Journal of Big Data

SURVEY Open Access

Exploring AI-driven approaches for unstructured document analysis and future horizons

Supriya V. Mahadevkar[1], Shruti Patil[2*], Ketan Kotecha[2], Lim Way Soong[3*] and Tanupriya Choudhury[4,5]

*Correspondence:
shruti.patil@sitpune.edu.in;
wslim@mmu.edu.my

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology Symbiosis International (Deemed University), Pune 412115, India
[3] Faculty of Engineering and Technology, Multimedia University, Cyberjaya, Malaysia
[4] School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand 248002, India
[5] CSE Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India

Abstract

In the current industrial landscape, a significant number of sectors are grappling with the challenges posed by unstructured data, which incurs financial losses amounting to millions annually. If harnessed effectively, this data has the potential to substantially boost operational efficiency. Traditional methods for extracting information have their limitations; however, solutions powered by artificial intelligence (AI) could provide a more fitting alternative. There is an evident gap in scholarly research concerning a comprehensive evaluation of AI-driven techniques for the extraction of information from unstructured content. This systematic literature review aims to identify, assess, and deliberate on prospective research directions within the field of unstructured document information extraction. It has been observed that prevailing extraction methods primarily depend on static patterns or rules, often proving inadequate when faced with complex document structures typically encountered in real-world scenarios, such as medical records. Datasets currently available to the public suffer from low quality and are tailored for specific tasks only. This underscores an urgent need for developing new datasets that accurately reflect complex issues encountered in practical settings. The review reveals that AI-based techniques show promise in autonomously extracting information from diverse unstructured documents, encompassing both printed and handwritten text. Challenges arise, however, when dealing with varied document layouts. Proposing a framework through hybrid AI-based approaches, this review envisions processing a high-quality dataset for automatic information extraction from unstructured documents. Additionally, it emphasizes the importance of collaborative efforts between organizations and researchers to address the diverse challenges associated with unstructured data analysis.

Keywords: Artificial intelligence, Unstructured document processing, Printed and handwritten text recognition, Information extraction, Optical character recognition, Semantic segmentation, Robotics process automation, Named entity recognition, Large Language models

Introduction

Every day, a massive amount of unstructured and semi-structured data is created, but it remains unanalyzed. According to IDC, the total amount of data in the universe will expand from 33 zettabytes this year to 175 zettabytes by 2025, representing a 61 percent

compound annual growth rate. Many banks, government offices, and agencies collect information and inquiries from customers using handwritten documents [1]. Almost every business or office today employs a digital database in combination with their websites to store customer information and other details. As a result, every time a form is filled, staff must manually type the data into the database. This is inefficient and wastes a lot of time and effort. Manual data entry can lead to a few mistakes that go unreported and end up in the database [2]. Many of the offline information or request forms provided by governmental organizations and other businesses are filled out in text boxes. The use of a keyboard is required to manually insert the data from these manually filled-out forms into the machine database. Handwritten character recognition from filled forms is challenging since the style of writing, curve, size, strokes, and thickness of character vary from person to person.

Many companies still use paper-based working practices. Formal reports can always be made with written or printed records, which is why. The difficulty of tracking lost data, storage, and the time and money lost on re-keying data are the main problems in processing unstructured documents. The numerous application domains for unstructured document data extraction include commercial forms, tabular forms, government records, historical documents, engineering reports, postal documents, and bank documents, among others [3].

Automation makes it easier for businesses to access and manage relevant information. The digitally stored unstructured data could be automated to give enterprises a competitive advantage, boost productivity, and swiftly obtain an understanding of their businesses and develop new ideas. Thus, by realizing the significance of AI-based technologies like Computer Vision (CV) and Natural Language Processing(NLP), enterprises adapt to automation solutions. Text, pictures, and scanned documents are examples of unstructured data that AI technology can grasp and categorize better than conventional information extraction techniques. An AI-based framework for processing unstructured documents is needed because of the growing volume and the requirement to use unstructured data effectively. This framework would enable enterprises to automatically gain insights from unstructured data. Figure 1 shows why unstructured document processing is required. These issues can be solved with a system that digitizes the data contained in these paper-based unstructured documents. So there is a need to develop a framework that focuses on these existing difficulties of unstructured document processing utilizing AI.

To extract accurate and helpful information from a large unstructured document while removing extraneous or insignificant information, summarization of important information is required based framework that makes it possible to extract relevant information from unstructured documents quickly and automatically [4]. AI also aids in the creation of a more intelligible analysis of unstructured materials that businesses can employ in their important decision-making processes. Figure 2 shows the overall organization of this systematic literature review paper.

A. Significance

Unstructured data is essential to many businesses. Forms like invoices, client information, and insurance claims are used by a company as evidence and records of

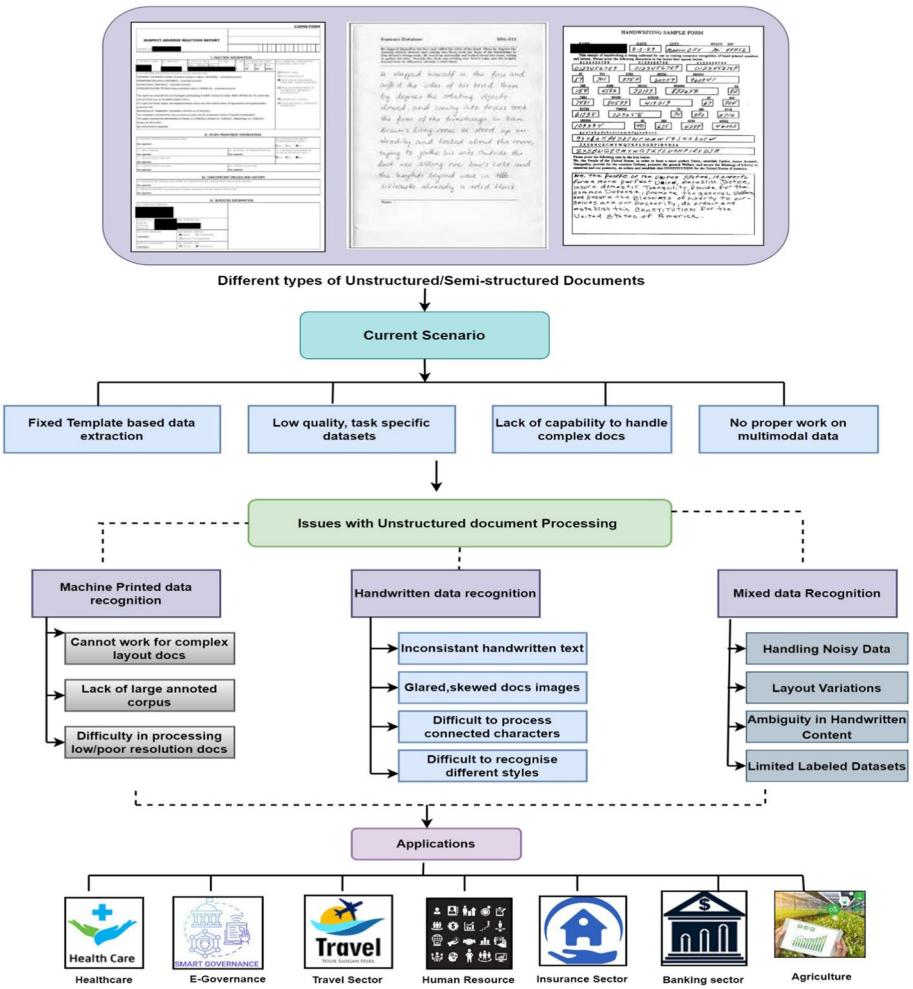

Fig. 1 Why unstructured document processing?

transactions and other important occurrences. It is crucial to process these forms properly. Manual processing may cause errors and delays, as well as possibly slow down the procedure. These problems are addressed by automatic data extraction, which provides an automated and digital solution to document processing. The monotonous and time-consuming processes can be automated to increase production and the expansion of the business. AI makes it possible to quickly and automatically extract relevant data from unstructured materials. Additionally, AI assists in producing an analysis of unstructured material that is easier to grasp and use by enterprises in key decision-making stages.

B. Motivation

A systematic literature review that focuses on information extraction strategies and examines their clear advantages, disadvantages, taxonomies, and comparative analysis is required. A thorough analysis of publicly accessible datasets and techniques for data preprocessing and data classification is lacking in the body of existing literature. Furthermore, the literature does not provide a comprehensive analysis of methods for automatically extracting information from unstructured materials.

- The major objective of this Systematic Literature Review (SLR) is to draw attention to the information that is currently available regarding The constraints of the existing information extraction techniques for processing unstructured documents.
- To show existing datasets that may be used to extract information from unstructured materials.
- To describe the techniques used for preprocessing, feature extraction, classification, and data preparation are used to assess the data's quality.
- To showcase different online information extraction tools and frameworks.
- Provide insight into the comparative examination of several AI-based unstructured document processing techniques. To help researchers working in this area create effective information extraction methods for unstructured documents, the presented SLR aims to provide insights into unstructured document processing in detail.
- It also focuses on the future directions in this domain which will be helpful for the researchers.

C. Evolution of techniques used for mixed text data processing

The approaches for automatically extracting information from unstructured documents, including machine-printed and handwritten text documents, have evolved throughout time, as shown in Fig. 3. It is still considered to be a difficult problem to solve to recognize handwriting (Handwriting OCR) or handwritten text (HTR) [5]. Making handwritten text machine-readable is difficult because of the wide range of handwriting styles across individuals and the bad quality of handwritten text compared to printed text. During the 1960s and 1970s, early research on handwriting recognition was done. Less research was done on handwriting recognition in the 1980s as a result of the unsatisfactory outcomes these systems were achieving at the time. Later in the 1990s, more effective handwriting recognition algorithms were developed, including Hidden Markov Models (HMM), Dynamic Programming Matching, Neural Networks (NN), etc. From the year 2000 onward, lexicons, dictionaries, and language models were implemented, in addition to the combination or collaboration of several independent recognizers. After that in 2010, RPA-based software robots were implemented in the recognition of handwritten and printed text recognition. In the year 2020 onwards the use of AI-based OCR evolved with the help of machine learning and deep learning techniques in this area.

In the past, in the case of machine-printed documents, enterprises used manual labor to enter data, process paper documents, and provide the necessary data to the subsequent business processing chain. The businesses began utilizing optical character recognition(OCR) as the initial step toward digitization. The scanned document's contents were converted into a digital format using OCR. The early iterations of OCR are constrained to recognizing a single typeface at a time and demand the provision of an image for each letter. It was suggested in the early 2000s to use "Omni-font OCR" to process text printed in nearly any typeface. Then it developed into a cloud-based service that was compatible with PC and mobile apps. Thanks to the numerous OCR service providers who supply OCR technology via APIs, almost all characters and scripts may now be identified with a reasonable accuracy rate. Businesses started using robotic process automation (RPA) to replace monotonous, rule-based, structured tasks with software bots as

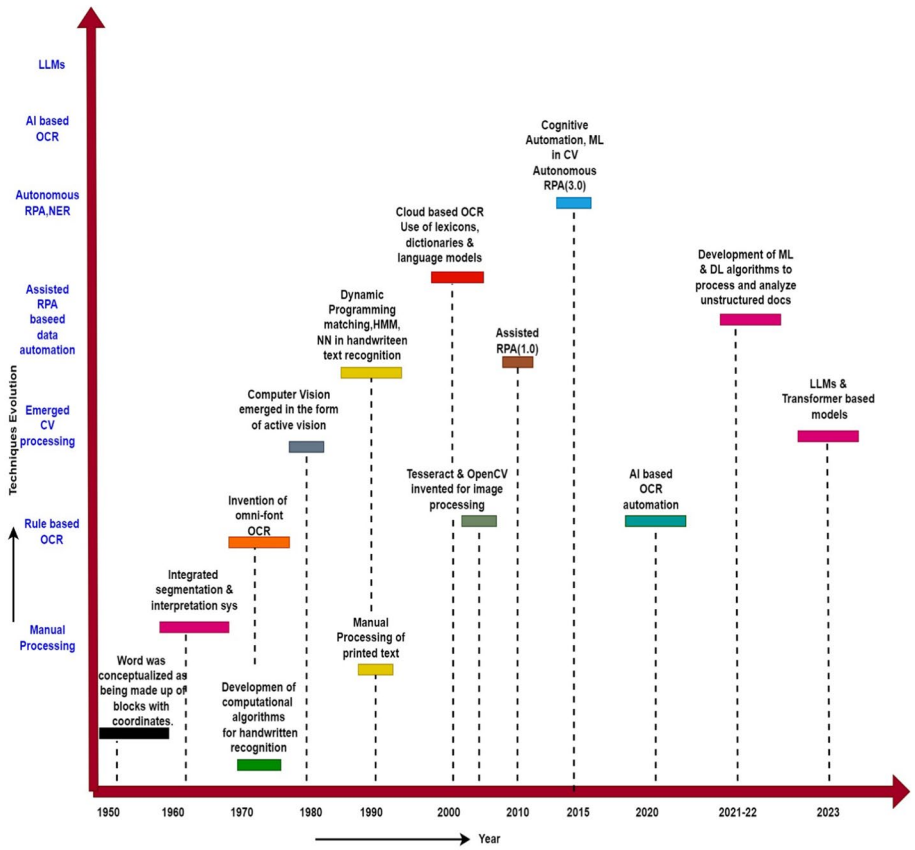

Fig. 3 Evolution of mixed text (printed and handwritten) based document processing

automation advanced. The majority of automated procedures consist of rule-based logic. It is logical software that follows predetermined rules to automatically do regular activities. From RPA versions 1.0, 2.0, through 3.0, RPA has developed [3]. To evaluate the significance of both structured and unstructured data, AI-based automation technology uses a range of promising techniques, such as Computer Vision, NLP, Machine Learning (ML), Deep Learning (DL), and Text Analytics. Unstructured data may now be sorted, processed, and analyzed to provide meaningful insights with minimal human work and involvement thanks to improvements in AI.

D. Related work

The availability of SLR in this area of study is quite limited. In the contemporary realm of Machine Learning and Computer Vision, the challenge of handwriting recognition frequently emerges as a significant concern. For those delving into these fields, the appeal to tackle such issues has intensified owing to both the inefficiency and tedious nature inherent in traditional manual transcription methods. A diverse range of methodologies and strategies have been explored by scholars and experts to surmount this obstacle. However, despite their endeavors, a conclusive solution has eluded the academic community thus far. While printed materials have successfully transitioned into digital formats, handwritten texts continue to pose distinctive difficulties that remain unresolved.

Identifying handwritten materials, such as forms, presents unforeseen difficulties due to handwriting style, readability, and some intrinsic noise, such as printed boundary boxes on form papers. This increases the handwriting recognition issues complexity. Characters with very similar forms, unexpected distortions that are distinctive of particular handwriting styles, and thickness differences among written characters as a result of the use of various writing materials and tools are just a few examples of common variances. Even using several scanners changes the resolution of the images used to train the models [6]. Very little research work done in the area of mixed text data recognition.

The research [3] is one of the most recent and important SLRs that offers a fair overview of text analytics for processing unstructured data in machine-printed invoice data recognition. The authors discussed different approaches used for printed types of unstructured document processing, various application domains, available dataset challenges and future directions. However, this SLR lacks in the discussion about the handwritten type of unstructured document processing approaches and its challenges. By providing insightful information from unstructured document processing for the banking industry, the study enhanced the literature on the subject. The writers covered the use of text analytics for client onboarding, forecasting market changes, preventing fraud, enhancing operational processes, and creating novel business models. The authors emphasized the importance of outside and internal unstructured data sources for text analytics employed in the financial industries. Data from log files, transactions, and applications are all contained inside data sources. Any data from social media and website data are examples of outside data sources. The survey also offers helpful text analytics techniques including topic extraction, keyword extraction, topic extraction, sentiment analysis, and named entity recognition [7]. The survey does not, however, include a thorough review of the techniques currently in use for automating the extraction of data from unstructured texts.

