InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol 3(No.1), 45-57 (2026)
DOI: https://doi.org/10.51173/ijds.v3i1.56
https://ijds.mtu.edu.iq

A DEEP LEARNING FRAMEWORK FOR EXTRACTING AND SUMMARIZING TEXT FROM IMAGES

Osama Emad Abd Alhussein [1]*, Abbas EL Dor [1]

[1] *Department of Computer Science, Faculty of Sciences, Lebanese University, Beirut, Lebanon*

** Corresponding author E-mail: abbas.eldor@ul.edu.lb (Abbas EL DOR)*

RESEARCH ARTICLE

ARTICLE INFORMATION

SUBMISSION HISTORY:
Received: 3 September 2025
Revised: 21 October 2025
Accepted: 26 November 2025
Published: 30 January 2026

KEYWORDS:

Text Summarization;
Image Text Extraction;
Deep Learning;
Natural Language Processing (NLP);
BiLSTM;
Optical Character Recognition (OCR).

ABSTRACT

In the digital era, substantial amounts of textual information are embedded in images, especially across news outlets, social platforms, and scanned documents. This presents a significant technical challenge: efficiently extracting and summarizing text from images in an automated way that preserves context and meaning. Traditional text summarization techniques are not directly applicable to image-based content because they depend on pre-structured input text. In this paper, we propose a framework that integrates Optical Character Recognition (OCR) and advanced Natural Language Processing (NLP) models to address this challenge. The proposed method implements OCR to extract raw text from images, followed by deep learning-based summarization using models such as LSTM, Bi-LSTM, BERT and T5. These models are trained on large-scale news datasets to enhance their ability to generate coherent summaries from unstructured text. To ensure accessibility and practical usability, our framework is deployed via an interactive web-based interface that allows end-users to upload images and receive concise summaries in real time. Experimental evaluation demonstrates the efficacy of the proposed approach, particularly with transformer-based models, in delivering high-quality summarization from visual text sources.

1. INTRODUCTION

In today's digital domain, the propagation of image-rich content across social media and digital archives has led to an explosion of multimodal data [1], [2]. Most of this content consists of embedded text in formats such as scanned documents, infographics, advertisements, and screenshots. According to previous research, extracting and understanding this textual information is a challenging yet crucial task for many downstream applications, including content indexing, summarization, information retrieval, and digital archiving. Although this need, traditional Automatic Text Summarization (ATS) approaches, primarily designed for structured and plain-text documents, struggle to effectively handle unstructured text based on images[3], [4]. These conventional methods commonly use rule-based extraction or shallow machine learning models that assume clean, well-formatted input. However, text in images often varies in font type, size, layout, and orientation, which may be obscured by background noise, making text detection and semantic interpretation significantly harder [5], [6]. Furthermore, although OCR systems are advancing, they remain prone to errors when handling low-resolution, stylized, or multilingual text, further degrading the performance of downstream summarization models [7].

Recent studies address these challenges by integrating frameworks that combine Optical Character Recognition (OCR) technologies with advanced Natural Language Processing (NLP) models, such as transformer-based summarizers like BERT, T5, and PEGASUS [8], [9]. These systems aim to address the gap between visual text acquisition and linguistic abstraction by first converting image content into machine-readable text, then generating concise, context-aware summaries using deep learning-based NLP techniques.

DOI: https://doi.org/10.51173/ijds.v3i1.56

Most automated solutions have lacked emphasis on the unsustainable growth of multimedia content. On the other hand, manual summarizations no longer scalable, particularly in domains where large-scale document digitization is ongoing, such as healthcare, education, journalism, and enterprise knowledge management [6]. The redundancy and noise ingrained in digital text store further complicate data consumption and hinder decision-making processes [10]. Hence, automated summarization not only reduces cognitive overload but also enables rapid information access, making it a valuable tool in time-critical environments [11]. In the deep learning domain, some studies have attempted to improve ATS systems by enabling contextual understanding, abstractive summarization, and even automatic domain adaptation[12]. However, several technical difficulties remain unresolved and are not more efficient. Chief among them is the challenge of selecting the most informative subcomponents from noisy OCR outputs and preserving logical coherence in the generated summaries [13]. Furthermore, there is a need to develop interfaces that democratize access to these tools for non-technical users and facilitate their real-world applicability.

This study proposes a comprehensive framework for text summarization from an image that integrates OCR with deep NLP models. It aims to (i) evaluate the effectiveness of various summarization models, including LSTM, Bi-LSTM, BERT, and T5. When applied to text extracted from images, (ii) compare performance using established metrics such as ROUGE scores, and (iii) provide a user-friendly interface that enables end-users to upload images and receive summaries in real-time. The proposed solution aims to address all gaps in existing methods and to provide a scalable solution for multimodal text summarization.

