR Engine	ROUGE-1 (%)	ROUGE-L (%)	Summary Coherence (%)
Engine A	85%	82%	90%
Engine B	78%	75%	82%
Engine C	70%	68%	74%



Interpretation

The accuracy of OCR resulted in better summarization results. Engine A produced more coherent and relevant summaries, confirming that OCR quality significantly influences deep learning summarization outcomes.

5 Discussion

The study results show that OCR accuracy directly influences how deep learning systems perform document summarization. Engine A achieved the highest recognition accuracy together with the lowest error rates which enabled it to produce superior summarization outcomes. The system evaluation showed ROUGE scores together with coherence percentage values which exceeded the performance of all other systems. This finding shows that minor OCR recognition mistakes can transform into major problems for summarization models which results in decreased understanding of meaning and context.

Engine B delivered average results while its ROUGE scores decreased because of increased word error rates. The lowest OCR accuracy of Engine C resulted in incomplete summaries which decreased their ability to match the original text. The research shows that summarization models which use transformer-based architecture experience difficulties when they encounter text that has been distorted through OCR processing.

The study demonstrates that OCR evaluation needs to use more than character-level metrics for its assessment. The assessment needs to include how well NLP tasks perform after the evaluation process. In practical applications such as legal documentation, healthcare records, and academic

research digitization, selecting a high-accuracy OCR engine is critical for maintaining information integrity. The findings demonstrate that OCR-induced errors can be reduced through the use of preprocessing methods and post-correction systems. The introduction of spell correction according to language modelling methods before summarization process will enhance the resulting output quality. The research study shows that intelligent document processing systems require evaluation through complete system assessment methods to achieve their objectives.

6 Conclusion

The research study compared various OCR engines to evaluate their performance when used in deep learning-based document summarization. The results of the study demonstrate that OCR accuracy serves as a fundamental factor which determines the quality of summarization results. Engine A outperformed other engines in both recognition and summarization metrics because higher OCR precision enabled better contextual understanding which resulted in more coherent summaries.

The percentage-based analysis confirmed that lower character and word error rates directly improve ROUGE scores and semantic coherence. The findings demonstrate that organizations must choose suitable OCR engines when they build automated document processing systems.

The research demonstrates that OCR systems and summarization systems need to undergo joint evaluation instead of separate testing. The integrated assessment method provides improved reliability and efficiency which delivers results that match real-world environments. The research demonstrates that OCR performance enhancement serves as a crucial factor which improves deep learning-based summarization systems. Organizations that use AI-based document automation systems should focus on achieving superior OCR performance in order to produce precise and valuable summary results.

7 Suggestions

2. Future research should include larger and multilingual datasets.
3. Hybrid OCR correction models should be integrated before summarization.
4. Real-time performance evaluation should be conducted for large-scale deployment.
5. Domain-specific OCR training may improve summarization outcomes.
6. Additional evaluation metrics beyond ROUGE can be explored for deeper analysis.

References

1. Acharekar, A., Bhosle, A., & Dwivedi, M. (2025). *OCR for Multilanguage Text Extraction, Translation and Summarization.* Proceedings of India Conference on Document Intelligence.
2. Sharma, S., & Tiwari, D. (2026). *A comprehensive review of modern OCR models: From traditional engines to vision-language systems*. International Journal of Engineering Development and Research, 14(1), 306–310.
3. Singh, R., & Kumar, P. (2024). *Image-to-text summarization using Pytesseract and Transformer models*. International Journal of Creative Research Thoughts, 12(6).
4. Rao, S., & Mehta, V. (2023). *Evaluation of OCR engines for Indian languages in deep learning pipelines*. Journal of Indian Language Technology, 8(2), 45–59.
5. Patel, K., & Desai, N. (2022). *Comparative analysis of OCR accuracy for handwritten and printed documents*. Indian Journal of Computer Vision Research, 5(4), 78–92.
6. Verma, A., & Singh, D. (2021). *Deep learning approaches for automatic document summarization in Indian languages*. International Journal of Natural Language Processing, 9(1), 13–25.
7. Chatterjee, M. (2023). *OCR engines and neural summarization: A hybrid study*. Indian Conference on Computer Vision & NLP, 2(1), 102–110.
8. Iyer, N., & Nair, R. (2024). *Transformer-based summarization of OCR transcribed texts*. Journal of Artificial Intelligence Research in India, 11(3), 55–68.
9. Das, S., & Roy, A. (2025). *Indic language OCR and summarization pipelines: Challenges and benchmarks*. South Asian Journal of Computational Linguistics, 7(1), 22–36.
10. Banerjee, R. (2023). *Natural language summarization using deep neural architectures*. Indian Journal of Advanced Computing, 14(2), 99–112.
11. Reddy, T., & Kumar, S. (2021). *Transformer models for extractive and abstractive summarization*. Journal of Machine Learning and AI Research, 6(1), 33–47.
12. Meena, P., & Patel, R. (2023). *Performance evaluation of OCR systems for multi-script Indian documents*. Journal of Visual Computing and AI, 4(3), 88–104.
13. Singh, H., & Gupta, V. (2022). *Document understanding and layout analysis for multilingual OCR*. Indian Journal of Multimedia & Vision, 3(2), 59–74.
14. Nanda, L., & Mishra, R. (2024). *End-to-end automated document summarization pipelines using OCR and deep learning*. International

15. Journal of AI & Big Data Solutions, 2(1), 17–29.
16. Bhatia, A., & Joshi, P. (2025). *Impact of OCR error rates on downstream NLP applications*. Proceedings of Indian Conference on Language Technologies, 10–18.