The limitations of various unstructured data forms, such as images, text, audio, and video, are covered in the surveys [1, 8]. The authors have noted the difficulties associated with representing and converting unstructured data into structured data, as well as the exponential volume increase and diversity of the data kinds.

A lot of existing RPA-related challenges, themes, and difficulties are identified by this structured literature review for further investigation [9]. The study focuses on how RPA has experienced widespread adoption in businesses looking to boost operational productivity. By using less human effort in typical business procedures and enhancing job quality, the authors emphasized how RPA might increase organizational performance and save costs. The survey also included several RPA suppliers who sell RPA products. However, the use of AI-based RPA techniques for handling unstructured data is not included in the paper. The role of RPA and the requirements for RPA implementation are highlighted in this SLR. The manuscript [10] emphasizes current improvements made using knowledge graphs in named entity disambiguation (NED), named entity recognition (NER), and named entity linking (NEL) (KG). The nodes in this network are represented by named entities, and the edges demonstrate their semantic relationships. However, the NER from the unstructured materials was not the author's main concern.The survey [11] focuses on apps that leverage electronic health records(EHR) to derive clinical information.

The study lists a few frameworks for extracting information from EHRs, including Unstructured Information Management Architecture (UIMA), General Architecture for Text Engineering (GATE), and Medical Language Extraction and Encoding (MedLEE). A mention of AI-based information extraction methods is missing from the survey.

Convolutional neural networks(CNN) are one of the most used Deep Learning Architectures for complicated image-processing tasks. These methods work by automatically extracting a wide range of features, from straightforward features like edges and curves to more intricate features like textures. CNN is therefore an extensively used technique for handwriting recognition. Researchers [12] used a modified CNN to get an accuracy rating of 99.59% on the MNIST dataset. To solve the issue of handwriting recognition on form papers, Darmatasia and Fanany [13] suggested combining CNN with Support Vector Machines (SVM). The authors first extracted the features from handwritten forms using CNN, and then they transferred that information to SVM so that the characteristics could be divided into alphabets and words. The accuracy percentages provided by the authors were 98.85% for numerical characters, 93.05% for uppercase, 86.21% for lowercase, and 91.37% for the fusion of number and uppercase characters. According to reports, this is a step forward from the currently used standalone CNN-based approaches.

The field of unstructured document processing for handwritten and printed text recognition has seen significant advancements, driven by the integration of advanced deep learning techniques and Large Language Models (LLMs). Attention mechanisms allow models to focus on relevant parts of the input, significantly improving recognition accuracy, especially in cluttered or noisy images.Transfer learning involves fine-tuning models pre-trained on large datasets (like ImageNet for CNNs or large text corpora for LLMs) to specific OCR tasks, reducing training time and improving performance.Multimodal models that integrate visual features from images with textual information to improve recognition accuracy and contextual understanding.Training models to understand and leverage information across different modalities (e.g., text and images) for more robust OCR systems. The integration of LLMs further pushes the boundaries by adding powerful contextual understanding and error correction capabilities, leading to more intelligent and reliable document processing solutions.

Highlights of a research work shortcomings that may be summarized as follows:

1. Existing reviews or surveys are task or domain-specific.
2. The existing corpus of research does not address the generalizability of information extraction algorithms to deal with different layouts or formats of unstructured documents. Multiple layouts refer to papers that have a flexible structure and can be written in different formats.
3. There is no existing SLR discussion about the multimodality of unstructured document data such as printed and handwritten document data images.
4. A very small number of researchers have examined the frameworks or tools that are available for the automated extraction of data from unstructured documents.

E. Research goals

This systematic literature review aims to locate and assess recent studies' findings in connection with the topic of attention. Overall research paper's contents organization is displayed in Fig. 2. The research questions that were developed to help this SLR study become more focused are listed in Table 1 below. The research goal is to comprehensively analyze existing literature on unstructured document processing with AI, aiming to identify key methodologies, challenges, and advancements. This systematic review seeks to provide a consolidated understanding of the current state of the field and potential directions for future research.

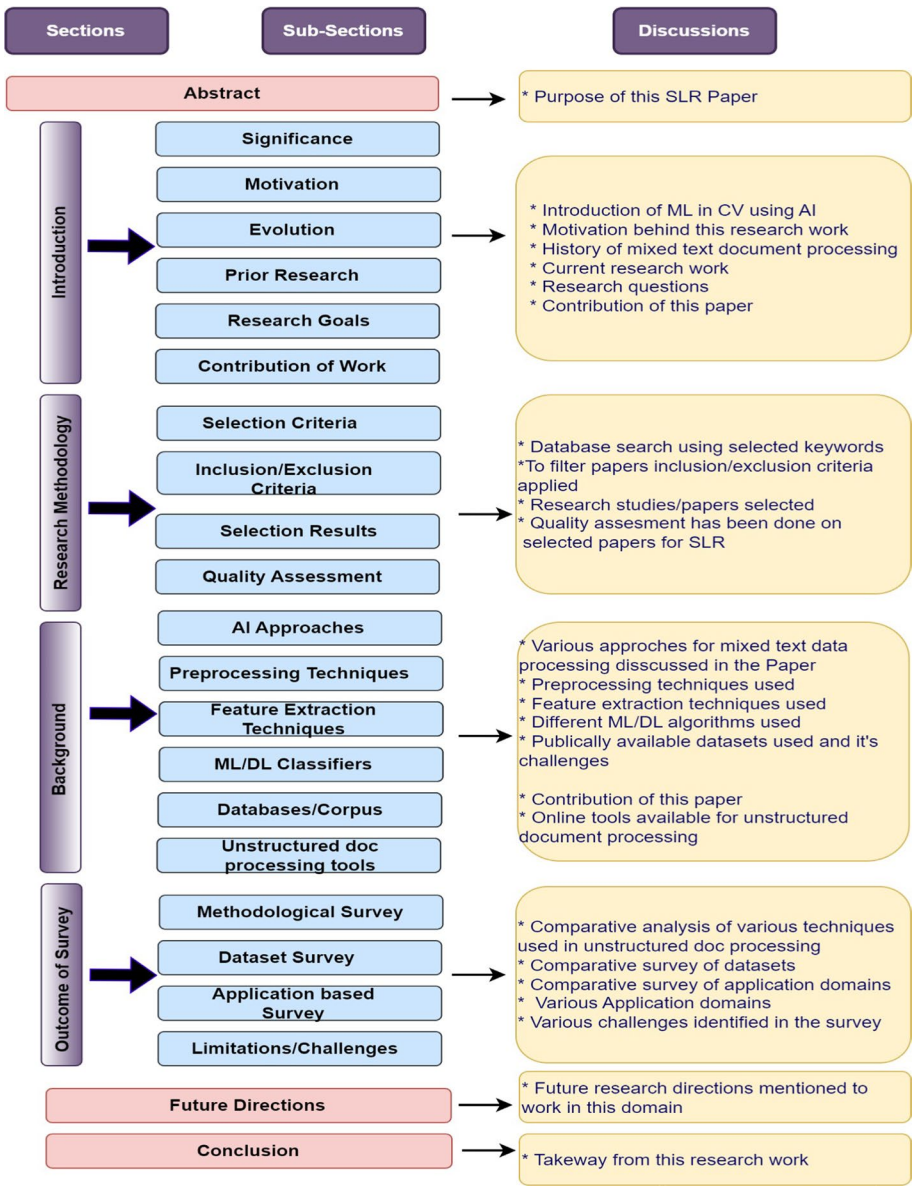

Fig. 2 Organization of paper

Table 1 Research questions

Sr. No	Research questions	Objective
RQ-1	What are the AI-based approaches used in unstructured document data recognition?	The goal is to examine several Artificial Intelligence (AI) strategies for mixed text data recognition, their benefits, and drawbacks, and to demonstrate a comparative comparison of various techniques
RQ-2	What are the domains and applications that utilize unstructured data processing?	Different application domains demand various approaches and datasets since they have various requirements. The goal is to examine the various application domains that require unstructured document processing, their specifications, and comparative analysis
RQ-3	What are the various data preprocessing techniques associated with unstructured document processing?	The aim is to study various data preprocessing techniques used for printed, handwritten, image-based multimodal data processing and analysis
RQ-4	What are the many issues that information extraction algorithms face while processing data from unstructured documents?	To find and analyze different AI-based information recognition and extraction methods used for unstructured document processing
RQ-5	What types of unstructured data processing datasets are available?	By examining application domains, data sources, data size, printed and handwritten text labels, and data imbalance, the objective is to examine the publicly accessible datasets for mixed text data recognition

F. Contribution of the work

This literature review has made the following contributions, which are highlighted:

- An in-depth review of the research literature on unstructured document processing based on handwritten and printed data recognition and processing, focusing in particular on approaches, datasets, applications, related difficulties, and future possibilities is delivered.
- A summary of several publicly accessible datasets that may be utilized to help this field of study is given.
- An overview of the systems in place for automatically extracting information from unstructured documents is described in brief.
- Researchers and financial institutions will be able to select the best method for automatically extracting crucial information from unstructured documents using AI methods with the help of the stated research requirements in this field. The probable future directions for this field of research are described.

Research methodology

SLR is a type of research that uses a certain technique to properly and (to some extent) consistently locate, evaluate, and interpret information that was previously available and relevant to a given research issue. The systematic literature review(SLR) in this article adheres to the PRISMA standards. A set of guidelines for the planning and structuring of systematic reviews and other data-based meta-analyses is provided by the Constructed Response Items for Systematic Reviews and Meta-Analysis (PRISMA) approach [14]. Moher and colleagues published these suggestions in 2009. The selection criteria, inclusion/exclusion criteria, selection outcomes, and quality evaluation are the f elements that comprise the authors' study strategy.

A. Selection criteria

To conduct a literature search about mixed text unstructured document processing, authors mainly used the IEEE, Science Direct, Scopus, and Web of Science databases. To obtain the pertinent articles from various database searches, a specific query was first created. The article selection is next subjected to the filtering procedure to enhance the outcomes that satisfy SLR's main objectives.

Duplicate items are removed, criteria for inclusion and exclusion are used, title and abstract-based filtering is applied, and full-text screening is the concluding step displayed in Fig. 4.

B. Inclusion/exclusion criteria

To search for significant research articles for a systematic review, the authors devised a set of inclusion criteria for research article selection and exclusion criteria for research

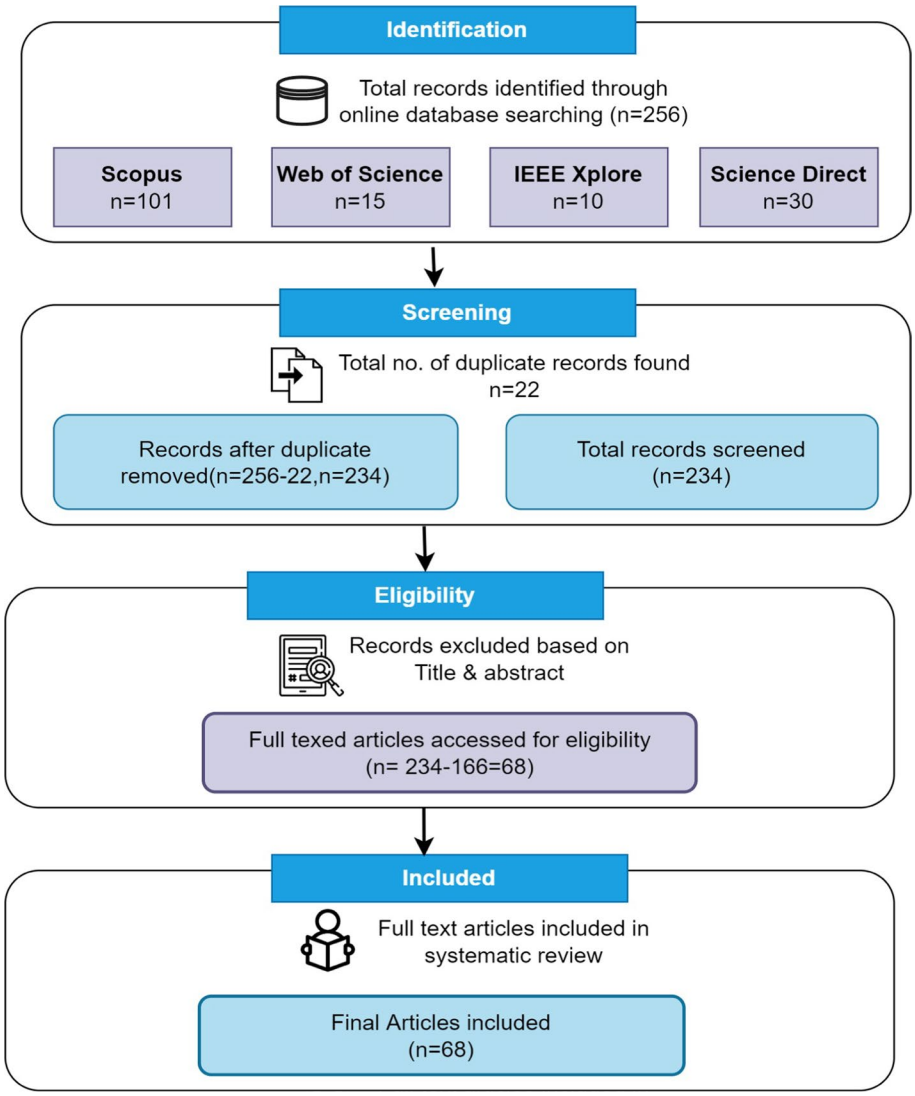

Fig. 4 Stepwise in detail process of literature review

article rejection. After finding all relevant publications, the authors eliminated those that didn't adhere to the criterion by using the exclusion criteria listed in Table 2.

C. Selection results

A total of 1008 articles were found using the first search engine result. IEEE Xplore supplied 24, Science Direct offered 225, SCOPUS discovered 722, and Web of Science offered 37 papers, articles, and research publications associated with text-based emotion identification. Thereafter, 1008 articles applied inclusion/exclusion criteria, and the results were obtained. A further search that focused on the publications' titles, abstracts, and keyword relevance turned up 256 articles. The search query given to all these databases with the search results count of each is shown in Table 3.

After applying full-text filtering and quality evaluation criteria to each item, 68 articles were selected for the comprehensive literature study. Studies from conferences, journals, and reviews were compiled from 2012 through 2023 [14].