2. LITERATURE REVIEW

Nowadays, a growing number of studies have focused on automatic text summarization by leveraging advances in deep learning and NLP to enhance the fluency, coherence, and relevance of generated summaries. These methods categorized into two models: recurrent models, attention-based networks, and transformer-based architectures. In 2019, Zhong et al. [14] examined data bias in neural extractive summarization models. They analyzed how dataset properties affect model generalization and established connections between dataset priors and model design. Their results showed that a deeper understanding of dataset characteristics can lead to significant improvements in summarization performance, even with relatively simple approaches. Another study in 2019, Liu and Lapata [15] implemented pre-trained transfer learning models such as BERT for text summarization tasks. They designed a general framework for both extractive and abstractive summarization that introduces a novel document-level encoder and a two-stage fine-tuning strategy. Their model achieved the best performance across three benchmark datasets, suggesting the effectiveness of pre-trained encoders for summarization. In a similar study on integrating extractive and abstractive summarization. Bae et al. [16] provided an approach for sentence rewriting. The authors implemented reinforcement learning to maximize ROUGE scores directly and incorporated BERT for stronger language understanding. Experimental results implemented on three different datasets, CNN/Daily Mail, New York Times, and DUC-2002 datasets, achieved an outperforming new state-of-the-art performance while generalizing effectively in deep learning-based abstractive summarization.

Suleiman and Awajan in [17] provided a detailed review, analyzing common datasets such as Gigaword and CNN/Daily Mail, and highlighted the importance of ROUGE metrics. In addition to the identified major challenges, there are out-of-vocabulary issues and factual inaccuracies. Their review found that RNNs with attention and LSTMs were the pre-trained models that achieved the best ROUGE scores. That extractive summarization was enhanced as a semantic text-matching problem rather than sentence extraction. In [18], Zhong et al. proposed a new framework matched between source documents and candidate summaries in a semantic space, leading to improved performance. In 2022, Gambhir and Gupta [19] proposed a deep learning model, WL-AttenSumm, for extractive summarization that uses a word-level attention mechanism. The authors integrated a Convolutional Bi-GRU architecture and attention to capture semantic and syntactic relationships. They evaluated system performance on the CNN/Daily Mail and DUC-2002 datasets, achieving

significant improvements in ROUGE recall and F1 scores compared to baseline methods. This study ensures the efficiency of word-level attention in extractive summarization.

In 2022, Mahalakshmi and Fatima [20] developed a deep learning framework for information retrieval, text summarization, and image captioning, which combined a BiLSTM for retrieval, template generation, and a Deep Belief Network (DBN) for summarization. On the Gigaword and DUC datasets, the model outperformed baseline approaches in precision, recall, and F-score, while BLEU evaluation confirmed the accuracy of the generated image captions. In 2022, Anand and Wagh [21] focused on automatic summarization of Indian legal texts using deep learning. The authors proposed two neural models: one based on word and sentence embeddings, which avoids the use of handcrafted features or domain-specific rules. This approach addressed the lack of labelled datasets by generating sentence classes from reference summaries, rather than relying on expert manual marking. The results indicated that these methods were effective and adaptable across legal and non-legal domains, leading to the use of this approach in environments with limited labelled data (Table 1).

Table 1. Summarization of related work

Ref	Year	Technology	Dataset(s)	Results / Contribution
[14]	2019	Neural extractive models; dataset bias analysis	Typical summarization datasets (e.g., CNN/Daily Mail)	Showed that dataset properties influence model design; simple models can improve with a better understanding of the dataset.
[15]	2019	BERT-based pretrained encoder; extractive & abstractive	CNN/Daily Mail, NYT, XSum	Achieved state-of-the-art results; two-stage fine-tuning improved summary quality.
[16]	2019	Sentence Rewriting + Reinforcement Learning + BERT	CNN/Daily Mail, NYT, DUC-2002	Directly maximized ROUGE scores; new SOTA performance; better generalization.
[17]	2020	Review of deep learning (RNNs, LSTMs, attention, pretrained encoders)	Gigaword, CNN/Daily Mail	Found pretrained encoder models achieve the highest ROUGE (43.85 / 20.34 / 39.9); highlighted challenges (OOV, repetition, fake facts).
[18]	2020	Semantic text matching framework	CNN/Daily Mail + 5 others	Achieved new SOTA extractive result (ROUGE-1: 44.41); showed effectiveness of the matching-based approach.
[19]	2022	WL-AttenSumm: Word-level Attention + Conv Bi-GRU	CNN/Daily Mail, Daily Mail, DUC-2002	Outperformed baselines; ROUGE recall: 55.9 (R1), 24.8 (R2), 53.9 (RL) on DUC-2002.
[20]	2022	BiLSTM (retrieval) + DBN (summarization) + image captioning	Gigaword, DUC	Outperformed compared methods; validated with ROUGE and BLEU for captions.
[21]	2022	Neural networks with embeddings (word/sentence)	Indian legal judgments	Effective summarization without handcrafted features; adaptable beyond the legal domain.
[23]	2023	LSTM, Seq2Seq (RNN-based abstractive summarization)	Kaggle News Summary	Produced appropriate abstractive summaries; results comparable to other methods.
[24]	2023	Transformer with Self-Attention (T2SAM)	In short news, DUC-2004	Reduced loss (from 10.30 to 1.82); F1-score: 48.50%; outperformed baselines.

