Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comprehensive Survey of Deep Learning-Based OCR and Information Extraction for Low-Resource Languages

Md. Ammar Hossain Ammar
United International University
Mehedi Hasan Talha Talha
United International University
Ananna Saha Ananna
United International University
Md. Al Emran Emran
United International University
Md. Mohin Khan Mohin
United International University
Tuba Hai Tuba
United International University
Jannatun Noor Jannat
jannatun@cse.uiu.ac.bd

United International University

Research Article

Keywords: Optical Character Recognition (OCR), Document Layout Analysis (DLA), Low-Resource Languages, Document Intelligence, Automated Extraction.

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11295807/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Comprehensive Survey of Deep Learning-Based OCR and Information Extraction for Low-Resource Languages

Md. Ammar Hossain[1], Mehedi Hasan Talha[2], Ananna Saha[3], Md. Al Emran[4], Md. Mohin Khan[5], Tuba Hai[6], Jannatun Noor[7]

[*]Department of Computer Science and Engineering, United International University, Dhaka, 1212, Bangladesh.

*Corresponding author(s). E-mail(s): jannatun@cse.uiu.ac.bd;
Contributing authors: mhossain221601@bscse.uiu.ac.bd; mtalha221312@bscse.uiu.ac.bd; asaha221117@bscse.uiu.ac.bd; memran221088@bscse.uiu.ac.bd; mkhan221091@bscse.uiu.ac.bd; thai213109@bscse.uiu.ac.bd;

Abstract

This survey reviews the progress made in Optical Character Recognition (OCR) and document understanding through a review of more than 100 peer-reviewed papers. The importance of OCR is that it helps in the digitization of different types of documents, including governmental documents, medical records, financial documents, and legal documents, among others. Modern OCR uses advanced machine learning models including Convolutional Neural Networks (CNNs) and transformers. These models provide OCR with greater abilities in text recognition, Document Layout Analysis (DLA), Information Extraction (IE), and Key Information Extraction (KIE). However, the research on DLA, IE, KIE, and OCR has been conducted independently because of different methods used, different datasets, evaluation measures, and application domains. Another challenge is that little research has been done on OCR and document understanding for low-resource languages like Bangla because of problems like scarcity of annotated data and the unique nature of the script, among others. The survey classifies the available OCR technologies and compares traditional OCR with advanced OCR and gives an overview of the commonly used datasets, evaluation measures, and implementation challenges. DLA has increasingly become important in recent

times; therefore, future research directions in OCR and document understanding will include multi-modal learning.

Keywords: Optical Character Recognition (OCR), Document Layout Analysis (DLA), Low-Resource Languages, Document Intelligence, Automated Extraction.

1 Introduction

OCR and DLA are the two key components in document image understanding that allow us to transform scanned and born-digital documents into machine readable and semantically structured forms. In combination, they form the basis for intelligent document processing systems applicable to domains such as digital libraries, e-governance, financial document processing, health care, and legal document management. Where the former OCR is concerned with extracting textual information contained in the document, the DLA analyzes the document structure, including text regions, tables, figures, and reading order. This way, DLA provides important contextual knowledge about the document required for its correct understanding (Patel and Bhatt 2020).

Early OCR systems were based on handcrafted features, heuristic preprocessing, and template matching, while conventional techniques of DLA made use of geometry-based rules and manually crafted layout assumptions (Singh 2025). While providing reasonable accuracy under certain conditions, they were characterized by low robustness to various factors such as degraded documents, fonts, multilingualism, and complex layouts, limiting their use in different collections of real-world documents (Yang et al. 2022), (Huang et al. 2026). Moreover, incorrect layout analysis often causes errors in subsequent stages of OCR, thus contributing to further mistakes and decreasing the performance of the whole system (Hasan et al. 2024).

The area of OCR and DLA has seen numerous achievements of deep learning methods in recent years. Using convolutional neural networks, transformers, vision-language models, and other techniques that allow extracting sophisticated features from images and layouts, deep learning approaches have significantly improved results of document text recognition, layout analysis, and even reasoning about the context of documents (Ding et al. 2026), (Yu et al. 2021). Yet, there are still some problems with the existing research efforts in the area of OCR and DLA; they are fragmented in terms of datasets used, languages they cover, evaluation setups, and application fields. Such problems as lack of coverage for low-resource languages, lack of annotated datasets, inappropriate benchmarking, and insufficient generalization prevent researchers from creating efficient and universal solutions for document understanding (Huang et al. 2026), (Prakash and Hanumanthaiah 2022). It is obvious that the existence of these issues proves the necessity of conducting a survey on this topic, which will help to analyze existing research on this issue, identify its methodological advances, and provide future research directions for OCR and DLA.

1.1 Motivation

Modern OCR and DLA architectures often encounter significant reductions in their performance once applied in real-world environments that include noisy scans, multilingual documents, historical archives, and complex layouts (Henge and Rama 2016). This problem is especially evident in the case of low-resource languages and non-Latin scripts due to a lack of annotated data and language-specific resources for building and testing the corresponding models (Ali et al. 2023). Another major issue concerns inconsistencies related to the datasets and evaluation methods (Ding et al. 2026). Quite often, researchers use proprietary datasets or narrowly defined experimental setups that do not allow reproduction of results and make unbiased comparisons between different approaches difficult (Favish 2019), (Nida), (Rahman et al. 2023).

Even publicly available benchmarks may differ in terms of their annotations, data distribution, and evaluation metrics, leading to inconsistent performance estimates (Yang et al. 2022). At the same time, although recent models based on transformers and vision-language approaches managed to significantly improve their recognition accuracy, less attention was paid to their computational efficiency, domain adaptation, interpretability, and error propagation in document understanding pipelines (Sharma et al. 2025), (Naquib 2025). These limitations highlight the necessity for a well-structured and scientifically based survey that will evaluate existing methodologies and identify gaps in scientific research.

1.2 Research Scope and Focus

The survey provides an exhaustive review of more than 100 papers published in peer-reviewed journals and conferences, gathered in a validated Excel database, including structured information on methodology, datasets, languages, evaluation approach, advantages, drawbacks, and research gaps. Through the examination of the entire collection of literature, the survey intends to give an impartial account of the recent advances made in the field of OCR and DLA.

This survey reviews the progress of OCR technology from rule-based approaches to deep learning, transformer-based, and vision-language approaches (Omee et al. 2012). It also reviews DLA approaches for document segmentation, structure analysis, reading order prediction, and joint OCR layout methods (Hasan et al. 2024). Moreover, the survey reviews the existing publicly available datasets, benchmarking techniques, evaluation metrics, and languages, with particular focus on low-resource languages and historical document collections (Nguyen et al. 2021). Studies taken into consideration in the present survey were chosen according to the novelty of their methodological contributions and the availability of detailed information on their strengths, limitations, and research gaps (Maung et al. 2025), (Sakib et al. 2022).

1.3 Contributions of This Survey

This review is a detailed critique of the current developments in OCR and DLA on the basis of systematically verified literature. Firstly, it offers a unified taxonomy framework that classifies the existing methods on the basis of methodology, data sets, and linguistic coverage for comparative analysis across various lines of research (Nida).

Secondly, it discusses the advantages and drawbacks of the existing methods from the perspective of recognition accuracy, scalability, robustness, computational efficiency, and generalization ability, pointing out the current methodological tendencies and technical challenges (Ding et al. 2026), (Chen et al. 2025). Finally, it summarizes the identified gaps in the existing scientific research, such as bias in datasets, inconsistent benchmarking, lack of annotations, and limited support for low-resource languages, and provides a structured roadmap for future research (Noor 2005), (Chowdhury et al. 2025).

1.4 Organization of the Paper

The study is systematically structured into ten sections, covering the theoretical foundations, methodology of review, research trends, applications, challenges, and future directions. Section 1 provides an introduction to the research background, motivation, and scope of the present literature review, with special attention to deep learning-based OCR and document understanding in Bangla and other low-resource languages. Section 2 introduces the main concepts of OCR, DLA, IE, and KIE, along with major challenges related to processing Bangla documents and low-resource languages. Section 3 introduces the methodology of the literature review, which includes the review methodology, the inclusion and exclusion criteria, and the limitations of the chosen review approach.

Section 4 constructs a taxonomy of OCR and IE systems based on the level of document processing, task type, document type, learning paradigm, and evaluation criteria. Section 5 includes the bibliometric and trend analysis of the existing OCR and IE research, such as publication trends and dataset utilization in high-resource and low-resource languages. Section 6 discusses the comparison of OCR and IE datasets, including dataset language type, document type, number of images, limitations of the dataset, and bias issues. Section 7 deals with the analysis of performance and errors in OCR and IE, including language-specific challenges, evaluation metrics, and comparative analysis of KIE techniques.

Section 8 provides a discussion of the major real-life applications of OCR and document understanding, including land records and legal documents, financial statements, healthcare, history and culture archival, and government and public services digitization. Section 9 covers the discussion of the major challenges and future research directions, including document quality and noise, mixture of handwritten and printed documents, low-resource and domain-specific languages, model generalization and fairness, and new trends, including self-supervised learning, few-shot learning, multimodal approaches, large language models, and scalability. Finally, the section 10 concludes the paper and outlines the key conclusions, contributions, and the long-term vision of the review study.

2 Background and Key Concepts

This section highlights the important basics of OCR and DLA. These are two vital concepts involved in the process of document image analysis. The section provides a

brief historical review of these concepts along with methodologies and their applications. This will establish the foundation for the comprehensive analysis presented in the following sections.

2.1 Optical Character Recognition (OCR)

OCR is an automatic technology for converting text from scanned documents, photos, or other visual media into machine-encoded and editable digital text. As one of the basic technologies for analysis of document images, OCR allows efficient digitization, information extraction , preservation, and automated processing of documents (Maliha et al. 2026). The classical approach to OCR focused on the development of methods of handcrafted feature extraction, pre-processing, character segmentation, and classification of characters using rules (Sharma et al. 2025), (Bhattacharya et al. 2016). Such methods showed good performance when used for clear and structured documents; however, their effectiveness dropped sharply for degraded images, complicated fonts, handwriting, multilingual documents, and low-resource languages (Jabde et al. 2025), (Pal et al. 2024).

The advent of deep learning has changed the approach to OCR by substituting manually engineered features with data-driven feature extraction (Rombach and Fettke 2025), (Wang 2025). CNN-based models provided the improvement of feature extraction and character recognition, whereas recurrent neural networks and attention mechanisms made sequence modeling more effective for word and line text recognition (Yu et al. 2021), (Hasan et al. 2023). Recent advancements in OCR include models based on transformer architecture and vision-language models, which utilize long dependencies and multimodal representations to provide better performance in case of irregular layout and complex document images (Chung and Choi 2025). Despite all the progress, the problems with limited availability of labeled datasets, poor cross-domain generalization, computational cost, and lack of models supporting low-resource and multilingual documents persist in all the available works (Prakash and Hanumanthaiah 2022). These limitations indicate that improving recognition accuracy alone is insufficient for developing reliable document understanding systems.

2.2 Document Layout Analysis (DLA)

DLA is aimed at recognizing the structure of document images and detecting logical and physical components such as regions of text, titles, tables, figures, forms, and reading order (Wang 2025). Accurate layout analysis is required to get the needed structural context for proper OCR and further analysis. Early approaches to DLA were largely based on heuristic geometry, projections, connected-component analysis, and manually constructed layout heuristics (Mahadevkar et al. 2024). While such approaches were sufficient for analyzing pages of rather simple structure, they failed to generalize to documents with complex page structure involving multi-column layout, embedded tables, various graphical elements, handwritten annotations, and irregular page design (Patel and Bhatt 2020), (Yang et al. 2022).