D. Quality assessment

The following steps were used tso assess quality:

1) Mixed text recognition: Research must be focused on machine-printed and handwritten text recognition, techniques, or datasets.
2) Unstructured document processing: Research must focus on printed or handwritten text detection on unstructured documents used in different application domains.
3) Research must concentrate on the many artificial intelligence methods utilized in the extraction of mixed text from unstructured documents.
4) Datasets: Publicly accessible datasets relevant to mixed text-based data identification and processing are the main focus of research activity.
5) Techniques for Classification—The research project has a strong emphasis on classification methods used in unstructured document processing

Table 4 gives the overall summary of the existing literature in the domain of unstructured document processing. It includes datasets used, application domains,

Table 2 inclusion and exclusion criteria list

Criteria No	Inclusion criteria	Exclusion criteria
1	Publication Years of articles must be between 2012 to 2023	Non-English research articles
2	The different types of Articles, Conference Paper, Conference Review	Duplicate research articles
3	Engineering, Computer science, Healthcare, Finance & Accounting, and Social sciences articles were taken into consideration	Research articles with non-availability of full-text
4	Articles should match at least one of the search keywords	Research articles that are not relevant/ not focused on mixed text data recognition for unstructured document processing and artificial intelligence
5	The research questions should be addressed in the articles	Articles that do not contain data pertinent to the study's question

Table 3 Literature database sources with search queries

Database	Query for searching	Number of articles acquired	Articles meeting the inclusion and exclusion requirements, count	Selection based on title and abstract	Final selection after quality assessment and duplication
Scopus	(TITLE-ABS-KEY (mixed text data recognition) OR TITLE-ABS-KEY (OCR) AND TITLE-ABS-KEY (artificial AND intelligence OR deep AND learning OR machine AND learning) OR TITLE-ABS-KEY (RPA OR Robotic Process Automation OR NER OR text recognition) OR TITLE-ABS-KEY (printed and hand-written text recognition OR unstructured document processing))	722	124	101	35
Science Direct	("mixed printed & handwritten text recognition") AND ("artificial intelligence or deep learning or machine learning")	225	179	130	20
IEEE Xplore	("mixed printed & handwritten text recognition") AND ("artificial intelligence or deep learning or machine learning") OR ("unstructured document processing")	24	13	10	7
Web of Science	("mixed text-recognition*") AND TOPIC: ("artificial intelligence" or "machine learning*" or "Deep learning*" or "computer vision*") AND TOPIC: ("Printed and handwritten text recognition*")	37	17	15	6
	Total articles	1008	333	256	68

Table 4 Summary of existing surveys for mixed text data recognition and processing

Refs.	Datasets used	Applications	Pre-processing techniques	Feature extractionm techniques	Multimodality	Performance metrics	Challenges discussed in the survey paper			Overview
							Mixed data analysis	Technique	Future directions	
[15]	×	√	√	×	×	×	×	×	×	This survey focuses only on English-based simple handwritten character recognition
[6]	√	√	√	√	×	√	×	√	√	In this, OCR based handwritten character recognition techniques with future directions are discussed
[101]	√	×	×	√	×	×	×	√	√	It focuses on advances in the techniques used for the detection & recognition of mathematical expression
[16]	×	√	×	×	×	×	×	×	×	Various ML algorithms used for handwritten documents and image data detection are discussed
[3]	√	√	√	√	×	√	×	√	√	It focuses on AI-based approaches to extract useful information from unstructured documents
[17]	√	√	√	√	×	√	×	√	√	Based on ML & DL techniques, this research suggests a survey for the identification of diverse handwritten documents
[11]	×	√	×	×	×	×	×	√	√	The authors presented a review of recently published research on clinical note-extraction applications
[10]	×	√	√	√	√	√	×	√	×	This review presents an overview of recent advancements in NER, NED(Disambiguation), NEL(Linking) for unstructured document processing
[1]	√	√	×	×	√	√	×	√	×	This review addresses this limitation and presents a systematic literature review of state-of-the-art techniques for a variety of big data, consolidating all data types. Recent challenges of IE are also identified and summarized
Proposed paper survey	√	√	√	√	√	√	√	√	√	The preprocessing, feature extraction, and classification methods for mixed data recognition with difficulties have all been compiled in this study, along with AI-based solutions

pre-processing techniques, feature extraction techniques, multimodality, performance matrices considered, various challenges covered, future directions, etc. It has been observed that no proper systematic literature survey present in this area which covers multimodal data consisting of mixed data with images.

Background

Digital text recognition is growing fast in the research world. A text recognition system is available that aids in creating an interface for communication between humans and computers. Printed or handwritten text is digitally encoded for text recognition. To better understand current research, this section contains a thorough qualitative analysis of pertinent literature. In this analysis, "(unstructured text-based AND data analysis AND data recognition or e-governance or clinical notes)" keywords were used to find the impact of unstructured document's data recognition of printed and handwritten text from unstructured documents. By reviewing the selected papers, appropriate articles were chosen for further in-depth analysis. Further, this research study investigated how unstructured document data recognition and summarization can be helpful in various domains like healthcare, insurance, e-governance, etc. in decision-making.

Interest in addressing the summarized data extraction challenge has emerged in fields where machine learning is used, such as Computer Vision (CV) and Natural Language Processing (NLP). Depending on the format of the unstructured document, various AI-based approaches are utilized to achieve the best possible performance. Massive amounts of unstructured data are now flooding the Internet and social media. Individuals find it difficult to rapidly locate useful information in such a large, unstructured document corpus. The proposed Unstructured document processing framework mainly divided into four parts is displayed in Fig. 5.

A. Data preparation & preprocessing

Obtaining an input unstructured document image with handwritten and machine-printed text is part of the data preparation stage. The image in this case needs to be

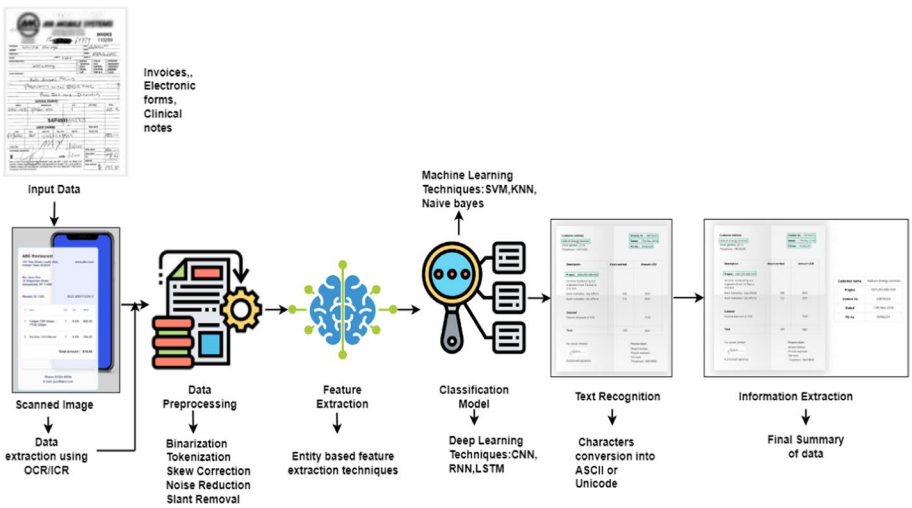

Fig. 5 Proposed workflow of printed and handwritten text recognition from unstructured documents

in a specific format, like PNG or JPEG. The image is taken using a digital camera, scanner, or other suitable input devices. On the other hand, during the digitization stage, the input document must be turned into an electronic format. The original of a document has to be scanned to produce a computer-savable image that can then be transformed. The digital image is required for the pre-processing phase [18]. Figure 6 illustrates the most frequently used datasets, tools, preprocessing techniques, feature extraction approaches, and machine learning and deep learning techniques used for classification.Before being pre-processed to reduce noise, the digitized image is first checked for skewing. For optical character recognition systems to identify data, it must first be preprocessed. Pre-processing is used to reduce background noise, highlight the subject of interest, and effectively separate the foreground from the background in an image [18]. In the case of unstructured or semi-structured document processing to isolate a character or word from the background of an image, pre-processing is required. After the data collection step in OCR for text recognition from unstructured documents preprocessing will be performed. It includes binarization, noise reduction, skew correction slant removal, etc.

- Binarization: using this method, an image is reduced to pixels that are only black and white. To do this conversion, a set threshold value is used. If the value is higher, a white pixel is assumed to exist; otherwise, a black pixel. The widely used algorithm for binarization operations is Otsu's Binarization.
- Noise reduction: it purifies the image by eliminating any undesirable spots or dots and patches.
- Skew correction: it aids with text alignment. A document may be slightly skewed while scanning (image aligned at an angle to the horizontal). Skew correction is accomplished using a variety of methods such as Topline, Scanline, Hough transformation method, and Projection profile method.
- Slant removal: using this method, the slant from the text that could be present in some photos in the dataset is eliminated.
- Border removal: in some cases, an input document with a border is considered as an extra character to avoid this preprocessing technique being used.

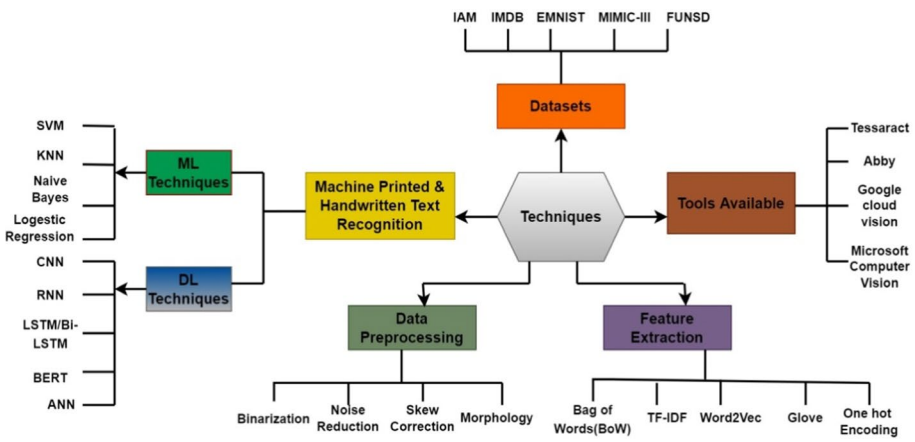

Fig. 6 Techniques used for unstructured document preprocessing

- Perspective transformation: it converts images with extra text to the proper image for processing.
- Negative image: negation is the process of turning bright regions of an image into dark ones and vice versa. Negation of the image after normalization changes the pixel values of 1–0 and 0–1 in an input image. This step is essential since it allows the next stage's identification of boxes. The preprocessing step converts the textual content into a format that ML or DL algorithms can better understand. Different preprocessing techniques used in form types of unstructured document processing are shown in Fig. 7. Some of the Preprocessing techniques included in NER are tokenization, normalization, and noise reduction Tokenization is the process of breaking the text up into smaller pieces, or "tokens." Stop words are eliminated during normalization, and all text is written in lowercase. By eliminating surplus white spaces, noise reduction conducts text cleaning.
- Tokenization: it is a process of breaking a sentence up into smaller, called "tokens." Words, sub-words, or letters are considered tokens. As a sample, the tokenization splits the statement "This is an unstructured document" into smaller units called tokens as shown: ["This," "is," "a," "unstructured", "document"].
- Stop word removal: the auxiliary verbs "a," "an,", "the" and "in," as well as conjunctions and articles, are examples of stop words. They take up processing time and memory in the database. Therefore, stop words must be excluded from sentences while processing a document.
- Part of Speech Tagging (POS): assigning "tags" or "parts of speech" to each token in a phrase, such as a "noun," "verb," and "preposition". For instance, the tag for a single noun is "AA". The POS tagging output for the tokens ['Natural', 'beauty'] is [('Natural', 'AA'), ('beauty', 'AA')].
- Lemmatization: it is the process of identifying the lemma, base, or root form using a dictionary. When the word's meaning is crucial for analysis, it is advised above stemming. For instance, "writing" has the root form "write," while "thinking" has the root form "think."
- Regular expressions: these are patterns or groups of characters used as instructions to a function to discover a substring, match a group of strings, or replace a group of strings. It employs specific notations. The symbol (-) designates a range for a pattern, while the character ('?') denotes zero or more instances of a certain letter in a word.

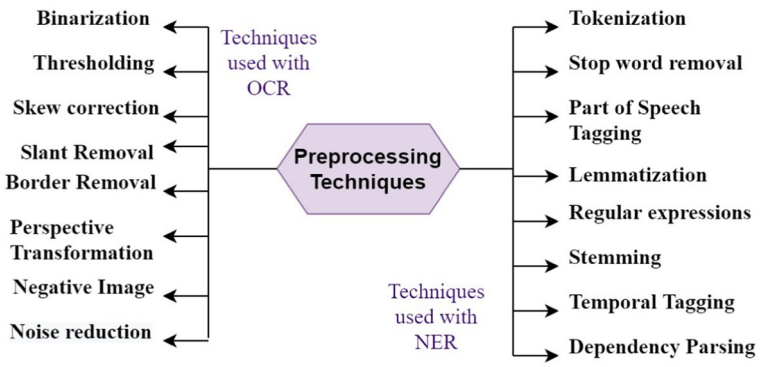

Fig. 7 Unstructured document preprocessing approaches

- Stemming: it is the process of shortening a word to its word stem, also known as a lemma, base, or root form. It could merely require removing known prefixes and suffixes. M Take, for example, is the root word meaning-making. Stemming involves removing the last suffix "ing."
- Temporal tagging: finding sentences with temporal expression or meaning is known as "temporal tagging." The temporal tag-detected for the example line "I may go to the university in the next 2 weeks" is "the next 2 weeks."
- Dependency parsing: it ascertains the connections between two words in a text that are symbolized by various tags. It says that "the" is employed as a "determiner" for "research work" in the sentence "I choose the morning time study for my research work," as an example.

B. Feature extraction

In every recognition technique, the selection of features is a crucial phase. Both should be simple to compute and reliable. The choice of features for training and classification tasks is then taken into consideration. GloVe, TF-IDF, and Word2Vec are a few feature extraction techniques used in text categorization and recognition. To obtain the vector or numerical representation of words, utilize Global Vectors (GloVe) [19].

The TF-IDF method is a well-liked way to give words weights that represent how important they are in the texts. The most popular way to use neural networks(NN) to learn word embeddings from very big datasets is Word2Vec. Continuous Bag of Words(CBOW) can be used for Word2Vec. To transform the text into a matrix (or vector) of features, feature extraction algorithms are needed. Data feature extraction is the common name for this transformation process. Figure 8 illustrates feature extraction methods. From the segmented subcomponents generated in the previous step, the extraction of important features from an input image is performed. There are several methods available to us for extracting characteristics like their form, strokes, etc. For every recognition methodology, feature selection is a crucial stage. Both of these

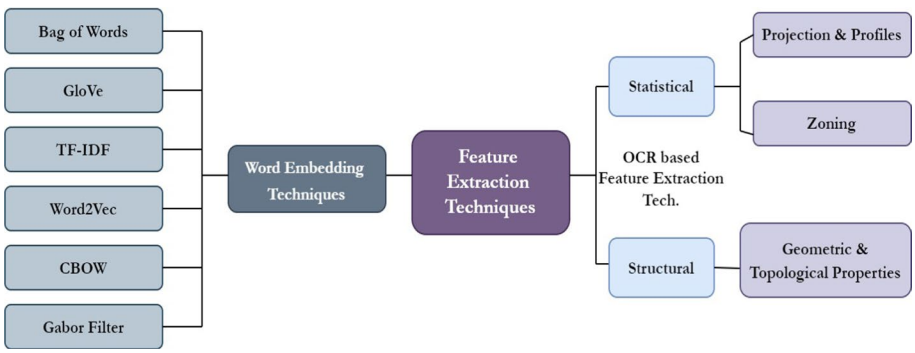

Fig. 8 Feature extraction techniques

attributes should be present. Then, feature selection is performed for both training and classification tasks [20].

Presently, mostly machine learning or deep learning models for feature extraction, which include a stack of CNN, RNN (Recurrent Neural Networks), and LSTM (Long Short Term Memory) layers, are used. Figure 9 shows the various feature extraction techniques used as word embedding techniques and techniques used in OCR for extracting important features.

For the feature extraction phase different methods used in handwritten and printed text classification and recognition are Global Vectors(Glove), Bag of words(BOW), Word2Vec, Term Frequency Inverse Document Frequency(TF-IDF), Continuous Bag of Words(CBOW), Gabor Filter, etc.