In 2023, Karuna et al. [23] proposed an abstractive summarization approach using a hybrid deep learning model that includes LSTM and Seq2Seq architectures. The authors trained this model on the Kaggle News Summary dataset, resulting in robust intensification of large textual inputs while preserving meaning. In summary, the results showed that this method produced summaries comparable to those of contemporary approaches, demonstrating the potential of Seq2Seq-based summarizers. In another 2023 study, Kumar and Solanki [24]developed the Transformer with Self-Attention Mechanism (T2SAM) for abstractive summarization, which addressed coreference issues by enhancing contextual understanding via self-attention. The researchers trained the model on the DUC-2004 dataset and evaluated it based on ROUGE metrics. The proposed model achieved an F1-score of 48.50% and reduced training loss from 10.30 to 1.82 within 30 epochs, confirming that T2SAM outperformed existing baseline models. Despite these advancements, most existing models assume structured, clean input text, often neglecting noisy or unstructured sources, such as OCR-extracted image text. Moreover, the models often fail to maintain coherence when summarizing fragmented or low-quality inputs. While transformer-based models, such as BERT and T5, have shown strong generalization, they still struggle when applied to multimodal scenarios without proper preprocessing or pipeline integration.

According to Table 1, almost all previous models assume the availability of clean, machine-readable text and focus on preprocessed datasets, such as CNN/Daily Mail, DUC, or Gigaword, overlooking scenarios where text is embedded in images (e.g., news infographics, scanned documents, social media posts). While some studies combined summarization with information retrieval or image captioning, they did not address the direct problem of extracting and summarizing text from images in a unified pipeline. In addition, most prior work is evaluated as research prototypes, reporting ROUGE scores on benchmark datasets. Few studies provide an interactive, real-time system that end users can use to summarize visual text sources (e.g., by uploading an image to get a summary). Collectively, the reviewed studies illustrate the evolution of text summarization from basic RNN models to powerful transformer-based architectures. The incorporation of attention, pre-trained embeddings, and reinforcement learning has notably improved summary quality across domains. However, few works address the specific challenges of summarizing text extracted from images, a domain characterized by OCR noise, layout variation, and semantic fragmentation. Moreover, much of the oldest research often assumes that text exists in a structured corpus, disregarding real-world applications in which text must first be extracted from image data. The proposed study addresses previous gaps by designing an end-to-end framework that integrates OCR for image-text extraction with deep NLP models for summarization, explicitly targeting noisy, unstructured inputs. In addition, most prior work focuses on dataset-specific tuning, whereas our framework ensures usability by deploying a real-time user interface that leverages automated summarization systems.

3. METHODOLOGY

The proposed model for summarization text based on image includes six integrated stages, which operate in a pipeline structure: (1) load dataset and preprocessing, (2) train and test NLP models, (3) evaluate models and select best model, (4) user interface deployment, (5) text extraction using Optical Character Recognition (OCR), (6) text summarization using deep NLP models. These stages collectively form an end-to-end system that automatically extracts and summarizes textual content embedded in images. The general system architecture is shown in Fig. 1. The proposed framework's steps include dataset loading, model building, model training, and inference, which were implemented in Python and run on Google Colab with seamless integration via the Kaggle API. These steps ensure reproducibility and scalability for researchers and developers working in similar domains.

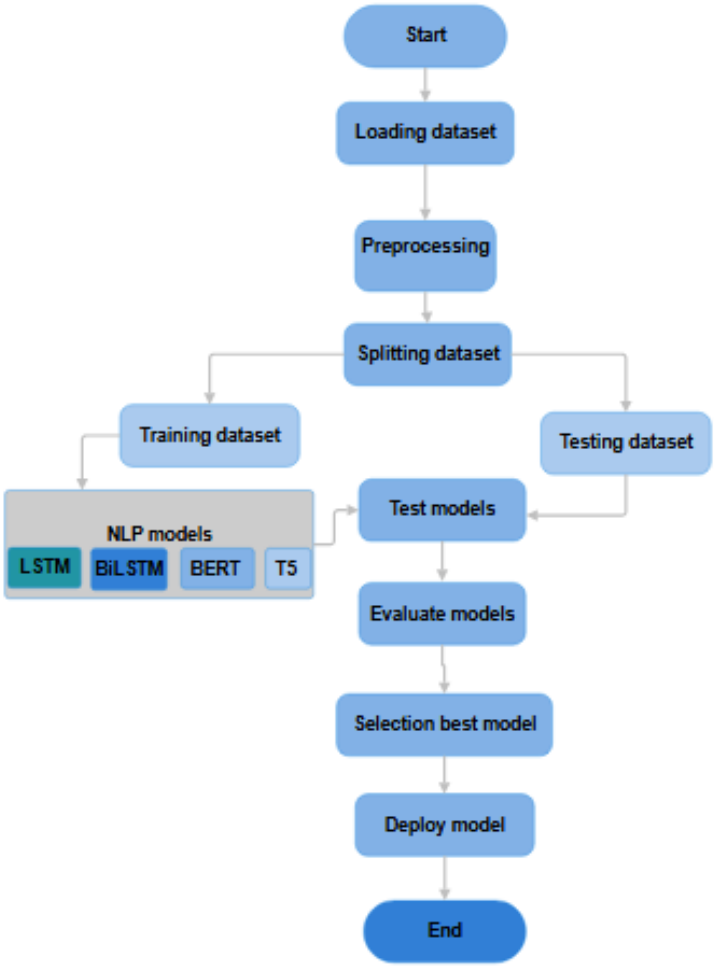

Figure 1. Flowchart for proposed solution.