Deep learning revolution in recent years brought significant progress in DLA via CNN-based object detection, semantic segmentation, graph representation, and transformer models (Ding et al. 2026). Such models allow achieving higher accuracy in region detection by analyzing local visual features and global structural patterns (Renda 2023). Moreover, OCR-DLA frameworks have shown that accurate layout analysis allows improving text recognition by preventing segmentation errors and maintaining document reading order, especially for historical or complex document collections (Qi et al. 2025), (Ly et al. 2025). However, modern research still faces numerous issues like insufficient annotated data for layout analysis, unreliable evaluation strategies, low efficiency, and generalizability across different document domains and languages (Favish 2019). Solving such problems is crucial for the development of advanced and fully automated document understanding systems.

2.3 Information Extraction (IE) and Key Information Extraction (KIE)

Information Extraction is an essential technology in natural language processing, the task of which is the transformation of unstructured natural language texts into a structured format (Yang et al. 2022). Information Extraction's goal is to identify and extract the specified types of factual information in the form of entities, relations between the entities, and events (Rombach and Fettke 2025). Structured output plays an important role in different applications for information retrieval and knowledge management (Atagong et al. 2025), (Omee et al. 2012). Basic tasks of Information Extraction include Entity Recognition, or identification and classification of named entities people, organizations, and locations in text, and Event Extraction, or identifying occurrences of certain events and their participants, time, and location (Yang et al. 2022), (Qi et al. 2025).

Multimodal Information Extraction is an emerging direction that uses the techniques of multimodal learning together with Information Extraction for extraction of information not only from textual data but also from audio, images, and videos (Cao et al. 2023). Table 1 presents a comparative study of our survey with recently published review papers based on differentiating features. Each review paper is rated for different features such as search period, number of review papers analyzed, analysis of OCR, DLA, Bangla language, dataset analysis, bibliometric trends, and analysis of LLMs or VLMs. For identification purposes, the review papers are assigned serial numbers from one to seven, which serve only as labels for the corresponding papers in the table and do not indicate ranking or importance. Star ratings are used to rate the level of analysis for each of the above features, and it can be seen that our survey has comprehensive coverage of all the features (Mahadevkar et al. 2024).

Key Information Extraction is a special direction of Information Extraction which is focused on extracting a set of key data points (also known as "key fields") from documents and structuring them (Yu et al. 2021). The importance of KIE is defined by the need for effective archiving, fast indexing, and advanced document analysis in many spheres of activity such as finance, accounting, and taxation (Rombach and Fettke 2025). KIE is characterized by its own challenges when working with visually rich documents such as receipts, forms, and financial reports (Stanisławek et al. 2021).

Table 1 Comparative analysis of the distinguishing characteristics of the proposed review and selected existing studies.

Evaluation Criteria	**Our Study**	**One (Ding et al. 2026)**	**Two (Sharma et al. 2025)**	**Three (Han et al. 2025)**	**Four (Rombach and Fettke 2025)**	**Five (Atagong et al. 2025)**	**Six (Ding et al. 2026)**	**Seven (Tuama and Mohamed 2025)**
Published year	- -	2026	2025	2025	2025	2025	2026	2025
Search Period	1994-2026	Not mentioned	2010-2024	2015-2025	2017-2024	2017-2025	Not mentioned	2020-2025
Papers Reviewed	100+	Not mentioned	94	Not mentioned	130	30	Not mentioned	31
Focus on OCR Task	★★★★★	★★★	★★★★★	★★★★	★★	★★★	★★★	★★★★★
Focus on DLA Task	★★★★★	★★★★	★	★★	★	★★	★★	★★
Bangla Language Coverage	★★★★★	★	★★★	★	★	★	★	★
Analysis of OCR or KIE Datasets	★★★★★	★★★★	★★★★	★★★	★★★★	★★★	★★★★	★★★★★
Bibliometric or Trend Analysis	★★★★★	★★★	★★★★	★★	★★★	★★★★	★★	★★
LLM or VLM Coverage	★★★★★	★★★★★	★★	★★	★★★★	★★★	★★★★★	★★★

One of the most relevant challenges is the proper use of text and visual information in documents for obtaining richer semantic representation and reducing ambiguity (Cao et al. 2023). Such factors as document visual and layout information, like positioning of the text, font size, and table layout, play an important role in understanding the content of documents (Ding et al. 2026), (Antonio et al. 2022). Then documents may have very diverse layouts, font sizes, and structures, which complicates the extraction process (Tewani et al. 2021). In order to extract text from scanned documents, KIE systems usually use Optical Character Recognition (Rombach and Fettke 2025). Errors during the OCR process inevitably lead to incorrect entity information.

2.4 Challenges in Bangla and Low-resource Document Processing

Research into the field of text recognition has had much attention paid to it by scholars for many years, and many studies have been conducted on character, word, and text recognition (Hossain et al. 2023). However, while much progress has been made on languages such as English, Latin, Japanese, and Chinese, the creation of a high-end Optical Character Recognition system for languages such as Bangla is still a work in progress (Maliha et al. 2026), (Mazumder et al. 2025). As the national language of Bangladesh and one of the important languages in parts of India, Bangla is one of the most widely spoken languages in the world, making it an important language from both academic and economic points of view (Saha and Rahman 2024). Despite this importance, there has not been much research done on Bangla handwriting recognition, with many text recognition systems designed for the English language (Islam et al. 2023). Looking at the history of the development of deep learning and neural networks for OCR reveals that their usage for low-resource languages is still a challenge (Rabby et al. 2024). These are some challenges:

2.4.1 Script Complexity and Characteristics of Bangla Language

The scripts in the Bangla language comprise very complex ligatures, variable spacing, and non-uniform word boundaries. Such characteristics make OCR systems more likely to generate high error rates. Bangla script has over 300 compound characters, and these characters can be written in many different ways (Saha and Rahman 2024). The complexity of the structure of the Bangla language is one of the main reasons that leads to unsatisfactory results of OCR (Rabby et al. 2024). The cursive form of Bangla characters and abundance of diacritic marks (combining vowels and consonants to create graphemes) create additional challenges in handwritten text recognition. There are more letters in the Bangla language, leading to complexity of its glyphs (Maliha et al. 2026), (Chaudhury et al. 2022).

2.4.2 Data Scarcity and Diversity

Datasets available for the English language for NLP tasks are richer and larger than those for Bangla (Islam et al. 2021). The reason behind the insufficient development in handwriting identification in the Bengali language lies in the lack of large and diverse datasets (Hossain et al. 2023). The major difficulty in Bangla OCR tasks is covering

various types of documents like computer compositions, letterpress, typewritten, and handwritten documents (Garrido-Munoz et al. 2025).

2.4.3 Low Resource Language Processing Specifics

In contrast to high-resource languages, low-resource languages are subjected to microscopic study (Han et al. 2025), (Maliha et al. 2026). The majority of text recognition algorithms are created for processing in English, and there is no specific algorithm for recognizing Bangla handwriting (Hossain et al. 2023). Processing degraded historical documents necessitates developing language models that can support lexicons from primitive indigenous vocabularies, as texts often contain older linguistic forms (Qi et al. 2025), (Chaudhury et al. 2022).

3 Review Methodology

This section outlines the methodology adopted to conduct this review and ensure a systematic and comprehensive analysis of the existing literature on OCR and document understanding. It describes the search strategy used to identify relevant studies, including the databases, keywords, and time frame considered during the literature collection process. Furthermore, the section explains the inclusion and exclusion criteria applied to select high-quality and relevant studies while maintaining the consistency and academic rigor of the review. Finally, it discusses the limitations of the adopted methodology, including potential publication bias, coverage gaps, and other factors that may influence the completeness and generalizability of the findings presented in this survey.

3.1 Literature Search Strategy

An extensive literature search was performed, and relevant research articles on Bangla OCR and deep learning-based document digitalization were collected mainly from Google Scholar. A variety of combinations of keywords were used to retrieve relevant research articles on Bangla OCR and deep learning-based document digitization using Boolean logic to narrow down the search results. The keywords used for this purpose were “Bangla OCR”, “Bengali OCR,” “Deep Learning” and “Bangla Document Recognition,” “CNN or Long Short-Term Memory (LSTM) or Transformer” and “Bangla OCR,” and “Bangla Document Digitization.” The research articles were filtered based on publication dates ranging from 1994 to 2025, considering that this period witnessed major breakthroughs in deep learning techniques for performing OCR tasks. Apart from this, backward citation tracking was also performed to retrieve relevant research articles from the list of references of selected articles to cover the entire domain of research.

3.2 Inclusion and Exclusion Criteria

To maintain the relevance and quality of the literature selected for review, inclusion and exclusion criteria are used during the literature review process. The literature review process considered studies related to deep learning technologies, including

CNN, Recurrent Neural Network (RNN), and Transformer networks, and specifically addressed Bangla or Bengali script recognition problems. The literature review process also considered studies related to printed and handwritten documents written in the Bangla language and addressed problems in document layout analysis and OCR systems. At the same time, studies related to traditional or rule-based OCR technologies are excluded from the literature review process. In addition, studies without proper experimental validation are excluded to maintain the academic integrity of the literature review process. Also, to maintain the originality of the literature review process, duplicate publications and extensions of previously published literature are excluded from the literature review process.

3.3 Limitations of the Review Methodology

The study mainly targeted research papers on deep learning-based OCR. Even though deep learning-based techniques are currently the best in performance, excluding traditional machine learning-based and rule-based techniques makes it difficult to conduct a comprehensive review. The review may also have publication bias because it mainly targeted articles that are published in indexed journals and conferences. We may overlook important contributions to Bangla OCR research from local, non-indexed sources. Another limitation of this review is that it was conducted within a specified time frame and with strict criteria when selecting research papers. Therefore, some important studies may have been excluded if they were published outside the specified time frame or did not meet the inclusion criteria. Therefore, the review may not be comprehensive in covering all possible research associated with the topic.

4 Taxonomy of OCR and Information Extraction

In this section, a taxonomy of OCR and IE systems is developed, with particular emphasis on deep learning document digitization and analysis. The taxonomy structures existing literature according to document processing levels, task attributes, and architecture, with a specific focus on Bangla and other under-resourced languages. Such a taxonomy is essential for comprehending the OCR pipeline and other document understanding systems.

4.1 Taxonomy Design Principles

The earlier OCR systems were characterized by a document processing pipeline, consisting of a sequence of rigid and rule-based modules, such as preprocessing, segmentation, feature extraction, and classification (Amini et al. 2026), (Smith 2007). Given the design, the systems were largely affected by noise and variations in the scripts, and as a result, more flexible frameworks were needed. Recent research suggests organizing the OCR and IE systems by processing granularity, document type, and learning paradigm to more accurately reflect contemporary deep learning frameworks (Ding et al. 2026), (Bhattacharya et al. 2014). For low-resourced languages such as Bangla, the design of the taxonomy must consider the intricate nature of the scripts, the limited annotated datasets, and the inconsistent document layouts (Maliha et al.

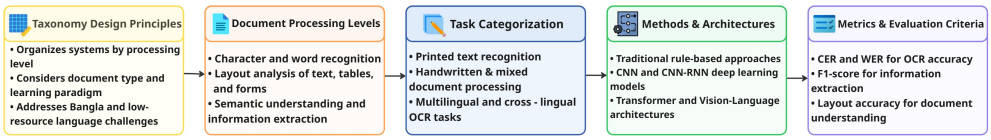

Fig. 1 Taxonomy of OCR and Information Extraction.