- Global Vectors(Glove)- to obtain the vector or numerical representation of words, apply GlobalVectors (GloVe). The count-based approach, also known as the encoding vector, encodes the ratio of the co-occurrence frequency of two words. The complete corpus is not used in this strategy, though. On tasks like word comparison, word similarity, and named entity identification, it enhances word representation learning and compiles global statistics [21].
- Bag of words(BOW)- the n-gram technique is another name for this technique. BOW produces a sparse vector that represents words that are present as one and absent words as 0 in the text by counting the frequency of words in the text. These vectors are used as input by machine learning algorithms. N-grams are collections of words that frequently appear together and are combined into a single vector of characteristics. N-grams are frequently represented as 1-, 2-, and 3-g units.
- Word2Vec- the most popular way to use neural networks to learn word embeddings from very big datasets is Word2Vec.
- Term frequency-inverse document frequency(TF-IDF)- The TF-IDF approach is a well-liked way to give words weights that represent how important they are in the texts.

Following is a definition of TF-IDF:

$$\mathbf{TF} * \mathbf{IDF} = \mathbf{TF} - \mathbf{IDF}$$

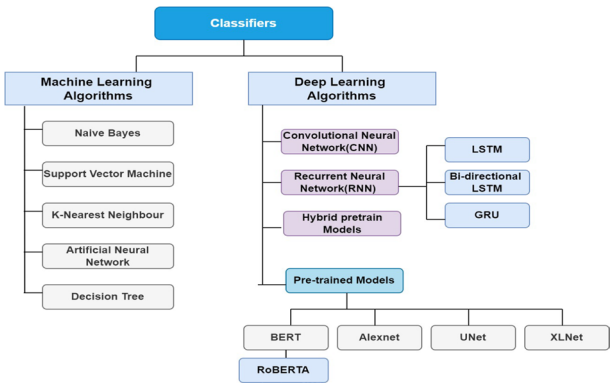

Fig. 9 Classifiers used for mixed text document data recognition and classification

Term frequency (TF) is calculated as the number of terms divided by the number of terms in the document [22].IDF stands for Inverse Document Frequency, which is calculated as Total Documents /Documents.

- Gabor Filter- Individual word images may be processed using the Gabor filter operator, and features can be extracted layer by layer via the Gabor filter feature extractor. Gabor wavelet transforms are the best for determining regional spatial frequencies because they have the dual qualities of multi-resolution and multi-orientation. Additionally, it has been discovered to produce distortion tolerance space for jobs involving pattern identification [23].

OCR includes two popular feature extraction techniques: statistical (which recognizes a character's statistical feature) and statistical(which recognizes a character's statistical feature) and structural (which detects structural features like horizontal and vertical lines, and endpoints) [24].

- Statistical feature extraction: By zoning, statistical feature extraction may be carried out. Images are separated into zones, and features are then retrieved from each zone to create the feature vector. Another approach for statistical feature extraction is projection and profile [25]. A character image's number of pixels moving in a distinct direction is determined by projection histograms. Characters like "m" and "n" can be separated using it. The amount of pixels needed to reach the outside border from the bounding box is determined using a profile. It is employed to specify the characters' outside forms. It makes it possible to tell between letters like "p" and "q."
- Structural feature extraction: The geometrical characteristics of a sign or character are retrieved during structural feature extraction. Character strokes, horizontal and vertical lines, endpoints, intersections of lines, and loops constitute the character's geometrical characteristics or structural aspects. It gives an understanding of the different elements that go into making up the persona.

C. Data classification techniques

A listing of the pattern classification models that have been widely and effectively used for character recognition is provided during the classification phase.

The several different types of methodologies include statistical techniques, ANNs, SVMs, structural methods, and multiple classifier methods, to name just a few. For statistical techniques, ANNs, and SVMs, the input feature vector is a fixed-dimensionality feature vector obtained from the input pattern. By elastically matching strings, graphs, or other structural descriptions, structural approaches identify patterns. The classification outcomes of several classifier algorithms are merged to modify the order of the classes [26].

While feature extraction intends to map input characteristics onto elements in a feature space, classification assigns each point in the space with a class name or membership score to the specified classes. Character recognition aims to get the class codes or labels of character patterns. Character patterns or words are split from document graphics, and then the task of recognition is to assign each character pattern or word to a class

from a given class set. Since many word recognition algorithms also utilize segmentation-based strategies with character modeling or character recognition integrated, the efficacy of character recognition is essential for document analysis. The classification of the unstructured document into printed and handwritten text types follows after producing feature vectors and word embeddings. The two classification methods most frequently utilized by researchers are deep learning and machine learning. In the image, an overview of the many classifiers employed in the unstructured document text recognition system is presented. This section provides a detailed explanation of the various deep learning and machine learning classifiers that are used to classify printed and handwritten t.

I. Machine learning algorithms

Machine learning classifiers are extensively and significantly used in printed and handwritten text-based classification. Because they utilize datasets that have been annotated or tagged, these classifiers are data-driven. Machine learning is the process of directly extracting information from a huge number of training instances or samples using computational techniques. These algorithms adaptively identify the optimum solutions and often get more effective as more instances are given for learning [27].

In the majority of OCR systems, an algorithm is trained on a given dataset and eventually learns to properly categorize or classify the character set and digits. Various machine learning and deep learning techniques used in unstructured document processing are illustrated in Fig. 9.

The following are the most often utilized ML approaches for categorizing printed or handwritten text literary studies:

- Naive Bayes classifier: this is the probabilistic approach to classification. To determine the class of novel or unrecognized data, it applies the Bayes theorem of probability [28]. Fields like patient ID and last visit date, among others, are recognized and categorized using it in EHR patient records [29].
- Artificial Neural network (ANN): by understanding the underlying connections between letters and words, it has demonstrated an excellent ability to automate text identification and data extraction. In the real-world Chinese passport and medical receipt dataset in [30], object recognition is carried out using region-based convolutional neural networks (R-CNN).
- Support vector machine: a supervised learning model is an SVM. It is employed to address classification and regression issues. Despite having the ability to handle non-linear data in a high-dimensional feature space, it is mostly employed to arrange linearly separable data. To categorize the data, it creates a collection of hyperplanes (decision boundaries) in high-dimensional space with the maximum margin possible. By employing the biggest margin, it seeks to increase the robustness of the categorization [29]. The most frequently used classification algorithm in OCR is the support vector machine. It outperforms all other classification techniques. Unlike the Naive Bayes approach, it is not predicated on any independence assumptions. In [31, 32], it is utilized for text classification or recognition.

- K nearest neighbor: this assigns a common characteristic of objects that are nearby to one another. It is used to separate, identify, and categorize Latin alphabets in uppercase and lowercase, as well as Devanagari consonants and vowels, for Indian scanned document pictures of the Latin and Devanagari scripts used in the research [31].
- Decision tree: a decision tree is a supervised classifier that works with no rules. It has nodes, branches, and directed edges like a tree. Each node represents a feature or attribute, each branch represents the criteria for making decisions or tests on attributes, and each leaf node illustrates a class label or result. Pruning, feature selection, and decision tree generation are the three stages of decision tree learning. The learning procedure makes use of straightforward decision-making principles from the features. A popular classifier for regression and classification is the decision tree. This approach uses a framework that resembles a tree to depict decision-making. Determining which characteristics to use, what criteria to use for splitting, and when to terminate are critical considerations when employing a DT. All characteristics are looked at throughout this process, and several division points are investigated and tested. To find the best route, the split with the lowest cost is chosen. A tree might perform even better after being pruned. Pruning is the act of removing less crucial components from a tree to decrease its complexity and, as a result, improve predictive power by eliminating overfitting [33].

II. Deep learning algorithms-

Deep learning classifiers are used during unsupervised learning. However, it could also be supervised or semi-supervised. Deep learning classifiers increase performance and accuracy by automatically learning and extracting information. Convolutional neural networks, recurrent neural networks, BERT, LSTM, Bi-LSTM, GRU, and pre-trained models like UNET, Alexnet, and XNet are the most widely used deep learning classifiers for the identification and categorization of handwritten and printed text.

- Recurrent Neural Network(RNN): The most popular Deep learning model for tackling the NER challenge is RNNs. For each iteration step in RNNs, all weights and parameters are utilized once again. Depending on the current input as well as the previously hidden state inputs, the next hidden state output is updated. The most crucial data from the prior inputs are stored in the hidden state, which serves as the neural network's memory. The study's BioNER (Biomedical Named Entity Recognition) methods employ RNN [25]. To detect the connections between proteins and medications or genes and illnesses, it extracts the relationships between biological entities. Convolutional Recurrent Neural Network (CRNN), a hybrid CNN/RNN network that extracts text from pictures of medical laboratory results, is suggested in [25] as a way for clinicians to review patient information.
- LSTM and GRU: There are two kinds of RNN networks: LSTM and GRU. Simple RNNs have extremely little internal memory. Complex Neural Network topologies are employed to address the short memory issue. The LSTM and GRU are the two most well-known. LSTM is made up of different memory cells and neural networks.

The information travels across various cell states. With these cell states, LSTMs may explicitly recall or forget information. The gates manage memory operations while the cells store information. The architecture of the Gated Recurrent Units (GRU) is a little bit less intricate and simpler. Data flow is controlled by gates and a hidden layer. GRU and LSTM have several characteristics. Both GRU and LSTM use a gating mechanism to perform memory-related operations.

The study [17, 34] claims that LSTM and GRU may be used to create supervised learning Natural Language Processing models or computer vision-based models to extract symptoms or illnesses from the unstructured clinical notes dataset. Only a few previous studies that performed information extraction tasks linked to memory have employed LSTM and GRU.

- CNN: CNNs are neural networks that are primarily employed in the study's identification documents dataset MIDV500, MNIST, and custom dataset to categorize pictures and perform feature extraction to identify and detect text lines [35, 36]. The research [37] suggested classifying invoices into three categories: receipts, machine-printed invoices, and handwritten invoices. A Deep CNN called Alexnet is utilized to extract features from invoices. On the Image-net dataset and subsequently, on the invoicing dataset, it serves as a pre-trained model. Utilizing OCR, Convolutional Networks can also digitize text. Information is extracted from business-related scanned documents, such as invoices and purchase orders, using a combination of CNN and computer vision techniques. The R-CNN was utilized in the study [38] to locate objects in real-time. Because of how quickly they can detect items, they are highly renowned for doing so. In the study [39], a CNN was utilized in conjunction with word embedding to extract named entities from clinical notes. To analyze the input and create a better feature set, the study [4] employed the convolution layer and pooling layers before passing the information to the LSTM to parse lengthy phrases. It aids in enhancing the efficacy and efficiency of the LSTM. An RPA application for dynamic object detection from the software applications interface was suggested by the study [40]. The input from several interfaces and menus is used to train a CNN for "dynamic object identification." Here, real-time software interface categorization is also done using CNN. Tensor Flow uses the CNN YOLO (You Only Look Once) tool to recognize objects in real time, extract features, enable decision-making, and carry out real-time action. The YOLO method outperforms all others because real-time object recognition and reaction need extremely less training and classification time. In complicated texts like books, CNN YOLO is also employed for object detection and feature extraction [41].
- BERT: it is appropriate for a range of NLP and computer vision tasks that include language processing. For transfer learning, the model is typically trained on an unannotated dataset. For various NLP tasks including sentiment analysis, phrase categorization, question answering, and others, this pre-trained neural network model may be fine-tuned. It is intended to pre-train Deep Bidirectional Neural Networks in Natural Language Processing using unlabeled text. The attention-based process known as a transformer, which is used by BERT, learns the contextual links between

the words in a text. A transformer encoder reads the overall word sequence concurrently in contrast to unidirectional models, which read the input text serially either from left to right or from right to left order [42]. Although BERT is a powerful NLP model, it cannot be modified for NER without using the NER dataset. In the article [42], it is discussed how to combine BERT with word embedding to improve NER performance in financial and biological documents. Improvement in the F1 measure is shown in [19] using DocRED, a big open-domain document-level information retrieval dataset, and fine-tuning the BERT model to accomplish document-level connection extraction.

- XLNet: both auto-regression and auto-encoding are features of the XLNet language model. While BERT hides the input and tries to forecast it using a bi-directional context, XLNet uses the permutation objective. Sentences containing entirely distinct word combinations are taught because Permutation Language can assist in the model's learning of bidirectional context. Positional eg and recurrence algorithms are then employed. To remember the predicted token's position, it employs two streams of self-attention [43]. Researchers used BooksCorpus, Giga 5, Common Crawl, Clueweb 2012–13, and English Wikipedia to pre-train and fine-tune XLNet.
- UNet: Olaf Ronneberger et al. created the U-Net for Bio-Medical Image Segmentation [44]. The design has two parallel paths: one for the encoder and one for the decoder. Standard convolutions and max-pooling layers are applied in the contraction route (Encoder) on the left. Here, the depth progressively rises as the picture size gradually decreases. The semantic context and characteristics of the picture are captured via the encoder path, sometimes referred to as the contraction path. Transposed convolutions are used in addition to regular convolutions on the expansion route (Decoder) on the right.
- The depth steadily lowers as the picture size gradually grows in the decoder. Accurate localization is achieved by the parallel decoder path, also known as the symmetric expanding path, which uses transposed convolutions. The Decoder instinctively extracts the "WHERE" information by gradually employing up-sampling or transposed convolutions (exact localization). To enhance, specific Skip connections are used at each stage of the decoder to combine the output of the transposed convolution layers with the feature maps from the encoder at the same level. To train the model to put together a more exact output, two successive regular convolutions concatenated. The symmetric U-shape that results from this lends the design its name, U-Net.

III. Hybrid approaches

To make optimal use of each type of neural network's unique characteristics, hybrid deep learning AI-based techniques for handwritten and printed text recognition and recognition from image datasets frequently integrate multiple neural network types and methodologies. Included here are some significant methods:

1. Convolutional Recurrent Neural Networks (CRNN)

Convolutional Neural Nets (CNNs) efficiently extract information from images by using spatial characteristics. Sequence modeling makes use of recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which govern the temporal relationships in the retrieved characteristics.

Example: For the creation of feature maps, a CRNN model passes an image through many convolutional layers. Following that, these maps are fed into an RNN, which interprets them as a sequence and generates word or character predictions.

2. Models based on transformers

Transformers are proficient at obtaining long-range dependencies through self-attention processes and are capable of handling data sequences. Transformers can improve OCR by offering improved erroneous correction and context understanding.

Example: combining a transformer model for sequence processing with a CNN for initial feature extraction. This configuration can be very useful for text recognition tasks where accuracy is greatly increased by context awareness.

3. Attention-based mechanisms

When formulating predictions, attention methods enable the model to concentrate on certain segments of the input, which is essential for text recognition in images where letters or phrases may be scrambled or distorted.

Example: Adding an attention layer to an encoder-decoder architecture, in which an RNN with attention serves as the decoder and a CNN commonly serves as the encoder.

The attention layer helps the model to selectively focus on relevant parts of the image.

4. GANs, or Generative Adversarial Networks

The robustness of OCR models is increased by using GANs for data augmentation and denoising, particularly when it comes to handwritten text recognition.

Better training data gets generated by the generator network, which generates synthetic data that the discriminator network attempts to categorize as real or false.

Example: To enhance the training dataset, use GANs to create artificial handwritten text imagery. This improves the OCR model's ability to generalize to different handwriting styles.

5. Mixed CNN-RNN-transformer methodologies

using the advantages of transformers, RNNs, and CNNs to build a reliable text recognition pipeline. Transformers improve contextual learning and prediction accuracy, RNNs handle sequence learning, and CNNs handle the initial feature extraction.

An example of a hybrid model would be one in which a CNN obtains features from the image, an RNN processes these data sequentially, and a transformer uses attention techniques to further improve the predictions.

6. Multi-task learning

By training the same model across several related tasks concurrently, Multi-Task Learning (MTL) enhances generalization by allowing task representations to be shared [29].

Language modeling, character recognition, and text detection are among the feasible tasks.