3.1 Dataset

In this step, load the CNN/DailyMail News Text Summarization dataset [25]. It is a widely adopted benchmark in NLP research that includes over 300,000 English news articles paired with professionally written summaries, where each row pairs a news article with its human-written highlights, as shown in Fig. 2. Although the dataset was created for machine reading comprehension and question-answering tasks, it has since been adapted for both extractive and abstractive summarization tasks. We simulated real-world scenarios by converting text documents into image formats, including screenshots, scanned documents, and stylized renderings. These synthetic image samples were then processed through our OCR module based on Tesseract, mimicking practical use cases such as summarizing text from infographics, scanned reports, and social media screenshots. The proposed model allows us to benchmark model performance in both clean text and OCR-derived noisy text settings.

	id	article	highlights
0	0001d1afc246a7964130f43ae940af6bc6c57f01	By . Associated Press . PUBLISHED: . 14:11 EST...	Bishop John Folda, of North Dakota, is taking ...
1	0002095e55fcbd3a2f366d9bf92a95433dc305ef	(CNN) -- Ralph Mata was an internal affairs li...	Criminal complaint: Cop used his role to help ...
2	00027e965c8264c35cc1bc55556db388da82b07f	A drunk driver who killed a young woman in a h...	Craig Eccleston-Todd, 27, had drunk at least t...
3	0002c17436637c4fe1837c935c04de47adb18e9a	(CNN) -- With a breezy sweep of his pen Presid...	Nina dos Santos says Europe must be ready to a...
4	0003ad6ef0c37534f80b55b4235108024b407f0b	Fleetwood are the only team still to have a 10...	Fleetwood top of League One after 2-0 win at S...

Figure 2. Sample Content from the Dataset.

3.2 Preparing Data

3.2.1 Text Preprocessing

There are some preprocessing steps to make the dataset clean and valid, as follows [26], [27]:

- Convert letters in articles to lowercase and strip them of HTML tags, punctuation, and redundant symbols using regular expressions and Beautiful Soup.
- Expand contractions and elongated words normalized.
- Remove stop words from articles to reduce noise, such as (the, a), while summaries were minimally cleaned to preserve meaning.
- Apply lemmatization to reduce vocabulary sparsity.
- Split it into two sets 80% for the training set and 20% for the testing set.
- Implement tokenization using the Keras tokenizer, generating vocabulary indices.
- Use pre-trained GloVe vectors and BERT contextual embeddings to capture semantics.
- Truncate input sequences to a maximum length of 512 tokens.

After implementing the previous preprocessing steps, the dataset is high-quality and clean. To ensure both robust model learning and fair evaluation, the dataset is split into two sets in an 80:20 ratio. Using the Keras tokenizer to construct a vocabulary index and transform the sequences into numerical representations. In addition, to enhance generalization, lemmatization is integrated to reduce words to their base forms, thereby lowering vocabulary sparsity. In parallel, word embeddings and contextual encodings were leveraged to capture deeper semantic dependencies within the data. In scenarios of class imbalance, sampling strategies were applied to maintain proportional representation across subsets. Finally, the preprocessed data was encoded as numerical values using word embeddings and contextual encoders, such as BERT, enabling compatibility with deep learning architectures, including LSTM, Bi-LSTM, BERT, and T5.

3.2.2 Image preprocessing for OCR

For enhancing the quality of the image and increasing density, we implemented some preprocessing steps as follows:

- Convert images to grayscale.
- Apply CLAHE (Contrast Limited Adaptive Histogram Equalization) to improve local contrast.
- Remove noise (median filtering) and contour-based ROI extraction to isolate text regions.
- Processed images were fed to Tesseract OCR, yielding machine-readable text that was then fed into the NLP summarization models.

3.3 Models

Long Short-Term Memory (LSTM): A kind of recurrent neural network (RNN) prepared to capture and learn long-term dependencies in sequential data [28]. It uses a group of memory cells and gating mechanisms, such as input, forget, and output gates, to control the flow of information and reduce the vanishing gradient problem. Bidirectional Long Short-Term Memory (Bi-LSTM): An extension of LSTM, which processes data in both forward and backward directions. This process enables the model to capture context from both past and future tokens, leading to a better understanding of sequences than a standard LSTM. BERT is A transformer-based language model pre-trained on large text corpora, which learns deep through bidirectional contextual representations by considering both left and right context in all layers. The BERT model is widely used for tasks such as classification and text summarization. Text-to-Text Transfer Transformer (T5): A model based on a transformer that treats every NLP task as a text-to-text problem. For

example, classification tasks are framed as generating labels in text form, and summarization tasks as generating shorter versions of the input text. This model is highly flexible and powerful in generative language tasks [29].

3.4 Training setup

- Optimizer: Adam with learning rate = 3e-5 (for transformers) and 1e-3 (for LSTM models).
- Batch size: 32.
- Epochs: 10–15 with early stopping.
- Loss function: categorical cross-entropy for sequence prediction.
- Evaluation: ROUGE-1 precision, recall, and F1-score.