2026). The surveys on Bangla OCR highlighted the recognition level and document type as important parameters for system categorization to ensure equitable cross-study comparisons (Inunganbi 2024). Figure 1 presents the taxonomy of OCR and Information Extraction, categorizing document analysis systems based on processing levels, task attributes, document types, learning paradigms, and resource settings.

4.2 Levels of Document Processing

Pipelines of OCR and document understanding systems are comprised of multiple levels. At the lowest level, individual glyphs are recognized. Segmentation-and-handcrafted-feature methods, which are the default for Bangla character recognition, are still able to achieve some success, but are limited in their ability to address compound characters and character modifiers (Parui et al. 2008). At this level, deep learning models, and more specifically CNNs, have proven their worth a great deal (Shi et al. 2016).

At the next level of word processing, individual character predictions are grouped as words in a sequence. While Connectionist Temporal Classification (CTC) loss, along with a CNN-RNN architecture, provides better recognition of both handwritten and printed texts in comparison to previous models, the implicit segmentation and context-aware decoding that CTC loss provides seem to improve recognition in the sequence of letters (Garrido-Munoz et al. 2025), (Breuel et al. 2013). At the document layout level, systems provide more focused analysis of the document by constructing the reading order and describing regions of text, tables, and forms. With layout-aware pre-trained models, such as Layout Language Model (LayoutLM) (Xu et al. 2020), layout analysis has more recently been incorporated into OCR systems. The highest level of document understanding systems is the semantic level (Ding et al. 2026). At this level, systems go beyond the mere recognition of documents to interpret and understand their meaning (Mahadevkar et al. 2024). The integration of text content and layout, which is made possible by Vision-language models and transformer-based architectures, provides the ability to support multiple functions, such as, but not limited to, document classification and extraction of key information (Qian et al. 2021).

4.3 Task Categorization

The OCR and IE tasks have been defined based on specific attributes of document content and linguistics. The recognition of printed texts has been studied and structured more thoroughly for printed documents than for others, and has been researched mainly in high-resource languages (Smith 2007). The recognition of handwritten texts brings major issues due to variability and has been the primary focus of Bangla OCR

research (Sarkar et al. 2012). Documents combining printed and handwritten texts are widespread in many real-world cases, for example, in administrative forms and educational resources (Shugaba et al. 2026). These documents require the construction of models to manage diverse visual patterns (Mondal et al. 2024). Multilingual documents present an added challenge as more than one script and language can be present in one document. Surveys show that the Bangla language tends to be underrepresented in multilingual OCR benchmarks, thus advocating the need for cross-lingual evaluation (Jabde et al. 2025), (Bhattacharya et al. 2014).

4.4 Methods and Architectures Overview

Most traditional OCR systems have relied on handcrafted construction features and handcrafted heuristics and rules. As a result, these systems are inflexible and cannot easily accommodate different kinds of documents (Smith 2007). The incorporation of deep learning, coupled with improvements in CNN-based feature extractors and RNN-based sequencing modalities, has provided a considerable boost in recognition accuracy (Shi et al. 2016). The newest models based on transformer architecture are proving to be the best models for OCR and document understanding (Amini et al. 2026). An example of the transformer architecture being nimble for end-to-end text recognition is Transformer-based Optical Character Recognition (TrOCR). For example, the LayoutLM model is layout-aware and incorporates layout information for some of the downstream tasks, such as information extraction (Li et al. 2023). These models are particularly useful for complex scripts and intricate layout designs, as they reduce the need for definite segmentation.

Table 2 provides a comparative review of various optical character recognition techniques based on five important generations. For each technique, it discusses sample models, advantages, disadvantages, and application areas. This table demonstrates the development of methodology from the basic template matching approach to the more complex multimodal and generative systems of end-to-end nature.

4.5 Metrics and Evaluation Criteria

The metrics used for OCR and IE system evaluations are based on task-specific evaluations. Approximately, Character Error Rate (CER) and Word Error Rate (WER) are commonly used for measuring the recognition accuracy of OCR systems (Shi et al. 2016), (Breuel et al. 2013). For precision extraction tasks, metrics such as recall and F1-score are used to measure extraction performance (Gu et al. 2022). Layout analysis accuracy, as well as accuracy in reading order, are used to reflect the importance of structural comprehension in document AI systems (Mondal et al. 2024). In the case of Bangla OCR, a persistent challenge for standard evaluation is still the absence of large-scale benchmark datasets (Inunganbi 2024).

5 Bibliometric and Trend Analysis

In this section of the analysis, a bibliometric analysis and trend analysis of Optical Character Recognition and Information Extraction is conducted, which is concerned

Table 2 Comparison of major approaches for OCR and document understanding.

Approach	Representative Models	Strengths	Weaknesses	Best For
Rule-based or Handcrafted	Template matching, HMM-based	Simple, low cost, works on clean printed text	Poor on degraded, noisy, handwritten documents, no generalization	Clean, structured, single-font documents
CNN-based	LeNet, AlexNet, VGG, ResNet	Automatic feature extraction, good on printed text	Limited long-range sequence modeling, needs segmentation	Character and word recognition
CNN-LSTM (CRNN)	CRNN, CNN-BiLSTM-CTC	End-to-end, handles variable-length text, no explicit segmentation	Not parallelizable, weak on long-range dependencies	Handwritten line recognition, scene text
Transformer-based	TrOCR, LayoutLM, LayoutLMv3	Global context via self-attention, handles complex layouts	High compute, needs large pretraining data	Complex scripts (Bangla), irregular layouts
Vision Language Models	Donut, Dessurt, Ocean-OCR	OCR-free, end-to-end, multimodal	High resolution needed, compute-heavy	Document understanding, KIE
LLM-based (KIE)	GPT-4, Gemini, DocLLM	Zero or few-shot, flexible reasoning	Hallucination, costly, variable output	Generative KIE, Q&A, document reasoning

with the change in research activities, prominent publication centers, and modification of methods used and data sources employed. The findings illustrate that document understanding research grows in complexity and global nature, especially with regard to the problems connected with resource-poor languages.

5.1 Publication Distribution over Time

Recent developments in the domain of OCR and Information Extraction have witnessed a considerable boost in research, largely attributed to progress in deep learning as well as a high need for automated document processing (Shugaba et al. 2026). Some authors have found that there has been a noticeable growth in the number of publications on the subject, especially since 2018, continuing until 2022 and 2023 (Abdallah et al. 2024). Thus, the research field remains relevant and constantly evolving. The fact that there is a lot of development indicates that the research domain

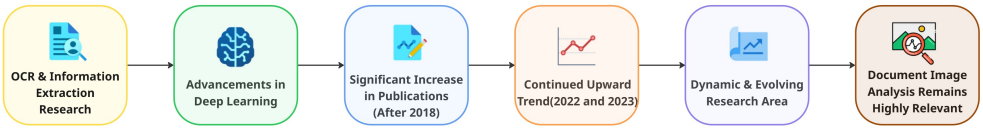

Fig. 2 Publication trend of OCR and Information Extraction research.

is mature but still evolving, where new techniques are constantly developed (Rabby et al. 2024). For instance, the domain of document image analysis attracts many publications and thus remains relevant. Figure 2 illustrates the trend of publications in OCR and Information Extraction research.

5.2 Research Venues and Communities

The spread of knowledge is mainly associated with scientific conferences and journals. International events where important new ideas are presented include the International Conference on Document Analysis and Recognition, the IEEE or CVF Conference on Computer Vision and Pattern Recognition, conferences of the Association for Computational Linguistics, and the International Conference on Pattern Recognition. All these events play a key role in informing the public about advances accomplished in Document Analysis, including text detection, processing, and retrieval (Subramani et al. 2011). It is possible to see the popularity of Document Image Analysis from the fact that papers presented at such conferences have received many citations lately (Wang 2025).

Journal publications play an important role, as their share tends to surpass or is almost equal to that of conference papers in the last couple of years, especially in 2022 and 2023 in the fields of study related to text line segmentation (Rabaev and Litvak 2025). Table 3 highlights major observations in bibliometric analysis and trends in OCR and information extraction research. Six major trends have been identified: increased publication, popular research platforms, change in the dataset from only English to multilingual and low-resource languages, changes in methods from conventional to deep learning to transformer to LLMs, geographical biases in contribution, and the gap between industry and academia. The table provides a concise overview of how the field has evolved and where gaps remain.

5.3 Dataset Usage Trends

Historically, studies in the fields of OCR and IE have mainly been dedicated to English-language documents, which have resulted in plenty of datasets available for this language, e.g., the IAM database (Ali et al. 2023) and different datasets for scanned receipts OCR and information extraction (Townsend et al. 2024). There is also an increasing understanding of the necessity of datasets in endangered and low-resource languages. Nowadays, there is an increasing trend towards developing and using multilingual and low-resource language datasets (Singh and Middleton 2025). The development of the "Gold Standard Bangla OCR Dataset" and vision-language

Table 3 Key trends identified in OCR and document understanding research.

Trend	Observation
Publication Growth	Sharp increase since 2018, peak in 2022–2023
Research Venues	ICDAR, CVPR, ACL
Dataset Shift	From English-only to multilingual/low-resource
Method Shift	Traditional → DL → Transformer → LLM
Geographic Bias	France, China, USA, and India lead; low-resource regions underrepresented
Industry vs. Academia	Industry leads model development, while academia leads datasets and low-resource research

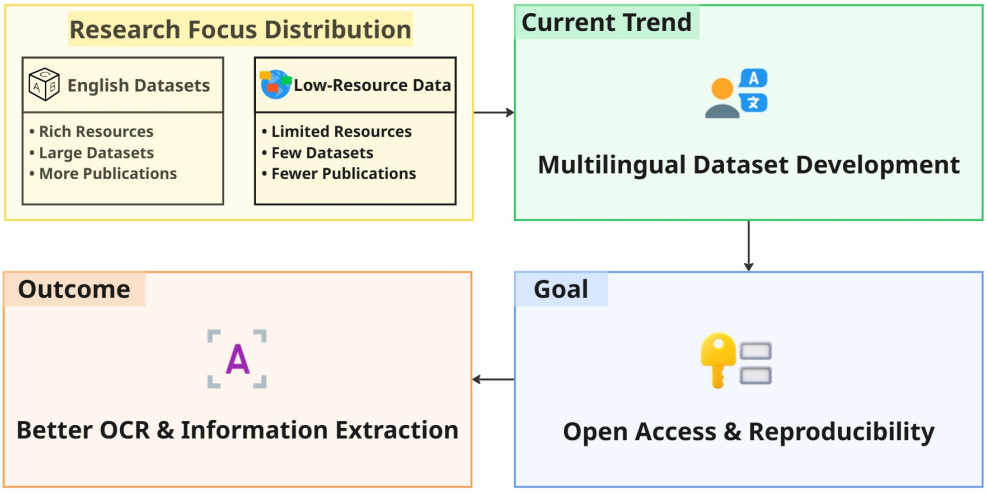

Fig. 3 Dataset usage trends in OCR and Information Extraction research.

model fine-tuning for low-resource languages like Manchu is very important at the moment (Ali et al. 2023), (Chung and Choi 2025).

Researchers urge the creation of open-access datasets in languages with fewer opportunities to promote transparency and further developments in this field (Rabaev and Litvak 2025). Nevertheless, there are still significant gaps for now, and the number of papers in ACL venues concerning many of the low-resource languages remains quite low. The influence of multilingual datasets on research in information extraction is being felt nowadays, which helps fill the gap (Yang et al. 2025). Figure 3 shows the progress in the development of datasets used in OCR and Information Extraction research.