Example: A model that learns to identify characters, detect text sections, and comprehend linguistic context all at the same time. This approach makes utilization of shared representations to enhance performance.

7. Self-supervised learning

Self-supervised learning (SSL) uses vicarious tasks like identifying missing data points to train models on massive volumes of unlabeled data. Before fine-tuning models on particular OCR tasks, this approach can be used to pre-train them on huge text corpora.

Example: Pre-training a model with SSL on a big text image dataset, where the goal may be to determine which letters or portions of the image are missing. Labeled OCR datasets can then be used to refine this pre-trained model.

D. Data recognition

Due to the growing use of digital technology for data storage and transfer in many industries and virtually all daily activities, machine-printed and handwriting character identification have emerged as major research areas. Data from an unstructured document could be printed or handwritten.

Image acquisition, digitization, preprocessing, segmentation, feature extraction, and recognition are all steps in the handwriting text recognition process. Whereas machine-printed text recognition includes steps such as scanned input document image acquisition, preprocessing of an image, and feature extraction followed by classification and data recognition [12].

The data processing step includes noise removal, skewness removal, binarization, tokenization, stemming, etc. Further different word embedding and Bag of words techniques are used to extract the important features from the document. Then various machine learning algorithms like KNN, SVM, and Naïve Bayes [29] and deep learning algorithms such as CNN, RNN, LSTM, and GRU will be used for classification and recognition. With advancement in this field, new hybrid techniques gets revolved such as CRNN, transformer-based techniques, combination of advanced machine learning techniques like meta learning, few shot learning and transfer learning, etc. Figure 8. Shows various AI-based approaches used in mixed text data recognition and extraction with advantages and limitations of each.

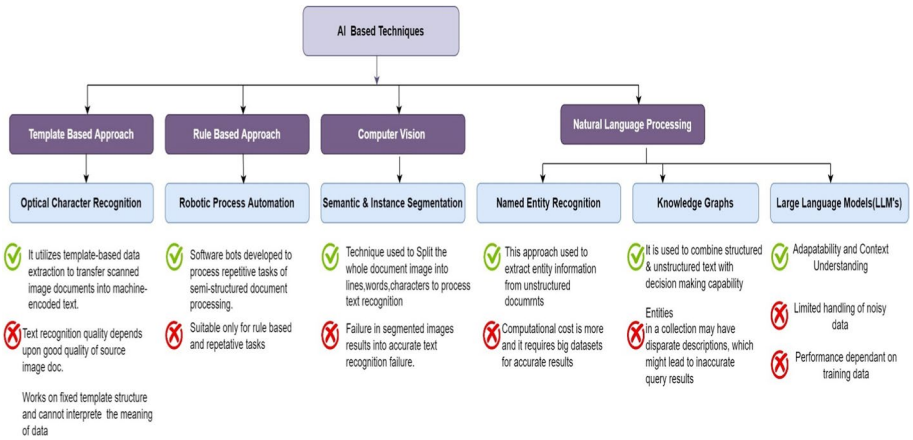

Fig. 10 Unstructured document processing approaches using AI

AI-based approaches for unstructured document processing-

Various AI-based approaches with their merits and demerits are discussed in this section shown in Fig. 10.

A. Optical Character Recognition (OCR)

Manual text extraction from unstructured documents like scanned Documents is labor-intensive and prone to errors since people are imperfect. Organizations have attempted to automate document processing by using template-based procedures like OCR. OCR is employed to identify text included inside an image, often a scanned copy of a printed or handwritten document [45]. OCR can automatically categorize various document types and arrange them according to specified rules. Keeping contracts maintained and up to date, for instance, based on the seller or the type of goods. OCR is employed in the case of electronic health records to keep track of a patient's earlier records [46].

Advantages-

- Efficient data input results in fewer data recognition errors.
- It converts scanned document data into easily editable forms, such as text and word.

Disadvantages-

- It can't comprehend how to analyze facts.
- A high-quality source image is also necessary for accuracy.
- Characters that are not text cannot be retrieved.

B. Robotic Process Automation (RPA)

Robotic Process Automation (RPA), a rule-based approach, is another way to automate extract, and analyze printed and handwritten data from unstructured documents (RPA). To use software bots to execute repeated tasks, this strategy aims for people to create

basic rules.RPA can work as automated software robots or cognitive RPA or Process selection as an RPA candidate for unstructured document processing.

The following are the important steps of Robotic Process Automation-

1. Determine which procedures to automate first step for businesses is to determine which processes may be automated.
2. Utilize software to streamline the procedures tasks that the recognized process completes are described in terms of a set of instructions. In this stage, the process automation workflow is defined.
3. Constructing/Implementing a controller for robot-Processes that are automated are transferred to a bot controller. In each RPA workflow, the bot controller is the most crucial stage. The process execution is prioritized and controlled by it. Users can plan, coordinate, and manage a variety of activities.
4. Integration with business software final stage, Business Process Management (BPM) solutions, or the organization's analytics tool can be incorporated with automated software bots.

Advantages-

- It can smoothly automate a lot of activities by writing software bots.
- It saves time, can perform repeated processes, and lowers operational costs.

Disadvantages-

- Multiple unstructured document styles and formats are difficult for RPA to handle.
- It necessitates establishing certain rules.

C. Semantic segmentation

For processing unstructured image-based documents, segmentation divides an image into segments that may be handwritten or printed in PDF format as text. Methods for segmentation include text line recognition and word extraction. Accurate segmentation is necessary for character recognition to be done correctly. Segmentation can be either instance segmentation or semantic segmentation [26].

The text-based image segmentation method is restricted by a variety of factors. Some of them are as below:

- Image quality: an important consideration in text segmentation is image quality. The accuracy and efficiency of the image are reduced when noise is present.
- Written or printed document: the majority of text line segmentation techniques are predicated on the notions that text lines are equally straight and that the distances between them are accurate. These presumptions, however, do not apply to handwritten documentation. Text image segmentation is a significant problem when dealing with handwritten documents. The printed text documents are examples from earlier [47].

- Text orientation: when extracting text from a handwritten document, the complexity of the task rises if the individual lines are curved or if the skew is present.
- Document with texture: the work of segmentation becomes complex due to the presence of texture in the text content, such as images, shapes, and other elements.
- Text type: because ligatures are present in the cursive text, character segmentation might be made more difficult.

In computer vision literature, the process of labeling each pixel in an image with the class of the item to which it belongs is referred to as semantic segmentation.

Advantages-

- It is still one of the most often used segmentation techniques because of the unique quality that distinguishes it from other image categorization techniques.
- It is not necessary to have prior visual conceptions or familiarity with the things being categorized.

Disadvantages-

- The same image will be segmented differently by several different individuals.
- While segmenting the image, human observers apply high-level knowledge and tackle high-level vision issues like recognition and perceptual completeness.

D. Named Entity Recognition (NER)

NLP employs NER to extract information from a given text-based resources utilizing sentence-level syntactic rules, such as applying grammar rules, and patterns at the word or token levels, such as regular expressions. To locate "identified entities" such as people, places like countries and cities, businesses, appointment dates for patients, and bill numbers on invoices, it automatically searches through the unstructured text.

Advantages-

- No need for a template or rules; learning ability depends on gathered data.
- NER is efficient for automatically extracting entity info from unstructured documents.

Disadvantages-

- For mixed text data recognition to produce better results, large datasets are required.
- The cost of computation is higher.

Table 5 gives a summary of approaches with examples used in unstructured document processing.

E. Knowledge graphs

The usage of knowledge graphs(KG) for idea modeling and retrieval of contextual information has grown in importance in recent years. However, building a knowledge graph

Table 5 Summary of approaches with their application areas in information recognition and extraction from unstructured document processing

Refs.	Approach used	Use cases of unstructured document processing
[3, 6, 26, 46, 48] [37, 45, 49–51] [52]	Optical Character Recognition (OCR)	• Academic document digitization • Receipt/Invoice data recognition • Handwritten and printed scanned image document text recognition and classification • Historical document data recognition • Scientific Article data extraction
[3, 40, 53, 54]	Robotic Process Automation (RPA)	• Automation of banking and financial records • Automation of patient healthcare records • Automate HR process in an organization
[55–59] [60]	Named Entity Recognition (NER)	• Invoice data extraction • Clinical notes information extraction • Legal Clause data extraction • Scientific article metadata extraction
[26, 35, 47, 61] [20]	Semantic Segmentation(SS)	• Forms data recognition and extraction • Image-based documents text, signature identification
[10, 62, 63]	Knowledge Graphs(KG)	• Clinical notes digitization • Cyber security for knowledge gain of attacks • Network analysis • Bioinformatics
[64]	Large Language Models (LLM)	• Agriculture document's information extraction • Customer feedback

for a new subject using extracted large, complicated, and unstructured documents in the absence of an established labeled entity-relation dataset or a well-defined ontology. It is considerably easier for the domain expert to extract the entities and relations from the specified relational schema if the data are stored in a structured relational database and data sources are well-integrated [62]. Unfortunately, the majority of domain knowledge is only preserved by human experts or may be found in a variety of data sources, such as unstructured texts that have been digitized or printed, product manuals, catalogs, online portals, legacy information systems, etc. It might be difficult to initially combine a lot of data from several sources before displaying it to find the important entities and relationships.

In this study, the authors examined the difficulties in creating a knowledge network for cyber security instruction out of unstructured materials. Cyber security is one of the cutting-edge fields that should take advantage of the potential of knowledge graphs to develop an interactive and adaptive learning environment to educate cyber security professional skills. Knowledge graphs have been successfully utilized in education [63]. Researchers working in this domain can now use unstructured text data in addition to simply structured data. This significantly increases the capability, reach, and use of knowledge graphs.

Advantages-

- It is used to combine data that is both structured and unstructured.
- It assists organizational decisions in making better responsible decisions.

Disadvantages-

- Entities in a collection may have disparate descriptions, which might lead to inaccurate query results.

F. Large Language models

LLMs are engineered to understand human-like language input and produce human-like text outputs. For document capturing, this implies that before deploying LLMs, preprocessing of input data images or PDF-like files, and converting them into text need to be performed [64]. Humans don't typically communicate information solely through prose. Use of paragraphs, tables, alignments, and more to convey information efficiently is required. The layout itself carries meaning. By default, LLMs don't process 2-dimensional information. They handle plain text, line by line. Simply using OCR to convert text and feeding it into an LLM might not suffice. Ideally, segmentation of the content into semantic chunks or regions to provide meaningful input for the LLM is the first step.

Disadvantages-

- LLMs have limitations on the volume of input and output text they can handle. Current LLMs can process only up to a certain number of tokens. For extensive documents, a strategy is needed to break the content into coherent segments.
- In the case of invoices, arguably the most common document globally. While certain fields on invoices are standardized (especially in Europe), they might contain company-specific information not available publicly, making it challenging for an LLM to capture.

Large Language Models (LLMs) such as GPT-4 have the potential to significantly improve deep learning-based methods for handwritten and printed text recognition and detection from image datasets. To increase text recognition's overall accuracy and understanding of context, these methods usually integrate LLMs with other neural network architectures. Various LLM-based deep learning artificial intelligence techniques are as follows:

1. Contextual error correction with LLMs:

By comprehending the context and semantics of the text, LLMs can be used to rectify problems after an OCR engine has initially recognized the text. This greatly increases text recognition accuracy, particularly in complicated or noisy photos.

Example: An OCR system generates a preliminary transcription after identifying the text in an image.

Use Case: Improving the precision of printed documents where OCR may have misinterpreted certain characters or words.

2. Recognition of handwriting using LLM-enhanced models:

LLM-equipped CNNs and RNNs: A hybrid model that employs RNNs (such as LSTMs) for sequence prediction, CNNs for feature extraction, and LLMs for contextual comprehension can be very successful in handwritten text recognition.

For instance: A CNN processes an image of handwritten text to extract characteristics, which are then input into an RNN to predict the letter sequence. By using contextual data to improve and correct the predictions, an LLM further refines the outcome.

Use Case: Identifying and scanning old documents or handwritten notes.

3. End-to-End text recognition with transformer models:

Sequence-to-Sequence Learning using Transformers: LLMs are built on transformers, which are immediately applicable to end-to-end text recognition tasks. Through self-attention mechanisms, they are able to integrate contextual awareness and effectively handle sequence prediction.

Example: A transformer model outputs the recognized text sequence straight from the input, which is an image of text. By assisting the model in concentrating on specific parts of the image, the self-attention mechanism increases the accuracy of recognition.

Use Case: Real-time text detection in images taken using cameras or mobile devices.

4. Multi-stage processing with LLM integration:

Step 1: Using conventional OCR techniques, perform preliminary text detection and recognition.

Stage 2: Error correction and context-based enhancements through refinement with LLMs.

Example: An OCR engine and a CNN-based text detector are used in the initial processing of the image. After that, the produced text is sent into an LLM, which improves the text's coherence and improves errors to further refine the recognition results.

Use Case: Handling scanned documents with potentially misleading initial OCR outcomes.

5. Text recognition by few-shot learning with LLMs:

Few-Shot Learning: Few-shot learning can be used for text recognition jobs where there is limited labelled data available by leveraging LLMs' capacity to adjust to new tasks with few samples.

Example:An LLM is fine-tuned with a small dataset of handwritten or printed text samples. It then applies this learning to recognize text in new, unseen images with high accuracy.

Use Case: Custom text recognition for specialized documents with limited training data.

6. LLM-assisted self-supervised learning:

Self-Supervised Learning (SSL): Using surrogate tasks (e.g., predicting missing text segments) to pre-train models on huge unlabeled datasets, and then fine-tuning on specific text recognition tasks.

Example: An LLM is pre-trained to fill in words or characters that are missing from a large corpus of text images. Then, labelled OCR datasets are used to refine this pre-trained model for handwritten and printed text recognition.

Use Case: Enhancing the OCR models' generalization and robustness for a variety of printed text formats and handwriting styles.

7. LLM-based attention mechanisms:

Including Focus in OCR Pipelines: Enhancing the emphasis on pertinent areas of the image by combining attention processes with LLMs will increase recognition accuracy.

Example:An attention-based model processes the image to highlight important regions (like individual characters or words). An LLM then uses this focused information to produce more accurate text recognition results.

Use Case: Text recognition in complex scenes where the text is cluttered or partially obscured.

8. Semantic understanding and information extraction:

Semantic enrichment refers to the process of using LLMs to understand and extract relevant information from text based on its semantic recognition.

Example: To increase the usefulness of the recognized text, an LLM processes it by extracting entities, summarising its meaning, or translating it into another language.

Automatic data extraction from receipts, invoices, and forms is the use case.

By using big language models' strong contextual understanding and error correction capabilities, these methods can greatly improve the accuracy and resilience of text detection and recognition systems.

Application domains

Recognizing and processing unstructured document data is widely applicable across many application disciplines. Management and marketing, user engagement, healthcare, education, finance, public monitoring, e-governance, and other application domains are included in this broader spectrum of application domains. In numerous fields, including online social media, consumer or product review systems, recommendation systems, academic records, and patient electronic health records, a large volume of text data is produced and processed to make recommendations. Therefore, there is a significant amount of potential in this field for unstructured data extraction and analysis.

- Healthcare: using information from unstructured documents is a significant change in the health industry that will optimize and enhance the standard of treatment and foster innovation and analysis in clinical procedures. To foster the process of patient

treatment, clinical note processing is essential which can be done using an AI-based framework of unstructured document processing [65].