3.5 Evaluation Models

To evaluate the quality of the model in text summarization, we used the ROUGE-1 metric, which measures unigram overlap between the generated summary and the human reference summary [30].

- Precision reflects the proportion of words in the generated summary that are also present in the reference summary, which indicates the accuracy of the model's output.
- Recall captures the proportion of important words from the reference summary that appear in the generated summary, showing the model's ability to retain essential information.
- F1-Score (Scale) provides a balanced measure by combining both precision and recall by ensuring that the evaluation considers both correctness and completeness of the generated summaries.

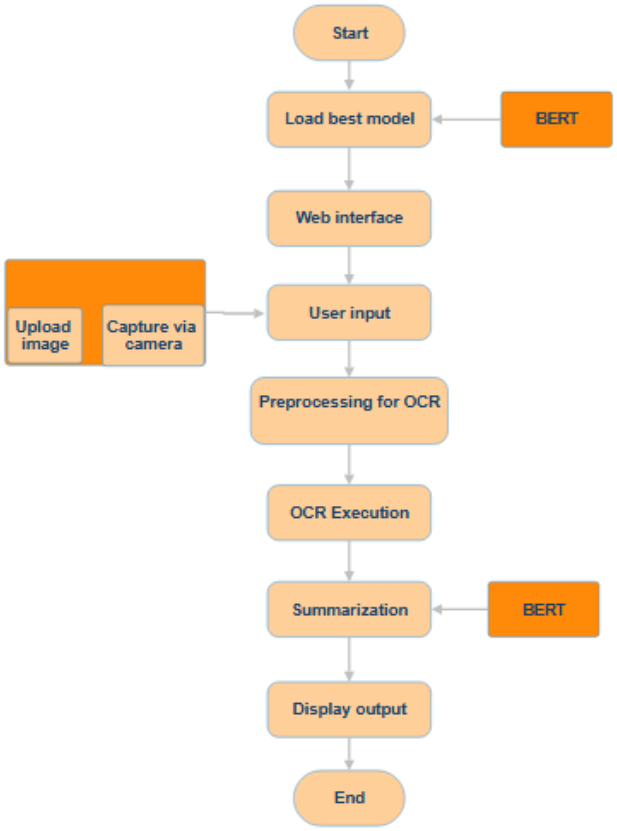

Figure 3. Workflow of the Deployment Process.

After building the model, train and test it, then evaluate it using previous metrics. Finally, select the best model for user interface deployment, which allows the user to interact with this system and summarise text as an image, as shown in Figure .

3.6 Deploy System

Figure shows the steps for deploying our system, which allows users to extract text from an image while preserving the original content with stable quality. The summary of the previous steps is defined below:

1. Load Best Model: Ensures high accuracy by using the top-performing model from the evaluation step. In our case, the best model is BERT.
2. Web Interface: using the Gradio platform to allow users to interact with the system without programming knowledge.
3. User Input: Users can upload existing images or capture new ones.
4. Preprocessing for OCR: Enhance OCR accuracy by enhancing the image and isolating text regions using contours.
5. OCR Execution: Converts visual text into machine-readable text for bridging image content and NLP processing.
6. Uses the best model BERT to analyze text based on an image and generate concise and coherent summaries.
7. Based on the analysis, the image provides real-time feedback to the user.

4. RESULTS

This section summarizes the findings obtained from implementing the proposed framework pipelines. These findings, which include results from each stage, ranging from dataset preparation and model training to final system deployment, are discussed. Discuss the performance metrics obtained during model training, and the usability and functionality of the final system were deployed through an interactive interface built using the Gradio platform. This interface enables users to interact with the proposed system by uploading images; our system then extracts textual content and generates concise summaries in a real-time user interface, showcasing the practical application and effectiveness of the proposed solution.

4.1 ROUG-1 Result

The comparison between models based on different metrics, such as f1 score, precision and recall, is shown in Table 2. The result of this comparison can help to select the most effective model. In terms of precision, the LSTM achieved the highest at 51%, resulting in concise summaries with fewer irrelevant words. But its recall was only 24%, suggesting it fails to capture enough relevant content from the original text. This means that this model selects a small number of sentences but is accurate at the expense of content comprehensiveness. The Bi-LSTM has slightly improved precision to 56%, making it the most precise model. However, recall dropped further to 22%, which negatively affected its overall F1-score 32%, indicating that although Bi-LSTM is highly selective. In summary, this model produces extremely short summaries that fail to include the essential information from the original text adequately.

BERT performs best among others, achieving the highest recall (45%) and F1-score (51%). On the other hand, its balanced performance between precision 57% and recall shows that it effectively captures important details while maintaining reasonable accuracy in selecting relevant content. This balance explains that this model not only selects accurate sentences but also focuses on the text's main content, which led to its being considered the best-performing model in this study. Finally, the T5 showed relatively balanced results, with a precision of 48%, a recall of 43%, and a solid F1-score of 46%. This reflects stable performance in summarization text. The T5 does not outperform the BERT model, but it is the best second option compared to Bi-LSTM and LSTM.

Table 2. Performance comparison of models.