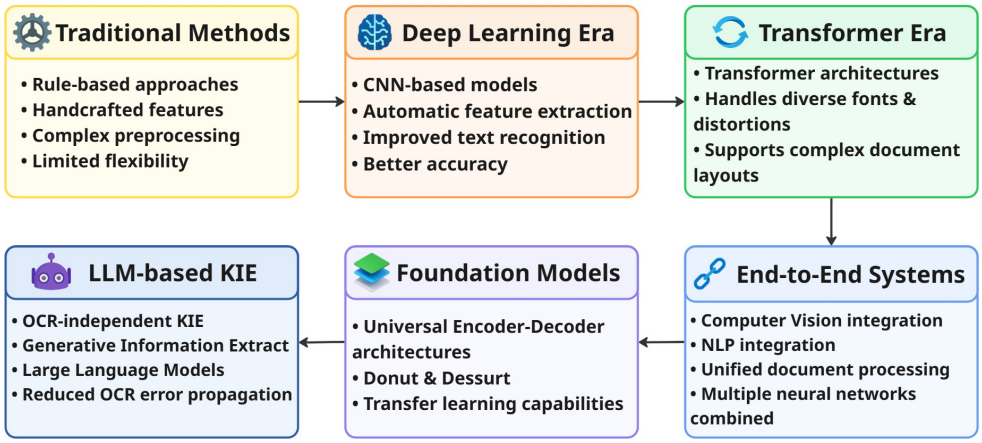

Fig. 4 Methodological evolution of OCR and Information Extraction.

5.4 Methodological Evolution

Methodologically, there was a radical change in the field of OCR and IE. Classical rule-based and machine learning methods, where feature engineering and heavy pre-processing steps were needed (Murad and Ali 2023), were outperformed by deep learning techniques (Rabby et al. 2024). Convolutional Neural Networks introduced novel methods for automatic feature generation and helped to increase the accuracy of scene text recognition (Murad and Ali 2023). The next milestone in this field was the development of transformer models that allowed building robust models able to handle various types of fonts, distortions, and complex layout structures (Constum et al. 2025). Contemporary end-to-end document understanding models incorporate different deep neural network architectures, including both computer vision and natural language processing components (Subramani et al. 2011). The tendency of the development is moving to the creation of universal encoder-decoder models, where large-scale models like Donut and Dessert prove their good language modeling skills using transfer learning (Constum et al. 2025). Some recent works include OCR-free and generative approaches to Key Information Extraction, where Large Language Models are used to prevent errors introduced by external OCR engines (Rombach and Fettke 2025). Figure 4 shows the methodological evolution of OCR and Information Extraction technologies.

5.5 Geographic and Institutional Contributions

Research on OCR and IE systems, including low-resource language cases, shows geographical and institutional inequality in contributions. There is no doubt about the substantial contributions to text line segmentation made by France, China, the USA, and India (Rabaev and Litvak 2025). Geographic citation gaps in NLP studies persist and concern less resourced languages in particular. As geographical location correlates with language use, overcoming these gaps will be crucial to advancing global language

technology (Joshi et al. 2020). Industry research laboratories play an increasingly significant role in transformer-based and large-scale pretrained model development, whereas academic organizations retain an essential role in creating datasets and innovating language-specific technology, including for low-resourced languages (Han et al. 2025), (Orife et al. 2020).

The development of Bangla OCR systems includes academic and regional research institutions contributions to creating essential datasets and developing tools. BanglaWriting offers data for Bangla handwriting recognition (Mridha et al. 2021). BaDLAD offers a dataset of a large-scale Bengali document layout covering several document types (Shihab et al. 2023). Open-source projects like bbOCR have contributed to Bengali document digitization and OCR studies (Zulkarnain et al. 2023). These contributions show the significance of work done by regional institutions in developing language and document-specific resources.

However, research on low-resource languages still faces numerous barriers connected with insufficient availability of datasets, computational power, funding, and research infrastructure (Han et al. 2025), (Joshi et al. 2019). Cooperation between universities, industry research laboratories, and the local language community may contribute to filling these gaps and the sustainable development of language technology (Orife et al. 2020). Development of multilingual pretrained models like mT5 gives opportunities to transfer language-specific knowledge and apply it to other languages, such as Bengali (Xue et al. 2021). On the other hand, projects like No Language Left Behind are aimed at developing technology for a wider range of languages (Costa-Jussà et al. 2022). However, a substantial research gap between high and low-resource languages still exists (Joshi et al. 2020). Hence, it is crucial to develop Bangla-specific datasets, benchmarks, open-source projects, and conduct further research.

6 Datasets and Benchmarks

Datasets and benchmark suites are a deciding factor in developing and comparing OCR and document understanding systems. The selection of the dataset will determine the generalizability of the system, its performance on the dataset, and the extent of progress achieved through research. Unfortunately, there is no uniformity in the datasets being used and in their evaluation.

6.1 General OCR Datasets

General OCR datasets are usually intended for use in text recognition at the levels of characters, words, or lines. Early OCR datasets were created using clean and high-quality scanned documents, and thus helped in the development of algorithms but could not represent the variation present in the wild (Pal et al. 2024). Later on, OCR datasets started to include images of degraded scans, scene text, and synthetic images to enhance robustness (Hasan et al. 2023). However, even though much improvement has been achieved with the help of these OCR datasets, there is still a bias in favor of Latin alphabets and structured documents (Taghva et al. 1994). The following Table 4 presents the comparison of several commonly used OCR datasets with respect to languages, document types, sizes, contents, and formats of images.

Table 4 Comparison of publicly available OCR datasets used for printed, handwritten, multilingual, and document image recognition tasks.

Dataset Name	Language	Dataset Type	No. of Images	Image Content	Image Format
Bengali OCR Dataset 1M	Bangla	Printed, Handwritten	1M+	Word	.jpg
SROIE Dataset v2	English	Printed	973	Receipts	.jpg
Page-level Annotated Text Recognition Assemblage	Bangla	Handwritten	996	Paragraphs (from 27 unique sources)	.png
ICDAR 2019 MLT OCR Dataset	Arabic, Bangla, Chinese, English, French, German, Italian, Japanese, Korean, and Latin	Printed	20,000	Different document sources	.jpg, .png
Tafsir Ibn Kathir Bangla	Bangla	Printed	6,413	Book	.png, .pdf
TextOCR	Mixed	Printed	25,100	Natural images	.jpg
OCR Dataset of Multi-type Documents	English and Vietnamese	Printed, Handwritten	4,266	Invoices, forms, ID cards, and real-life photographs	.png, .jpg

6.2 Key Information Extraction Datasets

KIE datasets move from just recognizing information to semantics-based labeling of entities like names, dates, totals, and IDs. Datasets for tasks related to invoices, receipts, and structured documents have resulted in joint modeling of OCR and layout analysis (Patel and Bhatt 2020). However, many KIE datasets utilize domain-specific layouts with minimal layout variations that make it impossible for the model to generalize to new layouts (Yang et al. 2025), (Ly et al. 2025). Furthermore, cost considerations may lead to very small datasets, thus increasing the possibility of overfitting and dataset optimization (Henge and Rama 2016).

6.3 Multilingual and Low-resource Datasets

Multilingual and low-resource datasets help to resolve the problem of unbalanced OCR research through the provision of data for the scripts or languages that lack coverage in current research (Singh and Middleton 2025). The availability of publicly available websites like Kaggle, GitHub, and Mendeley Data has made possible the publication of community datasets such as handwriting and historic documents (Nguyen et al. 2021). Examples of such datasets include BanglaLekha for Bangla handwriting, Mozhi-LR for low-resource scripts, and others (Noor 2005). However, despite the significance of the datasets, they are characterized by limited sizes, inconsistent annotation practices, and no benchmarks (Jabde et al. 2025), (Prakash and Hanumanthaiah 2022).

6.4 Dataset Limitations and Bias Issues

For both OCR and DLA datasets, biases within the data are major issues. Clean documents are more frequent in datasets than in reality, along with a majority language and certain document types, resulting in optimistic performance and poor generalization (Rahman et al. 2023). Even though synthetic data is valuable for extending the scale of training data, there might be problems in terms of simulating realistic noise distribution and layout variations (Yu et al. 2021). Moreover, the lack of consistent assessment procedures and data on metadata leads to challenges in replicability and comparability of results (Maung et al. 2025). Solving such problems would require larger and more varied data sets and the standardization of annotation procedures and benchmarking (Yang et al. 2025), (Sakib et al. 2022).

7 Performance Evaluation and Comparative Analysis

In this section, the critical analysis of OCR and KIE approaches used in this research is performed. The comparison is made among the conventional OCR engines, deep learning architecture approaches, transformer-based methods, and LLM-driven extraction approaches. Particular attention is paid to the script of the Bengali language due to its unique ligature-driven characteristics. The performance assessment is performed according to standard metrics quantification methods with respect to the literature.

7.1 OCR Model Performance

The OCR technology has progressed significantly from rule-based and handcrafted feature-based approaches towards deep learning architectures. The success of the OCR system greatly depends on the capability of the architecture to detect discriminative visual features, model character-level dependencies, as well as capture the contextual dependency of adjacent characters and words (Amini et al. 2026). Considering this, the widely used OCR approaches can be divided into the following categories: traditional OCR engines, convolutional and recurrent neural network-based architectures, and modern Transformer-based approaches. This section provides a review of the characteristics, pros, cons, and performance of the following four OCR approaches: Tesseract, TrOCR, CNN–LSTM-based OCR systems, and Transformer-based OCR approaches.

7.1.1 Tesseract OCR

Tesseract is one of the most popular open-source OCR engines and an example of the conventional OCR framework, which has experienced significant architectural progress over the years. The initial version of the Tesseract OCR engine was intended for printed text recognition using a series of image analysis, text-line detection, feature extraction, classification, and adaptive recognition algorithms (Smith 2007). Further developments of the engine included the implementation of neural network recognition modules, especially LSTM networks. Consequently, Tesseract has stayed an efficient OCR tool because of its openness, multilingualism, and low computational cost.

One of the strengths of Tesseract OCR is that it is capable of working well with clean, machine-printed texts having clear character edges and consistent layouts, thus being able to perform recognition without the need for a custom-trained deep-learning model. However, its recognition results can be rather poor for noisy, degraded, low-resolution, handwritten, or stylized texts. Recognition can be additionally affected by complex layouts, non-uniform character spacing, illumination, and background. Consequently, despite remaining applicable as a lightweight OCR engine and baseline for comparison in research papers, the performance of Tesseract OCR can be insufficient for some recognition tasks (Smith 2007).

7.1.2 TrOCR

TrOCR constitutes a breakthrough in end-to-end OCR through utilization of pretrained Transformer models for extraction of visual representations as well as for generation of text (Li et al. 2023). As opposed to conventional OCR architectures, which might include separate feature extraction, sequence modeling, and language-model-based postprocessing steps, TrOCR models recognition as an image-to-text generation task.

Due to the use of pretrained Transformers, TrOCR becomes capable of benefiting from large-scale pretraining, as well as fine-tuning on the task of interest. In the original TrOCR study, the architecture has been investigated for printed, handwritten, and scene-text recognition tasks and outperformed other recognition approaches (Li et al. 2023). Integration of visual and linguistic information in a single architecture makes it especially useful when character recognition depends on context.

On the other hand, the computation costs of TrOCR and other pretrained Transformer architectures tend to be higher than those of simple OCR algorithms. Their performance may also be affected by the availability of appropriate pretraining and fine-tuning data, especially if there are notable differences between the target and training languages, scripts, document domains, or visual characteristics. Thus, although TrOCR offers a robust end-to-end architecture, computation costs and domain adaptation are essential factors for deploying it (Li et al. 2023).