- Insurance: it is a different industry that may gain a lot from managing and analyzing unstructured data. In addition to being able to quickly identify issues and fix difficulties, it is also possible to extract all the information possible from the daily information that is created. To offer more adaptable and market-oriented procedures, for instance, travel insurance claims might be standardized to make judgments about services and to concentrate on the consumer [56].
- Banking: some industries have internal procedures that require the processing of a tremendous volume of data. Analyzing a person's work history and searching for information in a public deed are two concrete instances of processes that banks may need [37]. In these situations, using a technology that can retrieve customer information is especially crucial.
- Basic Services: it is the use of unstructured data in basic services allows for the identification of users' and customers' requirements, wants, and ambitions in order to create new services or enhance current ones, as well as streamline all internal personnel operations. Invoice/bill data digitization is also an important application in it [52].
- Government services: in various types of e-governance facilities unstructured document processing is essential. Automating address change requests and driving License renewal by verifying the user details are some of the important applications where unstructured document data recognition and analysis are performed.
- Agriculture documents: pest identification is a crucial aspect of pest control in agriculture. However, most farmers are not capable of accurately identifying pests in the field, and there is a limited number of structured data sources available for rapid querying. This work explored using a domain-agnostic general pre-trained large language model(LLM) to extract structured data from agricultural documents with minimal or no human intervention.
- Travel sector: in the travel sector, while traveling extracting and verifying ticket booking is the application of unstructured document processing. Extracting passenger details, bus timing, or information extraction is also an important application where automatic information extraction from unstructured documents is required.

ATMs are one of the use cases for this, allowing users to deposit their handwritten checks within the device. OCR is also utilized to process card transactions at payment terminals or for biometric card logins, as well as to evaluate OMR response sheets from entrance tests. Almost every industry now takes advantage of AI's recognizing data pattern function in unstructured document processing.

Datasets available

Every AI-based model is built on data. Building a better performance model will be assisted by obtaining contextual data that has been accurately obtained, is relevant, and has completed bias testing.

- MIXED-IAM labels dataset

The IAM database has a total of 1539 types of handwritten English text produced by 600 writers in two sections on a single page, with machine-printed text in the upper portion and space for the writer to fill in with their handwriting by transcribing the machine-printed text [20, 66, 67]. Each text line in the section on handwriting is tokenized and included in the dataset. The authors also present a paragraph-level transcription of the machine-printed section of the page in addition to the database. The authors integrated the two portions into a single transcription that conforms to the lines and word order in the page without respect to the wording, tokenizing the machine-printed portion of the page using the Natural Language Processing Toolkit (NLTK).

- NIST and MNIST dataset

The existing research has mostly employed publicly accessible datasets, including the MNIST collection, in languages as different as English and Arabic in unstructured document data recognition. The most popular and often referenced dataset for handwritten digit recognition in the English language is modified NIST (MNIST). It is referred to as a modified NIST or MNIST since it is a subset or component of the NIST dataset [17, 36]. Samples from MNIST are normalized. It shortens the time needed for pre-processing and organizing data. 10,000 testing sample photos are included and 60,000 training sample images.

- EMNIST dataset

The EMNIST dataset is a collection of handwritten character digits taken from NIST Special Database 19 and rendered in an image format of 28X28 pixels. All of NIST's model training resources for character and handwritten document identification may be found in the parent NIST Special Database 19. The EMNIST Balance dataset is used to evaluate and compare the two models among the six datasets that are a component of EMNIST. This collection has 47 classes and 131 600 characters. The dataset consists of 10 numbers, 26 capital letters, and 11 small letters. In this dataset, several lowercase and uppercase alphabets have been combined. The closeness between some capital and lowercase letters is a fascinating issue that arises in the categorization of handwritten numbers, and this is done to handle it [12, 68]. To avoid inherent bias, it is preferable to use the balanced dataset rather than alternative, more populated datasets. An overfit and biased model is frequently the result of an uneven dataset. This dataset is reduced to 36 classes after the classes for the lower case were eliminated since such forms are anticipated to only include upper case alphabets.

- PAW dataset

The PAWs (Printed Arabic Words) collection is used for OCR in Arabic. All of the terms in PAWs are in Arabic. Arabic nouns often include many subwords (PAWs). It has 83,056 PAW pictures in total, which were taken from a total of 5,50,000 different words. Five distinct font styles—Naskh, Thuluth, Kufi, Andalusi, and Typing Machine are used

for each word sample image [69, 70]. It is applicable for tasks involving document image analysis and Arabic input recognition.

- CEDAR dataset

The comprehensive literature review on handwritten OCR has provided a list of publicly available datasets for Chinese, Indian, Urdu, and Persian/Farsi languages, including CEDAR (English), CENPARMI (Farsi), UCOM (Urdu), PE92 (Korean), and HCL2000 (Chinese) [6, 71–73].

The Center of Excellence for Document Analysis and Recognition dataset was developed by the University of Buffalo in 2002. (CEDAR). It is an online collection of over 200 writers' handwritten English text lines, all of which are kept up to date in an electronic version.

- HCL 2000 dataset

A dataset of Chinese handwritten characters is called Handwritten Chinese Language character-2000 (HCL2000) [74–76] 0.3,755 of the most popular Chinese handwritten character photos, captured by 1,000 different individuals, are included. Because it primarily consists of two subcategories—the metadata for the relevant writer information dataset and a collection of Chinese handwritten characters—it is exclusive. The character image recognition task and the writer's metadata extraction task may both be applied to these categories of the dataset. For instance, information about a writer's age, gender, profession, and education can be gleaned.

- PRImA and UvA datasets

Document photographs of various types and layouts are included in the Pattern Recognition and Image Analysis(PRImA) [77] dataset, which is used to analyze printed document layouts. The UvA dataset [78] is another dataset constructed from scanned color document pictures with various layouts. The process of layout detection and segmentation uses the color document image analysis dataset UvA.

- PubMed dataset

The dataset PubMed [11, 79] contains 26 million citations for MEDLINE publications in life science journals, online books, and biomedical literature. These references also provide links to the full-text publications on publisher and PubMed Central websites. The articles' bibliographic references are also given together with their metadata data.

Due to this characteristic, the PubMed dataset is the most important and widely used dataset for activities involving metadata extraction.

- RVL-CDIP dataset

The RVL-CDIP dataset is designed for complex document information processing tasks, including document image classification. It was created by the Ryerson Vision Lab and is commonly used for benchmarking and evaluating document analysis and recognition algorithms. The dataset contains images of documents in various formats, including

newspapers, scientific articles, letters, and more. Documents are labeled into different categories, such as letters, form, email, handwritten, and scientific article.

The dataset consists of a large number of grayscale document images. It is commonly used for tasks such as document image classification, document layout analysis, and optical character recognition (OCR).

- MIMIC III dataset

The notes for this inquiry were extracted from the MIMIC-III database. De-identified clinical data from MIMIC-III reflect almost 53,000 adult hospital admissions to the intensive care units (ICU) at the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset contains a variety of clinical note types, including nursing notes (n = 812,128) and discharge summaries (n = 52,746). It concentrates on the discharge reports since they are the most useful for phenotyping patients [65, 80, 81].

After reviewing and analyzing the literature on dataset accessibility, the authors concluded that there are several issues with the present dataset, as seen in the above figure. Publically available datasets are domain-specific which includes task-specific, only handwritten/printed character datasets, and language-specific datasets where most of the data is not labeled properly. So the scarcity of data is the biggest challenge in this research area.

The quality of data, skewness of the document [24], poor quality noisy data with patches or dots present in the image of the document, old datasets, blurriness' of the document, and variations in lighting conditions are the major publically available datasets related issues [82] displayed in Fig. 11.

There are some security-related challenges present in publically accessible datasets. It includes ethical issues where an organization's permission is required to access confidential data, patient's concern while accessing patient history, etc. [83]. Overall very

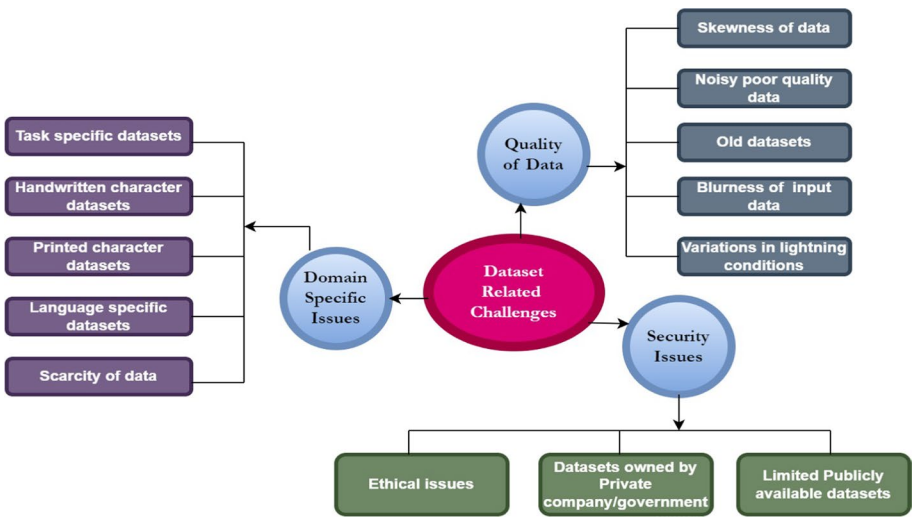

Fig. 11 Publically available dataset-related challenges

limited number of publically available datasets is present in this domain to do the research work.

Challenges related to available datasets-

Domain-specific issues-Very limited datasets are available in this domain which consist of mixed data with proper annotation/labels. Whatever datasets available are task-specific or language-specific datasets. So there is a need of own dataset creation and proper annotation process to be performed.

Issues related to the quality of the available datasets many document images are skewed. The angle of the document is not properly captured which leads to less accurate output data.

Blurriness or low/extra lightning conditions while generating input data becomes a problem in generating quality data. Many datasets available are too old to work on.

Some kind of patches, and dots available on input data requires a lot of preprocessing to achieve better results.

Security Issues-Some datasets are owned or created for a particular business or organization which is not available to work for the researchers.

If the data is sensitive such as clinical notes or any healthcare-related data then ethical issues occur before accessing that data is again a difficult problem. In summary, security, quality, and domain-specific problems have been noted in publicly accessible datasets.

Online tools available

Very little research work mentioned the existence of commercial technologies designed to meet the need for the automated extraction of valuable data/fields from unstructured texts. From unstructured records like invoices, clinical notes, orders, and credit notes, the businesses attempted to extract significant and particular information. These are commercial frameworks or tools. The literature is therefore lacking in information on the processes utilized to create these tools or frameworks, the datasets they used, the approaches they employed, or the assessment criteria they employed.

CULTIE- Convolutional Universal Text Information Extractor(CULTIE) is the instrument mentioned in the research study. CUTIE gains knowledge from the training data that is given to it. Significantly less human involvement is needed. It employs DL methods without establishing restrictions for any particular kind of document [84].

CUTIE can process texts' relative physical location coordinates and semantic data in parallel. It functions with "gridded texts." The grid's goal is to create a matrix-like or tabular structure in which the text is a "feature" that contains semantic data. The grid also keeps the text from the original scanned document's relative coordinate position connection. When OCR produces the recovered texts with their relative position coordinates, the gridded text is produced for the scanned document image.

Tesseract- Since 2006, Google has supported the development of Tesseract, an open-source text recognition engine that is available under the Apache 2.0 license. To extract the printed text from images, you may either utilize it directly through the API. The fact that it supports such a wide range of languages is its finest feature. Tesseract may be made to work with many frameworks and programming languages by using wrappers.

Tesseract works like C & C++ command line system. It accepts input files in the form of hocr /pdf for text extraction. Data preprocessing and page segmentation become challenges in Tesseract.

Abby- Abby Fine Reader is a paid tool used for text data recognition and extraction from unstructured documents. It takes input in the form of JSON,XML, and pdf and gives extracted text entities as output. Almost 50% accuracy has been achieved with synthetic noise in the case of mixed (printed and handwritten) data recognition. It gives low accuracy to handwritten text as compared with machine-printed text.

Google Cloud Vision- It is one of the popular tools used for information extraction. It is not good for tabular data extraction. It will give output in the form of JSON. Processing of the first 1000 pages each month is free; beyond that, payment is required to extract data from the document.

Microsoft Computer Vision- It is a leading OCR service provider in the market. Text detection, object detection, document label detection, landmark, and logo detection are the operations that you can perform using this tool. The image document is the input and the output generated using this tool is in the JSON format. Almost 85% of accuracy was achieved with synthetic noise document input.

This systematic literature review explored some of the more current tools on the market for extracting text from handwritten or machine-printed documents. However, there are still very few tools available for multimodal data recognition that are suitable for public use, meaning that they are limited to use by certain organizations.

Application case study

The patient's electronic health record (EHR) contains a variety of information formats that are commonly used in clinical settings, including tabular clinical data, image data (such as pictures, x-rays, MRI scans, and reports from pathological tests), and unstructured sequence data (such as clinical notes, forms, written reports, voice recordings, and videos). Recently, models using AI/ML and a variety of data sources are effective in the fields of cardiology, clinical note summarizing, pathology report digitalization, etc.

Figure 12 gives the stepwise workflow of clinical data processing. Input data can be tabular data of patient details clinical notes or images of various scans, and forms. It is then processed using the preprocessing techniques based on the type of input modality of data. Then the feature extraction modeling is performed. On the output word embedding techniques are applied on image and text data. After that model is trained using machine learning or deep learning algorithms. In the case of clinical notes summarization abstractive or extractive summarization techniques will be applied to get the summary of clinical notes which will be helpful for the physicians to make decisions related to the patient's further treatment within less time.

In this way, unstructured patient clinical records must be processed and automated.

Outcome of the survey

This section will address the results for every research question addressed in this study. The conclusions of primary research that served as the basis for developing research questions are discussed. There are tables added with the total results for each research topic as well as thorough justifications and analysis. It's important to keep in mind that

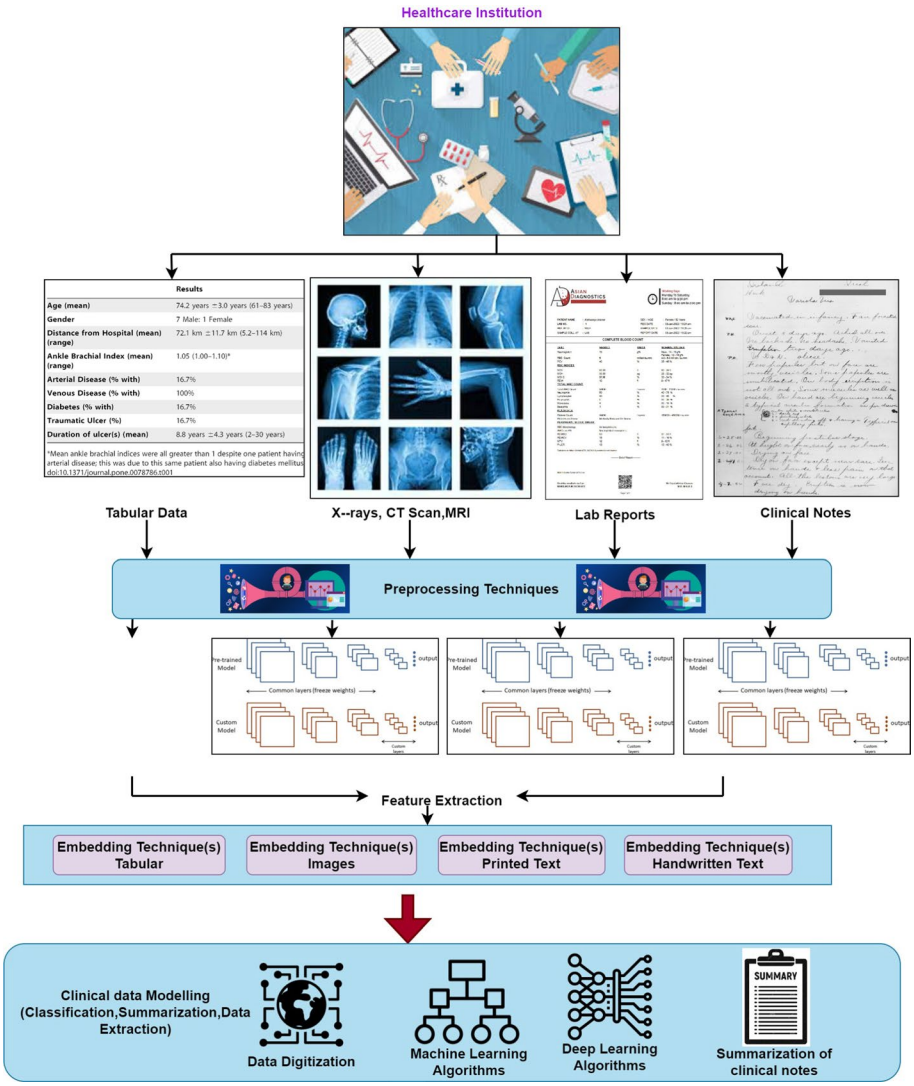

Fig. 12 Clinical unstructured data processing workflow

various tasks involving handwritten and printed data recognition require different features and approaches. As a result, a summary of each strategy is provided. All of the research question's findings will be summarized in this section.