Algorithm	F1 score (%)	Precision (%)	Recall (%)
LSTM	33	51	24
Bi LSTM	32	56	22
Bert	51	57	45
T5	46	48	43

According to the comparison of models based on scores, as shown in Figure 3. BERT achieved the highest precision by demonstrating its strong ability to filter out extraneous words and produce concise, meaningful summaries. This strength stems from its transformer-based contextual embeddings, which capture deep word dependencies and enable the model to retain only salient information, thereby minimizing redundancy. The Bi-LSTM achieved 56% precision, closely matching the baseline, thanks to its bidirectional structure, which captures both past and future context and yields overly concise outputs. The LSTM with 51% precision. It has moderate performance: although it can capture relevant words reasonably well, its unidirectional nature limits contextual understanding, often leading to noisier summaries than those from Bi-LSTM and BERT.

Finally, the T5 recorded the lowest score: It tends to produce longer summaries containing additional, sometimes irrelevant, content. In conclusion, these results showed that BERT is the most effective model for generating precise and relevant summaries. At the same time, Bi-LSTM produces highly focused but incomplete outputs, LSTM offers a reasonable baseline, and T5 emphasizes broader coverage at the expense of precision.

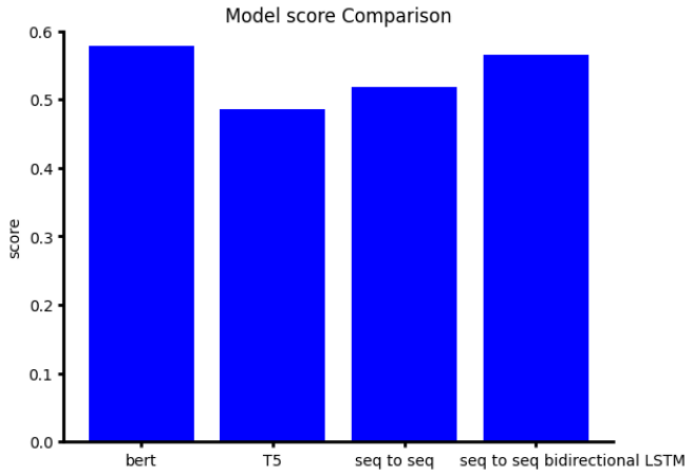

Figure 3. Comparative Analysis of Model Scores.

4.2 Test model via image

According to the test models shown in Fig. 5 and Fig. 6, the deployed BERT-based summarization system appeared promising when applied to text extracted from images. The OCR module effectively handled scanned documents and flyers, performing preprocessing steps such as grayscale conversion and ROI detection to improve text recognition clarity. While social media images with stylized fonts introduced occasional noise, the summarization model still generated coherent, concise summaries. In summary, BERT produces meaningful summaries even when OCR output is slightly noisy. But the final summary quality was strongly dependent on OCR accuracy, suggesting future work should focus on enhancing text extraction under challenging image conditions.

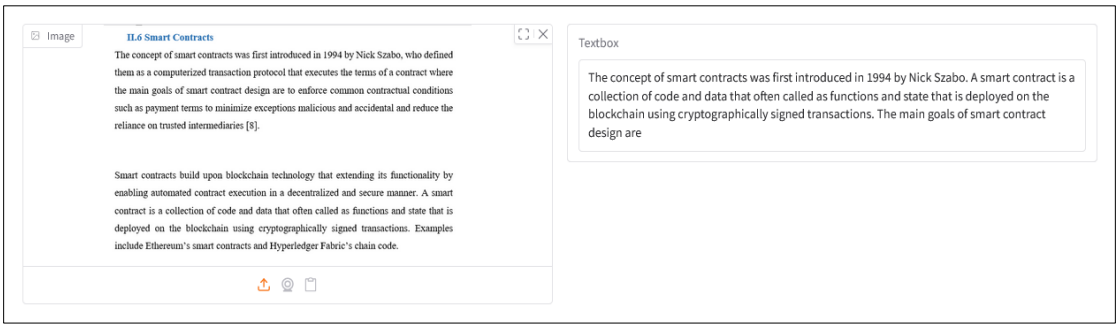

Figure 4: Example of Summarized Output (Short Text Case).

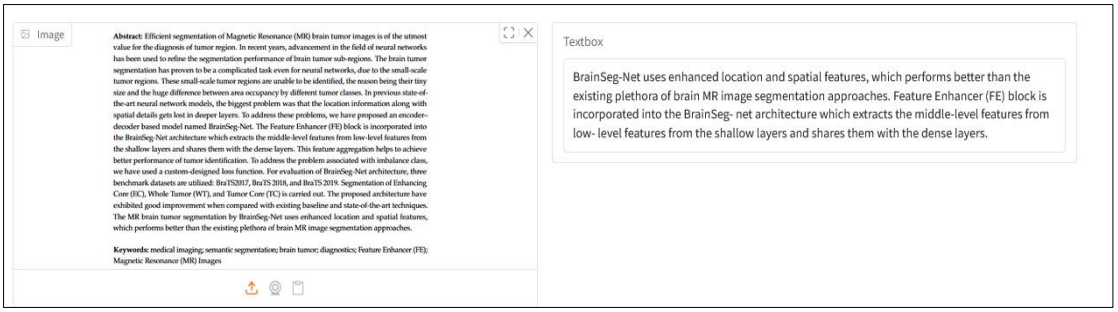

Figure 5: Example of Summarized Output (Long Text Case).