7.1.3 CNN-LSTM Based OCR Models

The combination of a CNN and an LSTM network allows integrating feature extraction and sequence modeling capabilities within a single model architecture. The Convolutional Recurrent Neural Network (CRNN) model represents one such architecture in

which feature extraction, sequence modeling, and transcription are combined into a single end-to-end system (Shi et al. 2016). The role of the CNN part is to map the input text image into a sequence of high-level features, while the recurrent layers are used to model interdependencies between neighboring characters (Wang 2025). CTC is widely used for transcription without explicit character segmentation.

CNN-LSTM-based approaches have been quite successful in scene-text and handwriting recognition due to their joint exploitation of spatial and sequential properties. Unlike traditional OCR algorithms, they are capable of learning discriminative visual representations in an end-to-end manner using training data; they are also capable of handling variable-length text sequences and thus can be employed in practice. However, due to their nature of processing sequential information in a recurrent manner, such networks cannot be easily parallelized, and consequently, processing time increases in the case of long sequences. Moreover, while LSTMs are capable of capturing context dependencies within extended sequences, their sequential nature might be inferior to the state-of-the-art self-attention approach (Shi et al. 2016).

7.1.4 Transformer-Based OCR Models

The impact of Transformer-based architectures on sequence modeling is associated with the use of self-attention modules (Dong et al. 2021). While RNNs operate on sequential inputs in a recurrent way, Transformers are able to model relations between several positions with the help of attention operations performed in parallel. The ability of the self-attention module to weigh input elements allows modeling contextual relations between them even at rather large intervals (Vaswani et al. 2017), (Dong et al. 2021). It is important in the context of OCR since the recognition of a particular symbol or token may be influenced by the context. Modern Transformer-based OCR systems can use encoder-only, encoder-decoder, and vision-language approaches depending on the type of OCR task. In the case of vision-based OCR, image encoders can encode image regions or patches into contextualized representations, while Transformers can be used to decode the textual sequence corresponding to the encoded visual input. TrOCR is an excellent example of such an architecture: the system includes an image Transformer encoder and a text Transformer decoder (Li et al. 2023).

Nonetheless, OCR Transformer architectures have several issues associated with them. The computational resources required for their training and inference operations might be quite significant for high-resolution images and long sequences, for instance. Moreover, pre-trained models can demonstrate inferior results when applied to different languages, scripts, document structures, and conditions compared to those encountered during the pre-training process (Li et al. 2023), (Vaswani et al. 2017). Thus, domain-specific fine-tuning and proper choice of training data become critical for reliable performance of specialized OCR tasks.

7.1.5 Discussion of Comparative Performance

The development of OCR from traditional OCR engines like Tesseract to CNN-LSTM architectures and then to Transformer-based solutions shows the trend of moving from

hand-crafted recognition pipelines to end-to-end representation learning and contextual sequence modeling. Tesseract is still useful owing to its simplicity, availability, and lower requirements for computational resources compared to other models. While they work fairly well for clean, printed texts in the English language, they perform poorly when processing Bangla scripts due to their complexity (Smith 2007).

CNN-LSTM models combine convolution-based feature extraction and recurrent sequence modeling (Shi et al. 2016). Such models perform averagely on multilingual data and fail to represent long-range contextual dependencies. TrOCR uses a vision transformer for encoding images and an autoregressive decoder (Li et al. 2023). The ability to model global context using the self-attention mechanism results in decreased CER and WER in low-resolution and noisy images. In general, OCR models based on transformers are more efficient for recognizing text, including those written in complicated scripts, such as Bangla. Table 5 presents a comparative summary of representative OCR and document information extraction studies.

7.2 OCR Evaluation Metrics

The following are the measures that are applied to measure the efficiency and effectiveness of OCR and KIE in relation to text recognition and information extraction, respectively.

7.2.1 Character Error Rate (CER)

The CER is an evaluation metric that computes the accuracy rate at the character level. It determines the ratio of substitutions, deletions, and insertions to the total number of characters in the ground truth text. The CER is especially relevant to Bangla OCR because of the complexity of compound characters and diacritics (1).

$$CER = \frac{S + D + I}{N} \tag{1}$$

where S denotes substitutions, D denotes deletions, I denotes insertions, and N represents the total number of characters in the ground truth.

7.2.2 Word Error Rate (WER)

WER is used to measure the accuracy of OCR on a word basis by measuring the number of substitutions, deletions, and insertions done to the words against the total number of words in the reference texts(Karpinski et al. 2018). Word error rate determines the correctness and readability of the OCR output (2).

$$WER = \frac{S + D + I}{N_w} \tag{2}$$

where N_w represents the total number of words in the ground truth.

7.2.3 Precision

Precision can be defined as the ratio of the information that has been extracted correctly to the total amount of information that has been extracted by the system (3).

Table 5 Comparison of representative OCR and information extraction studies based on datasets, methodologies, and reported performance.

Paper	Dataset	Methodology	Performance Result
Li et al. (2025)	1,720 scanned pages with 8,225 labeled regions	YOLO detects layout, CNN recognition, augmentation training, and projection-based segmentation	Segmentation and Recognition Accuracy: 97–99%
Niloy et al. (2025)	CMATERdb 3.1.1 and BanglaLekha-Isolated	CNN for spatial feature extraction and RNN/LSTM for sequential modeling	CNN slightly outperformed RNN
Singh (2025)	Various benchmark datasets	Survey of CNN, RNN, and hybrid methods for text detection and recognition	Detection F1: 0.92; Overall Accuracy: 90.5%
Opu et al. (2024)	Combined CMATERdb and ISI Bangla handwritten datasets	Comparison of LeNet, AlexNet, VGG, and lightweight CNN with augmentation and normalization	Accuracy: 98%
Pal et al. (2024)	Custom isolated Bangla characters and handwritten word images	Two-stage SSD with transfer learning. Stage 1 fine-tuned on characters; Stage 2 reused the backbone for handwritten word recognition	F1 Score: 82.86%
Cao et al. (2023)	SROIE, FUNSD, and CORD	Transformer integrating text, layout, and visual features with prompts	SROIE F1: 91%; Better than LayoutLMv3
Li et al. (2023)	684M synthetic text lines and multiple benchmark datasets	Transformer encoder-decoder with ViT/DeiT/BEiT encoder and RoBERTa/MiniLM decoder	SROIE F1 Score: >96%
Monsur et al. (2023)	11K real NID images and 50K synthetic images	OCR-free Donut with synthetic documents and multilingual tokenizers	Bangla Field Accuracy (TED): 81.53%
Rahman et al. (2023)	788 full-page handwritten documents	Unsupervised line segmentation using Otsu, Canny, morphology, Hough, and OPTICS	Line Segmentation FM: 81.57%; mAP: 0.547
Dergachov et al. (2021)	Scanned and photographed documents	Denoising, binarization, skew correction, segmentation, and normalization	OCR Accuracy Improvement: 10–25%
Islam et al. (2020)	Custom dataset of 600 natural scene images	Bounding-box text detection, word segmentation, character extraction, and CNN-based recognition	Detection: 89.25%; Extraction: 94.50%; Overall: 95.76%

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

7.2.4 Recall

Recall is the ratio between the correct extraction of information to the total amount of information that is relevant in the ground truth (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

7.2.5 F1-Score

F1-score is the harmonic mean of both and serves as an equilibrium evaluation measure for the information extraction process (5).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

7.3 Comparative Analysis of KIE Techniques

KIE reformulates the unstructured OCR results into structured objects. This process is normally realized in two broad paradigms. The first is structured extraction models with deep learning, whose architectures are generally designed to jointly infer from text, layout, and visual information, identify entities, and extract relevant fields. The second is generative extraction systems with LLMs, which rephrase extraction as a text-to-text generation task, and we instruct or fine-tune an LLM to generate the structure directly from the OCR text and context (Wang et al. 2025).

7.3.1 Deep-Learning-Based KIE

Traditional approaches use sequence labeling approaches, e.g., BiLSTM-CRF and frame entity extraction as a token classification task (Lample et al. 2016). Better approaches like LayoutLM fuse text and spatial layout information for better understanding of document images for structured documents (Xu et al. 2020).

Some strengths of these approaches include their relatively high accuracy on domain-specific tasks; the structure of the predicted outcome can be predetermined and rendered deterministic as well; they lead to efficient inference time.

Some limitations, on the other hand, include their need for a large-scale labeled dataset and low flexibility to accommodate document design changes for models trained.

7.3.2 LLM-Based KIE

LLMs like GPT-4 and Gemini frame KIE not as a token labeling task but as a problem of generation-reasoning, enabling them to achieve coherent outputs through a pre-specified document format from an in-context prompt without the need for individual

token-level labeling. This paradigm comes with several pros: zero-shot or few-shot performance, generalization over diverse document formats. This paradigm also has several cons: hallucination, higher costs, and variable output format (Wang et al. 2025).

7.4 Error Analysis by Language Type

Bangla script poses unique OCR problems that can be blamed on its orthographic level.

7.4.1 Ligatures

The Bengali language has many complex characters, which are a result of a mixture of consonants. These ligatures are often confused with segmentation-based systems, thus enhancing substitution and deletion errors.

7.4.2 Matra and Diacritics

The horizontal matra is used to link the characters in words. This structure can also have noise or partial discontinuities that typically lead to segmentation errors. The diacritics used above or below words are often misaligned, resulting in wrong identification.

7.4.3 Stroke-Level Similarity

A few of the Bangla characters have shapes that are almost identical, with minor differences in the strokes. Such characters can be confused with CNN-based models in low-resolution images. The transformer-based systems address this problem by using contextual inference. The analysis of errors shows that script-specific peculiarities have a strong impact on CER and WER, which makes the use of script-aware training tactics relevant.

7.5 Cross-Script Generalization Capabilities

Cross-script generalization represents the ability of a system to perform across multiple languages. In traditional approaches like Tesseract, there is a need for language-dependent training data and the inability of these systems since the performance decreases in relation to the usage of unknown scripts (Smith 2007). Architectures based on Transformers allow multilingual pretraining and the ability to obtain language-independent representations, which improves zero-shot transfer (Dong et al. 2021). Models that are layout-aware, like LayoutLM, are more generalized as they include spatial data and become resistant to any particular format of documents (Xu et al. 2020). Systems that rely on LLMs have a high level of adaptability due to the use of large multilingual training corpora (Wang et al. 2025). However, there may be a risk that performance in low-resource scripts underrepresented in pretraining will be low.

7.6 Strengths and Weaknesses of Current Approaches

A critical comparative analysis shows that different OCR systems have different trade-offs. In older systems like Tesseract, it is possible to distinguish smallness, ease of implementation, and cost-effectiveness. Nevertheless, there are no complex contextual modeling components in them, and it is difficult for them to cope with complicated scripts and noise. The architecture of CNN-LSTM models contributes to improving the accuracy due to the combination of visual and sequential learning components (Shi et al. 2016). However, repetitive patterns prevent parallelization and long-dependency modeling. Transformer-based OCR models like TrOCR represent cutting-edge systems with a global self-attention model (Li et al. 2023). It improves multilingual recognition but requires high computational power. GPT-4 (LLM-based KIE) has a high level of reasoning and can extract information flexibly. The problem with hallucination and high costs of computation calls for additional mechanisms of validation in order to ensure reliability. In conclusion, hybrid systems with transformer-based OCR and LLM assisted KIE modules represent the best trade-off.