RQ.1. What are the AI-based approaches used in unstructured document data recognition?
It is complex, time-consuming, expensive, and erroneous to effectively use and analyze unstructured data. Unstructured data consists of different types of printed and handwritten forms, financial documents, bills, patient records, clinical notes, academic records, blogs, etc. This unstructured data is relevant for further analysis since the information extraction techniques assist in retrieving insightful and important information from this data in various domains.

This SLR provides various information extraction methods along with a comparison of each. In light of varied domains and tasks, it has been found that different information

Table 6 Summary of handwritten text recognition systems in different application areas with objectives, methods used, datasets used, and challenges

Refs.	Application area	Description	Techniques/methods used	Dataset/data utilized	Challenges
[86]	Invoice/Bills	Developed an approach to identify text from printed bills	OCR, Adaptive character recognition, Tesseract	Bills used in over 100 languages	Inaccurately recognizes handwritten characters
[87]	Handwritten forms images	Proposed image segmentation-based model for handwritten character recognition	CNN, Open CV, Neural Networks	NIST	The system does not recognize cursive handwriting
[88]	Printed/typed Text images	Proposed a system for printed forms data recognition	Otsu's Algorithm for segmentation	Verdana font type	Only recognizes Verdana font type of size 14
[89]	Handwritten documents by various users	Technique for classifying handwritten and printed text automatically in ICR developed	OCR & ICR	1663 form zones scanned (947 handwritten,716 printed zones)	The system can't recognize cursive text
[18]	Handwritten filled forms	The system is implemented to take text data from forms and transform it into text for computer processing	ANN, CNN, Pytessaract, Open CV2	EMNIST	Not predicting the empty box with no text written
[13]	Filled application forms for citizen of Indonesia	Proposed ML model for recognizing handwritten characters in forms	CNN, SVM	Identity application form for Indonesian citizens from NIST	The connected issue with character recognition, difficulty with bounding box on form document data,
[90]	Arabic character Recognition	Developed a model to recognize Arabic characters from documents	CNN	AHCD dataset for isolated Arabic characters	Accuracy can be improved with modifications in the CNN algorithm
[12]	Motor claim forms	Presented a comprehensive method for isolating text using anchors, then an MCNN model was used to digitize hand-filled forms	MCNN-Multichannel CNN, Anchor based object detection	EMNIST dataset, motor claim forms filled by 200 different individuals(Hybrid dataset)	Models trained only on EMNIST dataset perform poorly on test data

Table 7 Summary of machine printed text recognition and extraction techniques in different application areas with description, methods used, data utilized, and performance

Refs.	Techniques used	Dataset used	Description	Feature extraction methods used	Domain	Performance
[80]	CNN, RNN	Healthcare (Chinese character data)	Medical invoice recognition	Scale-invariant feature transform (SIFT)	Invoice words	Accuracy 70–85%
[42]	NER, OCR, BERT, BLSTM	BC2GM BioNLP09	Extracting texts from financial and biomedical documents efficiently	Word2Vec, BERT	Text from biomedical and financial documents	Document precision 0.96
[46]	RNN, RCNN	Chinese passport and medical receipts	Multilingual text document text detection	OCR, CNN	Passport and medical receipt textual contents	Accuracy 94.25%
[91]	RNN, LSTM, OCR	3,26,471 invoices	Information extraction and digitalization from documents	CloudScan text extraction	Text from invoices	F1:0.840 Precision:0.87 Recall:0.80
[39]	LSTM-CRF,NER	I2B2/VA	Information extraction from medical records	RNN, LSTM	Words from Medical records	F1:88.78 Precision:87.75 Recall:88.46
[55]	BLSTM,LSTM,BLSTM-CRF	3500 contracts with 11 types of contract elements	Contract element extraction	Word2Vec	Contract information	F1:0.97 Contactn:0.87 Recall:0.80
[65]	CNN	MIMIC-III dataset	Text Classification of the Clinical Dataset	CNN	Disease symptoms	F1 score-between 57 and 87%
[4]	BI-LSTM, CNN, hybrid	Internet movie dataset(IMDB)	Text Classification	Word2Vec	Sequence of words	F1:0.90 Precision:0.90 Recall:0.89
[55]	Bi-LSTM CRF	English legal contracts(3500-labelled,750,000-unlabelled contracts)	Legal contract entity extraction	Word embedding	Contract entity information	F1:0.80
[84]	CNN	ICDAR 2019 SROIE data set	Grid text with feature-specific spatial and semantic knowledge	Word2Vec	1000 whole scanned receipt images	Average Precision: 85–90%
[92]	RNN, Bi-LSTM	PubMed	Biomedical named entities (BioNER) recognition and extraction	Word2Vec	Biomedical entities information extraction	F1-score: 67.2%

Table 7 (continued)

Refs.	Techniques used	Dataset used	Description	Feature extraction methods used	Domain	Performance
[93]	Faster-RCNN	PDF,scanned docs of invoices,RVL-CDIP	Financial documents and invoices data extraction	–	Financial business	Precision-66.80%
[94]	SVM, LR,RF	Amazon Foods product reviews	NOSQL queries data extraction	TF-IDF	NoSQL databases	SVM-80% RF-78%
[95]	RPA,mBERT, XLM-ROBERTa,InfoXLM, Layout XLM	BRC(Business Registration Certificates) BL(Bill of Loading)	Business product documents data extraction	–	Financial	F1 Score- BRC-97% BL-81%
[106]	Shelf OCR engine, Bidirectional attention complementation mechanism(BiACM)	FUNSD, RVL-CDIP	Language-independent layout transformer for structured document understanding	–	Scanned printed documents	FUNSD- Precision—0.87 Recall—0.89 F1—0.88 RVL-CDIP- Accuracy- 96.17%
[107]	Faster R-CNN	FUNSD SROIE	Model interacts between text & layout information across scanned doc images	Image and text embedding	Scanned printed documents	FUNSD- Precision—76% Recall—81% F1—79% SROIE- Precision—95% Recall—95% F1—95%
[108]	Transformer-based Visual Document Understanding (VDU)	IIT-CDIP RVL-CDIP	OCR free document understanding transformer	–	Printed docs, Receipts, Tickets, Business cards	RVL-CDIP Accuracy- 95.30%
[109]	Masked Detection Modeling (MDM) Extension of masked vision-language modeling (MVLM)	IIT-CDIP CORD images FUNSD	It helps in detecting & associating entity bounding boxes in visually rich documents & pre-training is used to learn entity detection	Entity Linking by Anchor Word Association	Structured information extraction from visually rich docs	CORD- Accuracy—94% FUNSD – Accuracy—84%

extraction methods are applied. To effectively perform automatic information extraction from business documents, the information extraction methodologies are classified in the current study. Selecting a common information extraction technique might be difficult for many application domains and unstructured documents with complicated or diverse formats.

It is recognized as the information extraction research's major problem. Traditional methods for the analysis of unstructured data have mostly relied on rule- or template-based approaches, which are time-consuming and costly to adopt. So, to process the unstructured documents, academics and organizations started seeking a solution that incorporates AI-enabled algorithms. AI-based solutions offer end-to-end automation solutions by combining CV and NLP capabilities with RPA or OCR workflows [77]. AI-based ML/DL and DL pre-trained models are useful to automate information extraction from documents are incredibly important for applications in banking, healthcare, finance, and other industries. A crucial issue for CV-based systems is automatically extracting useful and accurate information from unstructured documents. Automated information extraction techniques are used in a variety of application domains for various goals [85]. Rule-based approach RPA, Template-based approach, OCR NLP-based approaches like NER, Knowledge Graphs, and Semantic segmentation-based approaches used for mixed data recognition and classification are discussed in Sect. 3. All the literature with merits and demerits of each are discussed in detail in the study. Tables 6, 7 shows the existing work with Machine learning or deep learning algorithms used.

Recently, DL approaches have gained popularity in this field. Deep Learning is a great approach for solving the automated information extraction problem since it automatically learns new features. DL models are capable of completing tasks from beginning to end, including automated feature extraction and final classification. The main challenge in deep learning today is choosing or developing the best neural network architecture for a given task, choosing the optimal cost function, and accumulating a large amount of training data.

RQ.2. What are the domains and applications that utilize unstructured data processing?

Nowadays in every domain, unstructured documents are generated on a daily basis. It becomes very essential to process this large amount of data and generate a profit out of it. Various domains where unstructured document processing is required with existing research work are discussed in detail in Sect. 3. Table 6 shows the existing research work done in various application domains. The financial industry produces a substantial quantity of data, including client information, financial product logs, transaction data, and external information from sources like social media and websites that may be utilized to help decision-making. Similarly in healthcare, the majority of data is in the form of unstructured data which includes the medical history of the patient, hospital admission forms, discharge summaries, and prescriptions of patients. As very little work has been done in mixed text data recognition in the different domains so there is a large amount of scope present to work in the domains of healthcare, finance, insurance, banking, marketing, share market, academics, etc. It has been observed that, still there is no proper framework or solution available in the existing work that will focus on the digitization of

multimodal unstructured documents, it is apparent that researchers working in this field of study have a large amount of opportunities at their disposal.

RQ.3. What are the various data preprocessing techniques associated with unstructured document processing?

On a daily basis, unstructured documents are generated in various domains. If this large amount of generated unstructured data is retrieved and analyzed well can be useful for businesses to grow and make important future decisions based on analyzed data. After the data collection/generation step, various pre-processing operations are carried out on the data at the preprocessing stage of data processing to prepare it for subsequent processing and analysis. Stop-word elimination, the conversion of all text data to lowercase, the elimination of non-alphanumeric characters, the elimination of blank rows, and the elimination of joint words are a few of these. The dataset's noise, such as inaccurate text, will be manually deleted. Another upcoming objective in this domain is to create high-quality datasets that utilize a few advanced pre-processing techniques and are openly used by other researchers. Different types of preprocessing techniques are discussed in detail with literature in Sect. 3. There are many opportunities for researchers to work in this area with quality preprocessing techniques revolution to achieve better results. To automate the unstructured documents data preprocessing of input data is very important. Researchers can produce more accurate findings from it if they have access to high-quality datasets.

RQ.4.What are the many issues that information extraction algorithms face while processing data from unstructured documents?

The key issues and challenges faced during the unstructured document processing are outlined here.

The review performed in this literature shows the following research gaps identified while processing unstructured documents:

- Existing techniques do not focus on recognizing multiple modality types of data in unstructured documents with good accuracy.
- There has not been much research done on AI-based methods for automatically extracting crucial info from unstructured materials.
- Publicly available datasets are task-specific and of poor quality. A fresh dataset that represents actual situations is thus required.
- It might be challenging to extract the relevant information from handwritten papers since many of them include confused wording, messy handwriting, and different writing styles.
- When there is noise, like a bounding box on a form page, handwriting recognition algorithms have a harder time identifying the writer's handwriting. The risk of producing summaries that convey different interpretations from the original clinical notes recorded by doctors makes abstractive summarization a difficult process in the medical field.

- Creating informative summaries using more complex language and knowledge engineering techniques, personalizing summaries, adjusting to new sub-domains, developing evaluation scenarios that simulate real-world situations, and integrating summarization technology into practical applications like the clinical workflow are among the challenges [104].
- The selection of feature sets for diversified scripts in a multilingual environment is a complex task.

Since each person's written characters are unique, some characters have remarkably similar shapes, other characters have disconnected or distorted letters, written characters have a range of thicknesses, and various scanners are used, handwriting detection presents many challenges.

Limitations in the area of unstructured document processing are as follows-

- **Data recognition**: in the case of Multimodal data, which consists of handwritten text, machine-printed text, symbols, checkboxes, and radio buttons data recognition which can the accuracy in form data recognition, is one of the underexplored areas in semi-structured document processing.
- **Manual data entry**: manual data entry of handwritten filled data is required in most industries. No proper automation is present for the digitization of data.
- **Page orientation**: OCR-based approaches are challenged by the skewness and incorrect page orientation that are present in the scanned document.
- **Inconsistent text**: words are often scrambled in handwritten texts, and improving one's handwriting is a difficult process. To correctly understand the term based on the overall context of the text, OCR techniques may be utilized in conjunction with NLP and ML algorithms.
- **Glared documents**: when captured with cell phone cameras, several semi-structured documents, including patient admission forms, leave applications, and e-governance forms, create glare images, increasing the likelihood of inaccurate data extraction [105].
- **Unanalyzed data**: a large amount of unstructured data is generated daily that remains unanalyzed, if properly evaluated and summarized, can be beneficial in making decisions in a variety of business disciplines.
- **Bounding box problem**: the bounding box of the document makes handwriting recognition challenging. Many unstructured documents have irregular spaces between words in a document.

RQ. 5 What types of unstructured data processing datasets are available?

The model has been trained using a variety of datasets for specific information extraction tasks or unstructured document analysis tasks, according to the authors. It is essential to collect the proper dataset that includes sufficient amounts and high-quality data for AI-based model training and testing to provide high-quality research outcomes.

Table 8 Overview of datasets

Refs.	Dataset	Application domain	Size of dataset	Balanced/ unbalanced	Annotated/ not annotated	Access type free/not free	Multimodality
[17, 36]	NIST	Handwritten images of forms	Handwritten sample forms from 3600 writers 810,000 char images of forms	√	√	√	×
[12, 68]	EMNIST	Handwritten char dataset	47 classes, 131600 character data	√	√	√	√
[96–99]	AHCD	Arabic handwritten text data	16,800 char written by 60 participants	√	√	√	×
[65, 80, 81]	MIMIC-III	Healthcare	53,000-patient admission notes 812,128-nursing notes 52,746-discharge summaries	√	√	√	√
[4, 81, 100]	IMDB	Films, television series	50 K movie reviews	×	√	√	×
[11, 79]	PubMed	Medicine	19,717 Metadata about scientific publications	×	√	√	×
[77, 78]	PRImA &UvA	Financial color doc images	Printed doc layout's scanned color docs	√	√	√	×
[74–76]	HCL-2000	Chinese handwritten data	3,755 Chinese handwritten character photos	×	√	√	×
[6, 71–73]	CEDAR	Healthcare	English text lines composed by 200 writers	×	√	√	√
[69, 70]	PAW	paraphrase identification dataset	83,056 images 5,50,000 diff words	×	√	√	×
[20, 66, 67, 101]	I AM	Mixed handwritten and printed images docs	1539 forms of handwritten English text	√	√	√	√
[36]	MOST	Handwritten doc images dataset	10,000- testing sample images 60,000-training sample images	√	√	√	×
[58, 102]	ICDAR-2019 SROIE	Receipt dataset	1000 printed scanned images receipt	×	√	√	×
[103]	MIDV-500	Mobile identity document video dataset	500 video clips for 50 different forms of ID, including 13 licenses, 14 passports, and 17 ID cards	√	√	√	×

The largest proportion of the datasets used in the research studies depend on language and domain. The literature research indicates that building a general-purpose information extraction model requires a comprehensive dataset with common, general-purpose, and normalized entity annotations. The current information extraction algorithms have difficulties due to data scarcity, including variation in language vocabulary, a wide range of acronyms, a wide range of rules to remove linguistic ambiguity, and many languages. The most representative, complete, and diverse publically accessible dataset, which is domain agnostic, is not present in the existing literature. Another problem identified in the study is the magnitude of the dataset. The scarcity of a significant labeled corpus is another issue for this study field. Datasets may also contain noisy, skewed, or missing data values. Such datasets are used to create extractions that are less useful and useless.