5. DISCUSSION

Based on previous experimental results, which provide a comprehensive view of the performance of multiple text summarization models that include LSTM, BiLSTM, T5 and BERT. BERT achieved the highest overall performance, with an F1-score of 51%, precision of 57%, and recall of 45%, clearly outperforming other models. While LSTM and BiLSTM appeared to have relatively higher precision (51% and 56%, respectively), their recall values were significantly lower, resulting in weaker overall effectiveness for summarization tasks. According to previous analysis, although these models could correctly identify some relevant information, they failed to capture the broader context needed for producing coherent summaries. Another model is the T5, which achieved a balanced F1-score of 46%, higher than LSTM and BiLSTM but still lower than BERT. In summary, it has moderate precision (48%) and recall (43%), which appear more reliable than recurrent models but less effective than BERT at preserving key contextual details. However, BERT's strength lies in its contextual embeddings, which enable it to capture semantic nuances across longer text sequences, leading to superior summarization outcomes compared to both recurrent and other transformer-based models.

In the deployment domain, integrating the best-performing BERT model with a Gradio-based system enables end users to upload or capture images for text summarization based on those images. The extraction of text from images using an OCR pipeline implemented OpenCV preprocessing followed by PyTesseract successfully bridged the visual and textual modalities by extracting text from diverse image sources, such as scanned documents and synthetic text-rendered images. While OCR introduced some noise, such as fragmented words, the preprocessing step mitigated these issues and ensured compatibility with the summarization model. The evaluation of this system showed that image-based summarization is both feasible and practical across different domains, such as academic flyers and PDFs, enabling concise summaries by making information retrieval more efficient. On the other hand, text-heavy social media posts were successfully condensed by reducing cognitive load for end users. The combination of BERT's high recall and robust OCR preprocessing ensured that the most relevant information was preserved, even at the cost of some precision. Overall, this balance shows the system's ability to serve practical needs in real-world scenarios where accuracy and efficiency are equally critical. In conclusion, the findings highlight three key implications:

- The transformer-like BERT and T5 models consistently outperformed RNN-based models, including LSTM and Bi-LSTM.
- Based on the recall metric, BERT and T5 offer a better balance, suggesting that the choice of model should depend on whether completeness or conciseness is prioritised.
- In deployment, the integration of OCR, preprocessing, and a Gradio interface demonstrates a complete pipeline that bridges raw image input to meaningful textual summaries, providing a strong foundation for real-world applications.

6. CONCLUSION

The proposed framework demonstrated the comparative performance of multiple summarization models using rich metrics, particularly ROUG-1. The results show that BERT was the most effective choice due to its superior F1 Score and balanced precision-recall trade-off. This BERT performance makes it the best option for successful integration into a Gradio-based application, further validating the system's ability to handle real-world data by extracting and summarizing text from images via OCR preprocessing. While recurrent models, LSTM and BiLSTM, demonstrated strengths in precision, their limited recall hindered their usability. Transformer models, practically BERT, are more robust in capturing contextual meaning and generating coherent summaries.

In summary, the project emphasized that combining advanced deep learning models with OCR and user-friendly deployment platforms can significantly enhance access to summarized information. The proposed framework has practical implications for fields such as education and document management, where rapid and accurate summarization can save time and reduce cognitive effort. Future directions could focus on improving OCR accuracy through fine-tuning summarization models for domain-specific applications and on expanding deployment features, such as multilingual support and adaptive summary length control.

CONFLICT OF INTEREST

The authors declare that there is *no conflict* of interest regarding the publication of this paper.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun ACM, vol. 60, no. 6, pp. 84–90, Jun. 2017, doi: 10.1145/3065386;CSUBTYPE:STRING:MAGAZINE;PAGE:STRING:ARTICLE/CHAPTER.

[2] L. Deng and D. Yu, "Deep Learning: Methods and Applications," Foundations and Trends in Signal Processing, vol. 7, no. 3–4, pp. 197–387, Jun. 2014, doi: 10.1561/2000000039.

[3] R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, May 2017, Accessed: Jan. 07, 2026. [Online]. Available: https://arxiv.org/pdf/1705.04304

[4] D. Karatzas et al., "ICDAR 2013 robust reading competition," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, pp. 1484–1493, 2013, doi: 10.1109/ICDAR.2013.221.

[5] A. Yadav, S. Singh, M. Siddique, N. Mehta, and A. Kotangale, "OCR using CRNN: A Deep Learning Approach for Text Recognition," 2023 4th International Conference for Emerging Technology, INCET 2023, 2023, doi: 10.1109/INCET57972.2023.10170436.

[6] W. Wu et al., "ICDAR 2023 Competition on Video Text Reading for Dense and Small Text," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14188 LNCS, pp. 405–419, 2023, doi: 10.1007/978-3-031-41679-8_23.

[7] M. Hasan, E. Rundensteiner, and E. Agu, "DeepEmotex: Classifying Emotion in Text Messages using Deep Transfer Learning," Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, pp. 5143–5152, 2021, doi: 10.1109/BIGDATA52589.2021.9671803.