8 Applications and Use Cases

In this section, we are going to talk about practical applications of OCR and IE in various domains. Special attention will be paid to practical applications of these technologies with low-resource languages like Bangla. The recent improvements in deep learning and document understanding are changing the way information based on documents is being digitized, managed, and analyzed.

8.1 Land Records and Legal Documents

Firstly, OCR technologies are helping the government and legal institutions to manage land records (Favish 2019). In particular, OCR is useful in converting written and printed text into a machine-readable format. This makes the process of processing data much more efficient since, at the moment, it is done manually, which leads to slow and inaccurate results (Ali et al. 2023). Also, with OCR technologies, the possibility of fraud analysis in legal documents and extraction of accurate information from archives becomes much higher (Hossain et al. 2022). Therefore, there is an evident necessity to create a robust system for recognizing handwritten words for governments and companies to analyze large amounts of handwritten land and legal documents. There is a need for special OCR software for complex script languages like Bengali in order to analyze deteriorated documents with fading, stains, hard-to-read handwriting, and complicated layout. For example, datasets such as BaDLAD, a multi-domain Bengali Document Layout Analysis dataset with human-annotated property deeds and public government documents, can be used to improve layout analysis and the accuracy of OCR in such cases (Shihab et al. 2023).

Modern OCR technologies in Bengali are able to analyze the layout of the document and detect embedded images and signatures in order to make it completely digitized. Open-source OCR tools like bbOCR are helpful in analyzing and digitalizing historical and legal documents in Bengali (Zulkarnain et al. 2023). Current research

is focused on information degradation and improving the text extraction accuracy in order to preserve and extend the legal and land record accessibility in the era of information. First of all, the process of OCR starts with scanning the document into a digital image. This image is pre-processed with steps like binarization, noise removal, and skew detection. At the last step, the system extracts the features and recognizes characters in order to create accurate, machine-readable text. Additionally, technologies for recognizing handwritten digits also help in the analysis of numerical information in legal and land documents (Islam et al. 2023).

8.2 Finance

The financial industry greatly benefits from OCR technology, which recognizes characters from invoices, receipts, and other financial documents to convert them into machine-readable form. It enhances efficiency and minimizes errors due to human factors (Rexhepi et al. 2025). Manual processing of financial documents is tedious and error-prone, but OCR technology automatically extracts the data required for financial activities like expense management, recordkeeping, and auditing. The recent developments in machine learning and cloud technology have greatly helped in improving the efficiency of OCR technology.

However, problems arise due to variations in layout, font, language, and script (Rabby et al. 2024). Multilingual financial document understanding requires more advanced models and datasets for layout analysis, such as BaDLAD, which contains annotated financial documents like invoices and receipts along with structural elements like tables, logos, and signatures (Shihab et al. 2023). A proper OCR system, including scanning, preprocessing operations like binarization, noise removal, skew correction, and layout analysis, feature extraction, and text recognition, ensures successful digitization (Islam et al. 2023). Advanced handwritten OCR for complex scripts like Bangla, with quality datasets and software like bbOCR, would greatly aid multilingual financial document analysis (Zulkarnain et al. 2023). Table 6 summarizes the overview of the five areas where the OCR technologies find their practical application. In each area, the practical uses of the technology, the technology types used (DLA, OCR, KIE, handwriting), the related datasets (BaDLAD, bbOCR, SROIE, CORD), and the problems faced are provided. It can be seen that the technology is widely applicable, but different fields have their own unique problems to tackle.

8.3 Healthcare

In the healthcare sector, OCR technology plays a very important role in the digitization of electronic health records, prescriptions, and other healthcare documents. This technology enables the reading of printed and handwritten documents to easily access the information (Ali et al. 2023). Handwritten prescription documents, especially those written in complex scripts such as Bengali with complex characters and marks, pose some challenges (Maung et al. 2025). In state-of-the-art OCR technology, the use of advanced models and DLAs identifies text regions, tables, images, and signatures even in damaged and faded documents. The accurate recognition of handwritten text and numbers is crucial to understand the doctor's comments, prescriptions, doses, and

Table 6 Applications of OCR and information extraction across different domains.

Domain	Use Case	Key Technologies	Datasets	Challenges
Land Records & Legal	Property deeds, legal documents, fraud analysis	DLA, OCR, handwritten recognition	BaDLAD, bbOCR	Degraded documents, complex layouts, handwritten text
Finance	Invoices, receipts, banking	KIE, layout analysis, LLM-based extraction	SROIE, CORD, BaDLAD	Layout variability, multilingual documents
Healthcare	EHRs, prescriptions, medical records	Handwritten OCR, DLA, digit recognition	Custom medical datasets	Handwriting variability, complex scripts (Bangla)
Historical & Cultural Archives	Manuscripts, rare books, heritage preservation	Deep OCR, DLA, preprocessing	BaDLAD, Gold Standard Bangla	Paper decay, ink bleed, faded text, complex typography
Government or Public Services	Forms, registration, postal codes, ID cards	OCR, DLA, signature recognition	BaDLAD, bbOCR	Mixed printed or handwritten documents, poor illumination

record numbers (Islam et al. 2023). Deep learning algorithms enhance the performance of the OCR technology in degraded historical documents by addressing stains, illumination, and faded text (Chaudhury et al. 2022). The systematic OCR process involving preprocessing, binarization, noise removal, skew correction, and layout analysis allows the accurate digitalization of documents, while datasets and open-source software like bbOCR aid in multilingual processing on a mass scale (Zulkarnain et al. 2023).

8.4 Historical and Cultural Archives

The application of OCR technology in the digitization of historical and cultural data is crucial for the preservation and improvement of the research aspect of such data (Ali et al. 2023). OCR technology helps in the scanning of historical documents, which are later transformed into an easily machine-readable form (Tuama and Mohamed 2025), (Machidon et al. 2026). The technology will help in reducing the storage space that is required for storing such documents, which is crucial in preserving such rare cultural data (Chaudhury et al. 2022). Historical documents are, however, challenging due to issues such as paper decay, fading in text, staining of text, bleeding of ink, complex writings, and handwriting styles (Tuama and Mohamed 2025), (Machidon et al. 2026). The scenario becomes worse if the documents are written in languages such as Bengali, which has compound characters, modifiers, and diacritics, especially historical documents that may have varying typography and layout designs (Maung et al. 2025).

With better OCR technology, DLA, and preprocessing, it is possible to overcome these challenges by segmenting irregular historical document pages into meaningful units (Singh and Middleton 2025). High-quality human-annotated datasets such as BaDLAD, as well as well-crafted 'gold standard' training datasets, are critical in improving the performance of deep learning-based OCR systems for low-resource languages such as Bengali (Shihab et al. 2023). The use of state-of-the-art open-source tools such as bbOCR will play an important role in the large-scale digitization of historical Bengali documents, while deep OCR algorithms are necessary for handling degraded and handwritten texts for historical data preservation and accessibility.

8.5 Government or Public Services Digitization

The digitization of government and public services has been greatly helped by OCR technology that allows printed or handwritten texts to be converted to digital information for efficient data retrieval (Rexhepi et al. 2025). For Bangladesh, the official language of documentation is Bangla, hence the need for having robust OCR systems for handling the digitization of handwritten texts for cultural preservation and public access purposes. Application of OCR technology has significantly helped improve the efficiency in the administration services of digitally processing various forms that have information regarding students, registration numbers, and postal codes (Islam et al. 2023).

Advanced optical recognition systems are therefore needed for the handling of different government official documents which may either be computer-composed, letterpress, typewriter, or handwritten (Rabby et al. 2024). With advanced technology, Bengali can be able to identify layouts, maintain original alignments, images and recognizing signatures. The complete optical recognition process of document scanning, preprocessing, feature recognition, and post-processing is essential in the digitization of official documents, especially those that are faded due to poor illumination conditions (Omee et al. 2012). Datasets such as BaDLAD have helped in developing accurate models for official documents used by governments, with open-source tools such as bbOCR being able to facilitate the digitization of documents at scale (Zulkarnain et al. 2023).

9 Challenges and Future Directions

The following sections will discuss some of the important obstacles that still hinder the efficiency and usability of OCR and IE methods, especially for Bangla and other low-resource languages. Some of the important problems that exist in the areas mentioned above include poor quality of the documents, mixed-format documents, data paucity, lack of generalizability, and bias. Some new trends such as self-supervised learning, multi-modal learning, and LLMs are also discussed below.

9.1 Document Quality and Noise Issues

The problem of sensitivity to the degradation of handwritten Bangla documents, such as low-resolution scans, uneven illumination, ink bleeding, and complex backgrounds,

is very serious, resulting in poor performance of segmentation and recognition methods. The performance of the combined YOLO and CNN pipeline for handwritten Bangla script is high, and correction and data augmentation methods were identified as key challenges in noisy conditions (Maung et al. 2025). At the same time, testing the capabilities of Tesseract and EasyOCR on the Bangla-CrossHair dataset also confirms that performance depends on the image quality and preprocessing techniques (Chowdhury et al. 2025). Research into historical documents reveals the necessity of specialized layout detection techniques and machine learning-based correction of the degraded document layout to achieve adequate recognition results (Fleischhacker et al. 2025). Additionally, advances in the area of document layout analysis and text line detection are important in order to prevent error propagation in the subsequent recognition modules (Fateh et al. 2024).

9.2 Handwritten and Printed Mixed Documents

Mixed-document contexts are difficult in the structural and semantic aspects because they include handwritten and printed text. Benchmark testing results show that a classic OCR system such as Tesseract is unable to achieve reliable recognition for two types of documents, as shown in the research (Chowdhury et al. 2025), (Nazeem et al. 2024). Systems such as Invizo for Arabic text documents show that the implementation of an integration pipeline is much better at coping with fragmentation as compared to conventional solutions (Shugaba et al. 2026). However, the accuracy of text recognition is lower for more complicated documents (Waly et al. 2025). Ocean-OCR and other systems show good results in mixed OCR processing through utilization of cross-modal representations; however, a low quantity and variety of training data negatively influence the performance of this approach (Sharma et al. 2025), (Chen et al. 2025). All these examples prove that OCR application for Bangla language requires a good layout analysis that would accomplish construction-aware segmentation and good modeling.

9.3 Low-resource and Domain-specific Languages

The challenges of Bangla handwritten OCR include limited annotated training datasets, lack of sufficient domain variation, and poor representation of ligatures and compound character classes. In the case of Bangla handwritten OCR, the hybrid framework recognizes data paucity and lack of augmentation as major research challenges, especially when dealing with unconstrained handwriting (Maung et al. 2025). Open-source OCR libraries comparative analysis reveals that low-resource languages are generally lagging behind English due to a paucity of pretraining data corpora as well as script optimization (Nazeem et al. 2024). The research into the use of vision-language transformers for multilingual low-resource Urdu OCR suggests the potential to adapt similar techniques for Bangla OCR (Cheema et al. 2024). Moreover, the unified end-to-end approach recommended in General OCR Theory provides insights about scalability of pretraining and holistic modeling in order to reach the OCR-2.0 era (Wei et al. 2024).

9.4 Generalization, Robustness, and Fairness

The major shortcoming that all OCR systems for handwritten Bangla text recognition share is a lack of proper generalization abilities. Although the Transformers provide better performance compared to other models with regard to context embeddings and sequence processing, they will still require large datasets to be trained on diverse writing styles and domains (Fujitake 2024). The disadvantages outlined in the OCR-Reasoning Benchmark support the notion that the cutting-edge multi-modal pretrained language models are unable to conduct complex reasoning procedures with images of text due to their own limitations when it comes to semantic generalization of the content being analyzed (Huang et al. 2025). Moreover, the comparative analysis of various OCR methods applied to natural scene images showed that robustness differs depending on various conditions and the complexity of the script (Prakash and Hanumanthaiah 2022). Thus, it is important to highlight the necessity of ethical evaluation and cross-script evaluation research, as well as the need to create representative datasets from various domains in order to achieve good OCR services for Bangla or any other low-resource languages (Mahadevkar et al. 2024).