[8] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, Accessed: Jan. 07, 2026. [Online]. Available: https://arxiv.org/pdf/1907.11692

[9] P. M. Lavanya and E. Sasikala, "Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey," 2021 3rd International Conference on Signal Processing and Communication, ICPSC 2021, pp. 603–609, May 2021, doi: 10.1109/ICSPC51351.2021.9451752.

[10] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," SN Computer Science 2022 4:1, vol. 4, no. 1, pp. 33-, Oct. 2022, doi: 10.1007/S42979-022-01446-W.

[11] S. Li and J. Xu, "HierMDS: a hierarchical multi-document summarization model with global–local document dependencies," Neural Computing and Applications 2023 35:25, vol. 35, no. 25, pp. 18553–18570, Jun. 2023, doi: 10.1007/S00521-023-08680-0.

[12] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 1797–1807, Aug. 2018, doi: 10.18653/v1/d18-1206.

[13] F. Ladhak, E. Durmus, H. He, C. Cardie, and K. McKeown, "Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization," Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1410–1421, 2022, doi: 10.18653/V1/2022.ACL-LONG.100.

[14] M. Zhong, D. Wang, P. Liu, Q. Xipeng, and H. Xuan-Jing, "A Closer Look at Data Bias in Neural Extractive Summarization Models," pp. 80–89, Nov. 2019, doi: 10.18653/V1/D19-5410.

[15] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 3730–3740, Aug. 2019, doi: 10.18653/v1/D19-1387.

[16] S. Bae, T. Kim, J. Kim, and S. Lee, "Summary Level Training of Sentence Rewriting for Abstractive Summarization," pp. 10–20, Sep. 2019, doi: 10.18653/v1/d19-5402.

[17] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," Math Probl Eng, vol. 2020, no. 1, p. 9365340, Jan. 2020, doi: 10.1155/2020/9365340.

[18] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive Summarization as Text Matching," Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 6197–6208, 2020, doi: 10.18653/V1/2020.ACL-MAIN.552.

[19] M. Gambhir and V. Gupta, "Deep learning-based extractive text summarization with word-level attention mechanism," Multimedia Tools and Applications 2022 81:15, vol. 81, no. 15, pp. 20829–20852, Mar. 2022, doi: 10.1007/S11042-022-12729-Y.

[20] P. Mahalakshmi and N. S. Fatima, "Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques," IEEE Access, vol. 10, pp. 18289–18297, 2022, doi: 10.1109/ACCESS.2022.3150414.

[21] D. Anand and R. Wagh, "Effective deep learning approaches for summarization of legal texts," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 2141–2150, May 2022, doi: 10.1016/J.JKSUCI.2019.11.015.

[22] G. Karuna, M. Akshith, P. S. Dinesh, B. V. Vardhan, Y. S. Bisht, and M. N. Narsaiah, "Automated Abstractive Text Summarization using Deep Learning," E3S Web of Conferences, vol. 430, p. 01021, Oct. 2023, doi: 10.1051/E3SCONF/202343001021.

[23] S. Kumar and A. Solanki, "An abstractive text summarization technique using transformer model with self-attention mechanism," Neural Computing and Applications 2023 35:25, vol. 35, no. 25, pp. 18603–18622, Jun. 2023, doi: 10.1007/S00521-023-08687-7.

[24] K. Moritz et al., "Teaching machines to read and comprehend," proceedings.neurips.ccKM Hermann, T Kocisky, E Grefenstette, L Espeholt, W Kay, M Suleyman, P BlunsomAdvances in neural information processing systems, 2015•proceedings.neurips.cc, Accessed: Jan. 08,

2026. [Online]. Available: https://proceedings.neurips.cc/paper/5945-teaching-machines-to-read-and-comprehend

[25] O. M. Al-Janabi, O. M. Alyasiri, E. A. Jebur, and S. M. Nafl, "Evaluating AI Language Models in News Retrieval: A Comparative Study Of ChatGPT-Plus and DeepSeek (R1)," InfoTech Spectrum: Iraqi Journal of Data Science , vol. 2, no. 2, pp. 13–19, Jun. 2025, doi: 10.51173/IJDS.V2I2.33.

[26] H. Fadhil Khalil, M. Fadhil Ibrahim, and H. Ataallah Hussein, "Evaluating The Impact of Feature Extraction Techniques on Arabic Reviews Classification," InfoTech Spectrum: Iraqi Journal of Data Science , vol. 1, no. 1, pp. 42–54, Jun. 2024, doi: 10.51173/IJDS.V1I1.10.

[27] F. M. Salem, "Gated RNN: The Long Short-Term Memory (LSTM) RNN," Recurrent Neural Networks, pp. 71–82, 2022, doi: 10.1007/978-3-030-89929-5_4.

[28] E. Lloret, L. Plaza, and A. Aker, "The challenging task of summary evaluation: an overview," Language Resources and Evaluation 2017 52:1, vol. 52, no. 1, pp. 101–148, Sep. 2017, doi: 10.1007/S10579-017-9399-2.

[29] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," SSRN Electronic Journal, May 2022, doi: 10.2139/SSRN.4120317.