9.5 Emerging Solutions

The new models will be useful in developing OCR for Bangla handwritten texts. For instance, the Ocean-OCR model employs the integration of images and language for a deeper contextual understanding and generalization of OCR (Chen et al. 2025). The use of decoder-only transformer models such as DTrOCR will also allow for better end-to-end modeling and context alignment and therefore avoid using the traditional procedures of data preprocessing, which have been developed manually (Fujitake 2024). Using self-supervised and generative methods such as TextPixs allows generating synthetic data, implying the use of diffusion conditioned on glyphs and OCR-guided supervision, which may be helpful for solving problems under the conditions of limited resources (Gillani et al. 2025). The last but not least method to increase the efficiency of OCR in Bangla manuscripts will be the use of large-scale language models for correcting the results of the OCR of historical manuscripts, which leads to considerable improvements in the transcriptions (Thomas et al. 2024). In general, the above-mentioned findings indicate ways to develop efficient solutions for Bangla OCR.

10 Conclusion

This paper provides a detailed discussion of the current state-of-the-art progress made within the areas of OCR, DLA, IE, and KIE, with a focus on deep learning-based approaches in the areas of OCR and DLA specifically in the domain of Bangla and other low-resource languages. To be more specific, the current state-of-the-art progress made within the area of OCR, with the evolution of OCR from its earliest rule-based methods to the modern transformer-based and vision-language models, is highlighted within this survey. The significance of integrating techniques from document layout analysis and text recognition is also highlighted in the survey with respect to achieving better results.

10.1 Summary of key insights

The objective of this survey was to analyze recent trends in the field of document analysis and processing in relation to the following sub-topics: OCR, DLA, IE, and KIE in the context of Bangla and other low-resource languages using deep learning-based solutions. The findings of this survey show that the field of document processing and analysis has experienced a paradigm shift from the traditional handwriting feature engineering-based methods to the advanced deep learning and vision-language techniques, including layout-aware language models and LLMs. Advanced techniques for document image recognition are more efficient at capturing contextual information, which in turn contributes to higher accuracy in both layout analysis and information extraction.

Furthermore, this survey reveals that modern-day document analysis systems tend to adopt an approach combining OCR and IE systems rather than relying solely on OCR in order to achieve better results in complex document layout while preserving reading order and extracting relevant information systematically. Nonetheless, several barriers have been identified that hamper the development of research in the domain of document processing and analysis, especially in low-resource languages like Bangla, including the non-existence of standardized annotated datasets, lack of benchmarking standards, high computational power requirements, and poor generalization capabilities. Lastly, this survey concludes by suggesting some research directions in order to further the research in document analysis.

10.2 Long-term vision for OCR and IE in low-resource languages

The future of OCR and IE is associated with creating systems that can extract information from images of documents with a minimal amount of human intervention and that can understand the structure and the content of the document at the level of an average person reading the document. To solve the problem with low-resource languages like Bangla, it is important to get more annotated data and create a benchmark. Moreover, it is necessary to create a multilingual foundation model capable of performing cross-lingual transfer. Thus, the future of research needs to be associated with developing an end-to-end system that recognizes text, analyzes the structure, and extracts information using vision-based language models and large-scale language models. Such a model is very necessary to digitize historical documents, legal entities, healthcare, financial, and public services. At the same time, it is extremely important that low-level languages get the same opportunities to benefit from new technologies as major languages.

Nowadays, OCR and IE are recognized as actively developing and researched fields of science. It is worth mentioning that recognition systems for text in the document do not stop developing, whereas the solution for languages with low resource is still under research. The future of OCR and IE will definitely bring more available and efficient systems thanks to the efforts of academia, industry, and government. In the future, OCR will be able not only to digitize texts but also preserve the world's cultural

heritage, optimize business processes, improve the quality of public services, and make people digitally literate.

11 Funding

No financial assistance was provided to the authors for conducting the study or publishing this article.

References

Patel, S., Bhatt, D.: Abstractive information extraction from scanned invoices (aiesi) using end-to-end sequential approach. arXiv preprint arXiv:2009.05728 (2020)

Singh, P.: Deep learning-based text detection and recognition in document images: A comprehensive review. International Journal of Creative Research Thoughts (IJCRT) **13**(4), 369–377 (2025)

Yang, Y., *et al.*: A survey of information extraction based on deep learning. Applied Sciences **12**(19), 9691 (2022)

Huang, K., Mo, F., Zhang, X., Li, H., Li, Y., Zhang, Y., Yi, W., Mao, Y., Liu, J., Xu, Y., et al.: A survey on large language models with multilingualism: Recent advances and new frontiers. Artificial Intelligence Review (2026)

Hasan, M.S., *et al.*: Enhancing bangla language text recognition in ocr using levenshtein distance and character grouping strategy. In: 2024 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON), pp. 1–6 (2024). IEEE

Ding, Y., Han, S.C., Lee, J., Hovy, E.: Deep learning based visually rich document content understanding: A survey. Artificial Intelligence Review **59**(4), 114 (2026)

Yu, W., *et al.*: Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4363–4370 (2021). IEEE

Prakash, P., Hanumanthaiah, S.K.Y.: A comparative analysis of optical character recognition models for extracting and classifying texts in natural scenes. Int J Artif Intell ISSN **2252**(8938), 1291 (2022)

Henge, S.K., Rama, B.: Comprative study with analysis of ocr algorithms and invention analysis of character recognition approched methodologies. In: 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp. 1–6 (2016). IEEE

Ali, H., Rabby, A.S.A., Islam, M.M., Mahamud, A., Hasan, N., Rahman, F.: Gold standard bangla ocr dataset: An in-depth look at data preprocessing and annotation

processes. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 460–470 (2023)

Ding, Y., Luo, S., Dai, Y., Jiang, Y., Li, Z., Sun, Q., Martin, G., Liu, W., Peng, Y.: A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. Findings of the Association for Computational Linguistics: ACL 2026, 13319–13340 (2026)

Favish, A.: Data capture automation in the south african deeds registry using optical character recognition (ocr) (2019)

Nida, B.R.: Smart automation: Elevating expense management experience with the power of optical character recognition (ocr) technology

Rahman, M.A., Tabassum, N., Paul, M., Pal, R., Islam, M.K.: Bn-htrd: A benchmark dataset for document level offline bangla handwritten text recognition (htr) and line segmentation. In: Computer Vision and Image Analysis for Industry 4.0, pp. 1–16. Chapman and Hall/CRC, ??? (2023)

Sharma, V., Nazim, M., Prakash Sharma, S.: A systematic literature review on handwritten text recognition: Approaches, challenges, and future directions. In: Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024) (2025)

Naquib, F.B.: Optimizing ocr performance through black-box approximation on full-resolution text images (2025)

Omee, F.Y., Himel, S.S., Bikas, M.A.N.: A complete workflow for development of bangla ocr. arXiv preprint arXiv:1204.1198 (2012)

Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-ocr processing approaches. ACM Computing Surveys (CSUR) **54**(6), 1–37 (2021)

Maung, A.T., Salekin, S., Haque, M.A.: A hybrid approach to bangla handwritten ocr: combining yolo and an advanced cnn. Discover Artificial Intelligence **5**(1), 119 (2025)

Sakib, A.M., *et al.*: Medical text extraction and classification from prescription images. In: 2022 25th International Conference on Computer and Information Technology (ICCIT), pp. 472–477 (2022). IEEE

Chen, S., et al.: Ocean-ocr: Towards general ocr application via a vision-language model. arXiv preprint arXiv:2501.15558 (2025)

Noor, N..A.: Bangla optical character recognition. Master's thesis, BRAC University (2005). Thesis on neural network approaches for Bangla OCR

Chowdhury, A.N., *et al.*: Performance analysis of tesseract and easyocr for bangla

optical character recognition on the novel bangla crosshair dataset. In: 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), pp. 716–721 (2025). IEEE

Maliha, F., Nil, S.G., Tabassum, S., Rahman, M.A., Rahman, A., Habib, M.T.: A comprehensive study of bengali language tools and applications for artificial intelligence research. Engineering Reports **8**(6), 70853 (2026)

Bhattacharya, S., Maitra, D.S., Bhattacharya, U., Parui, S.K.: An end-to-end system for bangla online handwriting recognition. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 373–378 (2016). IEEE

Jabde, M., Patil, C.H., Vibhute, A.D., Saini, J.R.: A systematic review of multilingual numeral recognition systems. Artificial Intelligence Review **58**(4), 106 (2025)

Pal, A., Hasan, M.S., Ahsan, S.M.M.: Improving character recognition in bangla handwritten words: a two-stage single shot detector approach. In: 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), pp. 1–6 (2024). IEEE

Rombach, A.M., Fettke, P.: Deep learning based key information extraction from business documents: Systematic literature review. ACM Computing Surveys **58**(2), 1–37 (2025)

Wang, X.: A review of text recognition in complex scene images based on deep learning. In: 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025), pp. 776–783 (2025). Atlantis Press

Hasan, S.R., *et al.*: Optical text recognition in nepali and bengali: A transformer-based approach. In: 2023 IEEE Fifth International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 01–06 (2023). IEEE

Chung, Y.H.M., Choi, D.: Finetuning vision-language models as ocr systems for low-resource languages: A case study of manchu. arXiv preprint arXiv:2507.06761 (2025)

Mahadevkar, S.V., Patil, S., Kotecha, K., Soong, L.W., Choudhury, T.: Exploring ai-driven approaches for unstructured document analysis and future horizons. Journal of Big Data **11**(1), 92 (2024)

Renda, H.E.E.: Implementation of deep learning models for information extraction on identification documents. Master's thesis, Universidade NOVA de Lisboa (Portugal) (2023)

Qi, H., Yang, H., Wang, Z., Ye, J., Xin, Q., Zhang, C., Lang, Q.: Ancientglyphnet: an advanced deep learning framework for detecting ancient chinese characters in complex scene. Artificial Intelligence Review **58**(3), 88 (2025)

Ly, N.H., *et al.*: Vietnamese receipt information extraction using ocr and deep learning: A hybrid approach with fuzzy c-means and phobertv2. In: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, pp. 296–307 (2025). Springer

Atagong, S.D., Tonnang, H., Senagi, K., Wamalwa, M., Agboka, K.M., Odindi, J.: A review on knowledge and information extraction from pdf documents and storage approaches. Frontiers in Artificial Intelligence **8**, 1466092 (2025)

Cao, P., Wang, Y., Zhang, Q., Meng, Z.: Genkie: Robust generative multimodal document key information extraction. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 14702–14713 (2023)

Han, Y., Ding, H., Cui, X., Zhang, X., Wilson, D.I., Yu, W.: Breaking through language barriers: A review of ocr technology for low-resource minority languages based on deep learning. Available at SSRN 5467297 (2025)

Tuama, B.A., Mohamed, F.: A systematic literature review of deep learning methods for handwritten text recognition in historical arabic manuscripts. Engineering, Technology & Applied Science Research **15**(4), 25772–25782 (2025)

Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: Kleister: key information extraction datasets involving long documents with complex layouts. In: International Conference on Document Analysis and Recognition, pp. 564–579 (2021). Springer

Antonio, J., Putra, A.R., Abdurrohman, H., Tsalasa, M.S.: A survey on scanned receipts ocr and information extraction. In: Proceedings of the International Conference on Document Analysis and Recognit, Jerusalem, Israel, pp. 29–30 (2022)

Tewani, R., *et al.*: An efficient extraction of information from indian government issued documents aadhar and pan card. Fusion: Practice and Applications **4**(2), 56–61 (2021)

Hossain, M.J., Zaman, S.S., Akash, F.R., Sarker, M.R., Huq, M.R.: Developing a bangla handwritten text recognition framework using deep learning. In: 2023 26th International Conference on Computer and Information Technology (ICCIT), pp. 1–6 (2023). IEEE

Mazumder, T., Nusrat, F., Mahi, A.B.S., Brishty, J.B., Rahman, R., Helaly, T.: Automated and efficient bangla signboard detection, text extraction, and novel categorization method for underrepresented languages in smart cities. Results in Engineering **26**, 105156 (2025)

Saha, C., Rahman, M.M.: Banglanet: Bangla handwritten character recognition using ensembling of convolutional neural network. arXiv preprint arXiv:2401.08035 (2024)

Islam, M., Shuvo, S.A., Nipun, M.S., Bin Sulaiman, R., Shaikh, M., Nayeem, J., Haque, Z., Sourav, M.S.U., Kareem, A.: Efficient approach to using cnn-based pre-trained models in bangla handwritten digit recognition. In: Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2022, pp. 697–716. Springer, ??? (2023)

Rabby, A., Ali, H., Islam, M.M., Abujar, S., Rahman, F.: Enhancement of bengali ocr by specialized models and advanced techniques for diverse document types. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1102–1109 (2024)

Chaudhury, A., Mukherjee, P.S., Das, S., Biswas, C., Bhattacharya, U.: A deep ocr for degraded bangla documents. Transactions on Asian and Low-Resource Language Information Processing **21**(5), 1–20 (2022)

Islam, M.M., Das, A., Kowsar, I., Rabby, A.S.A., Hasan, N., Rahman, F.: Towards building a bangla text recognition solution with a multi-headed cnn architecture. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 1061–1067 (2021). IEEE

Garrido-Munoz, C., Rios-Vila, A., Calvo-Zaragoza, J.: Handwritten text recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

Amini, A., Firouzkouhi, N., Alotaibi, S., Ali, O., Rashidi, M., Gholami, A., Zheng, Q., Zabalawi, I., Cheng, C., Davvaz, B.: Toward a unified governing framework for modern artificial intelligence: a taxonomy-based cross-domain survey. Artificial Intelligence Review (2026)

Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633 (2007). IEEE

Bhattacharya, U., Parui, S., Chaudhuri, B.: Recognition of bangla handwriting: Current status and future directions. Advances in Digital Document Processing and Retrieval, 287–321 (2014)

Inunganbi, S.: A systematic review on handwritten document analysis and recognition. Multimedia Tools and Applications **83**(2), 5387–5413 (2024)

Parui, S.K., Guin, K., Bhattacharya, U., Chaudhuri, B.B.: Online handwritten bangla character recognition using hmm. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008). Ieee

Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)

Breuel, T.M., Ul-Hasan, A., Al-Azawi, M.A., Shafait, F.: High-performance ocr for printed english and fraktur using lstm networks. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 683–687 (2013). IEEE

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192–1200 (2020)

Qian, Y., Jia, C., Liu, Y.: Bert-based text keyword extraction. In: Journal of Physics: Conference Series, vol. 1992, p. 042077 (2021). IOP Publishing

Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: Cmaterdb1: a database of unconstrained handwritten bangla and bangla–english mixed script document image. International Journal on Document Analysis and Recognition (IJDAR) **15**(1), 71–83 (2012)

Shugaba, F.A., Sheikh, U.U., Othman, M.A., Khamis, N., Abdulfattah, M.H.: Handwritten arabic text recognition: A review of recent trends and future direction. International Journal of Computer Vision **134**(5), 207 (2026)

Mondal, A., Mahadevan, V., Manmatha, R., Jawahar, C.: Icdar 2024 competition on recognition and vqa on handwritten documents. In: International Conference on Document Analysis and Recognition, pp. 426–442 (2024). Springer

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 13094–13102 (2023)

Gu, Z., Meng, C., Wang, K., Lan, J., Wang, W., Gu, M., Zhang, L.: Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4583–4592 (2022)

Abdallah, A., Eberharter, D., Pfister, Z., Jatowt, A.: Transformers and language models in form understanding: A comprehensive review of scanned document analysis. arXiv preprint arXiv:2403.04080 (2024)

Subramani, N., Matton, A., Greaves, M., Lam, A.: A survey of deep learning approaches for ocr and document understanding (2021). URL https://arxiv. org/abs (2011)

Rabaev, I., Litvak, M.: Recent advances in text line segmentation and baseline detection in historical document images: a systematic review: I. rabaev, m. litvak. International Journal on Document Analysis and Recognition (IJDAR), 1–37 (2025)

Townsend, B., May, M., Mackowiak, K., Wells, C.: Realkie: Five novel datasets for enterprise key information extraction. arXiv preprint arXiv:2403.20101 (2024)

Singh, L.G., Middleton, S.E.: Tabular context-aware optical character recognition and tabular data reconstruction for historical records: Lg singh, se middleton. International Journal on Document Analysis and Recognition (IJDAR) **28**(3), 357–376 (2025)

Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., *et al.*: Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21744–21754 (2025)

Murad, H., Ali, M.E.: Towards detecting, recognizing, and parsing the address information from bangla signboard: A deep learning-based approach. arXiv preprint arXiv:2311.13222 (2023)

Constum, T., Tranouez, P., Paquet, T.: Daniel: a fast document attention network for information extraction and labelling of handwritten documents: T. constum et al. International Journal on Document Analysis and Recognition (IJDAR) **28**(4), 573–595 (2025)

Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the nlp world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293 (2020)

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J.T., Abbott, J., Marivate, V., Kabongo, S., et al.: Masakhane–machine translation for africa. arXiv preprint arXiv:2003.11529 (2020)

Mridha, M.F., Ohi, A.Q., Ali, M.A., Emon, M.I., Kabir, M.M.: Banglawriting: A multi-purpose offline bangla handwriting dataset. Data in Brief **34**, 106633 (2021)

Shihab, M.I.H., *et al.*: Badlad: A large multi-domain bengali document layout analysis dataset. In: International Conference on Document Analysis and Recognition, pp. 326–341 (2023). Springer

Zulkarnain, I.M., Islam, S.B., Farabe, M.Z.A.Z., Shawon, M.M.H., Abedin, J.M., Hasan, B.R., Haque, M., Shihab, I., Mobassir, S., Ansary, M., et al.: bbocr: An open-source multi-domain ocr pipeline for bengali documents. arXiv preprint arXiv:2308.10647 (2023)

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M., Bali, K.: Unsung challenges of building and deploying language technologies for low resource language communities. In: Proceedings of the 16th International Conference on Natural Language Processing, pp. 211–219 (2019)

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498 (2021)

Costa-Jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)

Taghva, K., Borsack, J., Condit, A.: Evaluation of model-based retrieval effectiveness with ocr text. Technical report, Citeseer (1994)

Yang, Y., Eli, E., Aysa, A., Ubul, K.: Script identification in multilingual environment: A survey in recent years. Artificial Intelligence Review **58**(10), 294 (2025)

Dong, Y., Cordonnier, J.-B., Loukas, A.: Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: International Conference on Machine Learning, pp. 2793–2803 (2021). PMLR

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need, in advances in neural information processing systems": Annual conference on neural information processing systems 2017 (2017)

Li, D.-L., Lee, S.-K., Liu, Y.-T.: Printed document layout analysis and optical character recognition system based on deep learning. Scientific Reports **15**(1), 23761 (2025)

Niloy, D.B., Mahamud, S.A., Emon, I.H., Islam, A., Jony, M.E.H.: Comparative analysis of cnn and rnn models for bangla hand written character recognition. In: 2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), pp. 1–6 (2025). IEEE

Opu, M.N.I., Hossain, M.E., Kabir, M.A.: Handwritten bangla character recognition using convolutional neural networks: a comparative study and new lightweight model. Neural Computing and Applications **36**(1), 337–348 (2024)

Monsur, S.M., Kabir, S., Chowdhury, S.: Synthnid: Synthetic data to improve end-to-end bangla document key information extraction. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pp. 117–123 (2023)

Dergachov, K., Krasnov, L., Bilozerskyi, V., Zymovin, A.: Data pre-processing to increase the quality of optical text recognition systems. Radioelectronic and computer systems (4), 183–198 (2021)

Islam, R., Islam, M.R., Talukder, K.H.: Extraction and recognition of bangla texts from natural scene images using cnn. In: International Conference on Image and Signal Processing, pp. 243–253 (2020). Springer

Karpinski, R., Lohani, D., Belaid, A.: Metrics for complete evaluation of ocr performance. In: IPCV'18-The 22nd Int'l Conf on Image Processing, Computer Vision, & Pattern Recognition (2018)

Wang, W., Hu, H., Zhang, Z., Li, Z., Shao, H., Dahlmeier, D.: Document intelligence in the era of large language models: A survey. arXiv preprint arXiv:2510.13366 (2025)

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270 (2016)

Hossain, M.I., Rakib, M., Mollah, S., Rahman, F., Mohammed, N.: Lila-boti: Leveraging isolated letter accumulations by ordering teacher insights for bangla handwriting recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1770–1776 (2022). IEEE

Rexhepi, A., Hasi, E., Haxholli, A., Bytyçi, E.: Invoice and receipt optical character recognition: review on current methods and future trends (2025)

Machidon, O.M., Rejec, J., Vreš, D., Machidon, A.L.: Large language models for ocr in cultural heritage: a comparative study on slovene folkloristic texts: Om machidon et al. International Journal on Document Analysis and Recognition (IJDAR), 1–15 (2026)

Fleischhacker, D., Kern, R., Göderle, W.: Enhancing ocr in historical documents with complex layouts through machine learning. International Journal on Digital Libraries **26**(1), 3 (2025)

Fateh, A., Fateh, M., Abolghasemi, V.: Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. Engineering Reports **6**(9), 12832 (2024)

Nazeem, M., Anitha, R., Navaneeth, S., *et al.*: Open-source ocr libraries: A comprehensive study for low resource language. In: Proceedings of the 21st International Conference on Natural Language Processing (ICON), pp. 416–421 (2024)

Waly, A., Tarek, B., Feteha, A., Yehia, R., Amr, G., Gomaa, W., Fares, A.: Invizo: Arabic handwritten document optical character recognition solution. arXiv preprint arXiv:2502.05277 (2025)

Cheema, M.D.A., Shaiq, M.D., Mirza, F., Kamal, A., Naeem, M.A.: Adapting multilingual vision language transformers for low-resource urdu optical character recognition

(ocr). PeerJ Computer Science **10**, 1964 (2024)

Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)

Fujitake, M.: Dtrocr: Decoder-only transformer for optical character recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 8025–8035 (2024)

Huang, M., Shi, Y., Peng, D., Lai, S., Xie, Z., Jin, L.: Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. arXiv preprint arXiv:2505.17163 (2025)

Gillani, S.A., Baig, M.S.A., Khan, O.A., Shah, S.M., Mujeeb, U., Ali, M.: Textpixs: Glyph-conditioned diffusion with character-aware attention and ocr-guided supervision. arXiv preprint arXiv:2507.06033 (2025)

Thomas, A., Gaizauskas, R., Lu, H.: Leveraging llms for post-ocr correction of historical newspapers. In: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024, pp. 116–121 (2024)