The majority of self-built/custom document databases are sensitive or confidential because they include personal data about people, organizations, or businesses. Table 8 illustrates various types of datasets, application domains, size of the dataset, access type, whether the dataset is balanced or imbalanced, multimodality, and annotated in detail. Further Table 9 gives an overall summary of the type of input data used, preprocessing and feature extraction techniques used, and different machine learning and deep learning approaches used.

Conclusion & future directions

To find opportunities for improvements in this field, the SLR explores current information extraction methods for unstructured document processing. To conduct the PRISMA approach of literature search for this SLR, the guidelines suggested by Kitchenham and Charters were followed. A total of 68 research papers were ultimately chosen to address the study topics based on inclusion, exclusion, and quality evaluation criteria.

The scholars' contributions to publications have increased significantly during the past nine years, it may be said. It illustrates the significance and developments in this field of research study.

The SLR analyses and assesses automated information extraction from unstructured documents which includes mixed text documents processing research in-depth by:

- Identifying the issues that exist with current information extraction methods for handling the processing of unstructured documents.
- Recognizing the requirement for creating good quality, freely accessible unstructured datasets.
- Determining the procedures for validating data that are accessible for evaluating data quality.
- Investigating this field of study for many application areas, including the extraction of banking documents, clinically named entities, legal clauses, invoices, academic data, Social media data, etc.
- Examining text detection, preprocessing, feature extraction, classification, and recognition techniques.
- Outlining the difficulties and potential areas for future study in automated information extraction from unstructured documents.

Table 9 Overview of unstructured document processing approaches with different techniques used

Refs.	Type of information		Pre-processing techniques										Feature extraction techniques			
	Printed text	Handwritten text	Binarization	Noise removal	Normalization	Slant removal	Tokenization	Stemming	Stop word removal	Morphological O.angular alignment	Angular alignment	Gabor filter	Bag of words	Word2Vec	BERT	Scale invariant feature tranform
[4]	√		√		√									√		
[12]		√		√		√							√			
[13]	√				√					√						
[18]		√	√	√						√						
[20]	√	√	√		√						√		√			
[22]	√	√	√										√			
[23]		√	√									√		√		
[25]	√		√	√											√	
[26]	√	√	√	√		√										
[31]	√		√							√				√		
[32]	√						√	√	√				√			
[34]	√															
[36]	√													√		
[39]	√															
[42]	√		√		√									√	√	
[48]				√							√			√		
[50]	√	√		√							√					
[55]	√			√							√			√		√
[80]	√		√	√									√			
[84]	√		√			√					√			√		
[87]		√	√	√									√			
[90]	√	√	√	√							√					
[92]	√		√	√										√		

Table 9 (continued)

Refs.	Machine learning approaches					Deep learning approaches									
	SVM	Naïve Bayes	KNN	Random forest	CRF	CNN	RNN	LSTM	Bi-LSTM	CRF	BERT	MCRNN	Alexnet	Unet	VGG16
[4]						√			√						
[12]												√			
[13]	√					√									
[18]						√									
[20]								√							
[22]	√														
[23]	√		√			√									
[25]							√		√						√
[26]												√		√	
[31]			√												
[32]					√										
[34]							√				√				
[36]						√	√								
[39]								√		√					
[42]					√				√		√				
[48]															
[50]		√	√	√									√		
[55]									√	√					
[80]						√	√								
[84]						√									
[87]						√									
[90]						√									
[92]							√								

Table 10 Challenges and future directions

Sr. No	Challenges	Future directions
1	Less accuracy while processing multi-format documents	Create a hybrid model with great processing flexibility for many formats
2	Supports limited file types for processing	Future research in this area may go beyond TIFF, JPEG, PDF, and other file formats
3	Fixed template dependent approach	Design and implement an AI-based OCR template-free approach to deal with it
4	Low outcome as it depends upon the quality of input data	Advancements in pre-processing techniques using appropriate text segmentation methods give improved results
5	Insufficient labeled & balanced data	Making high-quality datasets that adhere to real-world standards and have a logical layout
6	Less performance of automation techniques	To meet the criteria for a template-free end-to-end solution, a hybrid model incorporating OCR, RPA, NER, semantic segmentation, and AI-based ML/DL approaches is being developed
7	Fixed rule-based and template-based approach	Implement a hybrid approach of Deep learning pre-train models with OCR/RPA/Semantic segmentation/NER using AI
8	A lot of human workforces for performing repetitive tasks	Scaling automation across all business processes through the use of RPA and business process management
9	No proper multimodal data processing solutions	Combine the visual characteristics, entity semantics, and handwritten, printed, or data with information extraction methods
10	Data preprocessing is primarily critical	Investigating different data quality enhancement strategies that enhance the effectiveness of information extraction methods

The outcome of this literature demonstrates that the researchers focus their attention on the hybrid model, which combines diverse approaches like ML/DL and CV/NLP, because of its effectiveness in managing large amounts of unstructured data. It is apparent through this study that the complex and heterogeneous patterns of unstructured texts are not as well addressed by template-free approaches. Table 10 shows the various challenges in the field of mixed text data classification and recognition with possible solutions to overcome these issues.

Therefore, developing an AI-based model independent of templates for text data detection and extraction of pertinent information from unstructured texts with a complex and changing layout is a potential prospect. For autonomous information extraction from unstructured documents, the study identifies areas that need more research in the disciplines of OCR, RPA, NER, segmentation-based techniques, Large language models, and AI-based techniques. This work will have a variety of rapid and widespread effects on how automatic information extraction is adopted in the legal, financial, and medical fields. This literature research indicates that while automatic information extraction has started to be employed in several different industries, more development is still needed to successfully extract information from complex and varied unstructured materials. To strengthen their position in the deployment of automation, successful enterprises may leverage their position to create a first-mover advantage in information extraction from unstructured documents employing end-to-end automation.

Abbreviations

AI	Artificial intelligence
CV	Computer vision
ML	Machine learning
SLR	Systematic literature review
OCR	Optical character recognition
KG	Knowledge graph
RPA	Robotic process automation
NER	Named entity recognition
TF-IDF	Term frequency inverse document frequency
RNN	Recurrent neural network
LSTM	Long short-term memory
ANN	Artificial neural network
HMM	Hidden Markov model
HTR	Handwritten text
HER	Electronic health record
SVM	Support vector machine
CNN	Convolution neural network
KNN	K-nearest neighbor
CRF	Conditional random fields
LLM	Large Language models

Author contributions

Conceptualization, SM and SP; methodology, SM and SP; software, SM; validation, SM, SP, KK; formal analysis, SM; investigation, SM; resources, SM; data curation, SM and SP; writing—original draft preparation, SM; writing—review and editing, SP, KK, and LS; visualization, SM; supervision, SP; project administration, SP, LS; funding acquisition, KK.

Funding

This study was funded by Research Management Centre of Multimedia University.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 23 January 2024 Accepted: 17 June 2024
Published online: 05 July 2024

References

1. Adnan K, Akbar R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0254-8.
2. Eberendu AC. Unstructured data: an overview of the data of Big Data. Int J Comput Trends Technol. 2016;38(1):46–50. https://doi.org/10.14445/22312803/ijctt-v38p109.
3. Baviskar D, Ahirrao S, Potdar V, Kotecha K. Efficient automated processing of the unstructured documents using artificial intelligence: a systematic literature review and future directions. IEEE Access. 2021;9:72894–936. https://doi.org/10.1109/ACCESS.2021.3072900.
4. Jang B, Kim M, Harerimana G, Kang SU, Kim JW. Bi-LSTM model to increase accuracy in text classification: combining word2vec CNN and attention mechanism. Appl Sci. 2020. https://doi.org/10.3390/app10175841.
5. Mehta N, Doshi J. A review of handwritten character recognition. Int J Comput Appl. 2017;165(4):37–40. https://doi.org/10.5120/ijca2017913855.
6. Memon J, Sami M, Khan RA, Uddin M. Handwritten Optical Character Recognition (OCR): a comprehensive Systematic Literature Review (SLR). IEEE Access. 2020;8:142642–68. https://doi.org/10.1109/ACCESS.2020.3012542.
7. Bach MP, Krstič Ž, Seljan S, Turulja L. Text mining for big data analysis in financial sector: a literature review. Sustain. 2019. https://doi.org/10.3390/su11051277.
8. Adnan K, Akbar R. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. Int J Eng Bus Manag. 2019;11:1–23. https://doi.org/10.1177/1847979019890771.
9. Syed R, et al. Robotic process automation: contemporary themes and challenges. Comput Ind. 2020;115:103162. https://doi.org/10.1016/j.compind.2019.103162.
10. Al-Moslmi T, Gallofre Ocana M, Opdahl AL, Veres C. Named entity extraction for knowledge graphs: a literature overview. IEEE Access. 2020;8:32862–81. https://doi.org/10.1109/ACCESS.2020.2973928.
11. Wang Y, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34–49. https://doi.org/10.1016/j.jbi.2017.11.011.
12. Chiney A, et al. Handwritten data digitization using an anchor based Multi-Channel CNN (MCCNN) trained on a hybrid dataset (h-EH). Procedia CIRP. 2021;189:175–82. https://doi.org/10.1016/j.procs.2021.05.095.
13. Fanany DMI. Handwriting recognition on form document using CNN-SVM. 2017; 3–5.

14. Kitchenham B. Guidelines for performing Systematic Literature Reviews in Software Engineering (Software Engineering Group, Department of Computer Science, Keele …. 2007.
15. Plamondon R, Srihari S. Online_Offline_2000.pdf. 2000.
16. SurShivanana I, Pathak K, Gagnani M, Shrivastava V, Mahesh TR, Madhuri SG. Text extraction and detection from images using machine learning techniques: a research review. Proceedings of the International Conference on Electronics and Renewable Systems, ICEARS 2022. 2022; 1201–1207. https://doi.org/10.1109/ICEARS53579.2022.9752274.
17. Sharma S, Gupta S. Recognition of various scripts using machine learning and deep learning techniques-a review. Proceedings of IEEE International Conference on Signal Processing, Computing and Control. 2021; 2021-Octob: 84–89. https://doi.org/10.1109/ISPCC53510.2021.9609404.
18. Shah A, Doshi N, Shah J, Goel K, Raut P. Extraction of handwritten and printed text from a form. J Phys Conf Ser. 2021. https://doi.org/10.1088/1742-6596/1831/1/012029.
19. Baviskar D, Ahirrao S, Kotecha K. Multi-layout unstructured invoice documents dataset: a dataset for template-free invoice processing and its evaluation using AI approaches". IEEE Access. 2021;9:101494–512. https://doi.org/10.1109/ACCESS.2021.3096739.
20. Medhat F et al. TMIXT: a process flow for Transcribing MIXed handwritten and machine-printed Text," Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018. 2019; 2986–2994. https://doi.org/10.1109/BigData.2018.8622136.
21. Zhu M, Cole JM. PDFDataExtractor: a tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. J Chem Inf Model. 2022;62(7):1633–43. https://doi.org/10.1021/acs.jcim.1c01198.
22. Zagoris K, Pratikakis I, Antonacopoulos A, Gatos B, Papamarkos N. Distinction between handwritten and machine-printed text based on the bag of visual words model. Pattern Recognit. 2014;47(3):1051–62. https://doi.org/10.1016/j.patcog.2013.09.005.
23. Hamida S, Cherradi B, Ouajji H. Handwritten Arabic Words Recognition System Based on HOG and Gabor Filter Descriptors. 2020 1st Int. Conf. Innov. Res. Appl. Sci. Eng. Technol. IRASET. 2020; 1–4. https://doi.org/10.1109/IRASET48871.2020.9092067.
24. Boiangiu CA, Dinu OA, Popescu C, Constantin N, Petrescu C. Voting-based document image skew detection. Appl Sci. 2020;10(7):1–12. https://doi.org/10.3390/app10072236.
25. Xue W, Li Q, Xue Q. Text detection and recognition for images of medical laboratory reports with a deep learning approach. IEEE Access. 2020;8:407–16. https://doi.org/10.1109/ACCESS.2019.2961964.
26. Patil S, et al. Enhancing optical character recognition on images with mixed text using semantic segmentation. J Sens Actuator Netw. 2022;11(4):63. https://doi.org/10.3390/jsan11040063.
27. Zaman G, Mahdin H, Hussain K, Atta-Ur-Rahman. Information extraction from semi and unstructured data sources: a systematic literature review. ICIC Express Lett. 2020;14(6):593–603. https://doi.org/10.24507/icicel.14.06.593.
28. Su, Sayyad, Shirabad, Matwin, Huang. Discriminative Multinominal Naive Bayes for Text Classification. http//www.site.uottawa.ca/~stan/csi5387/DMNB-paper.pdf. 30–11–2012.
29. Mahadevkar SV, et al. A review on machine learning styles in computer vision - techniques and future directions. IEEE Access. 2022;10(September):107293–329. https://doi.org/10.1109/ACCESS.2022.3209825.
30. Cao W, Zhou C, Wu Y, Ming Z, Xu Z, Zhang J. Research progress of zero-shot learning beyond computer vision. Lect Notes Comput Sci. 2020;12453:538–51. https://doi.org/10.1007/978-3-030-60239-0_36.
31. Sahare P, Dhok SB. Multilingual character segmentation and recognition schemes for indian document images. IEEE Access. 2018;6:10603–17. https://doi.org/10.1109/ACCESS.2018.2795104.
32. Kanya N, Ravi T. Named entity recognition from biomedical text -an information extraction task. ICTACT J Soft Comput. 2016;6(4):1303–7. https://doi.org/10.21917/ijsc.2016.0179.
33. Chowdhury S, Schoen MP. Research Paper Classification using Supervised Machine Learning Techniques. 2020 Intermt. Eng. Technol. Comput. IETC 2020, no. July 2021. 2020, https://doi.org/10.1109/IETC47856.2020.9249211.
34. Steinkamp JM, Bala W, Sharma A, Kantrowitz JJ. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. J Biomed Inform. 2020;102:103354. https://doi.org/10.1016/j.jbi.2019.103354.
35. Stewart S, Barrett B. Document image page segmentation and character recognition as semantic segmentation. ACM Int Conf Proc Ser. 2017. https://doi.org/10.1145/3151509.3151518.
36. Chernyshova YS, Sheshkus AV, Arlazarov VV. Two-Step CNN framework for text line recognition in camera-captured images. IEEE Access. 2020;8:32587–600. https://doi.org/10.1109/ACCESS.2020.2974051.
37. Artaud C et al. Receipt Dataset for Fraud Detection To cite this version : HAL Id : hal-02316349 Receipt Dataset for Fraud Detection. 2019.
38. Wu C, et al. Extra - 3. IEEE Access. 2019;7:117227–45.
39. Yang J, Liu Y, Qian M, Guan C, Yuan X. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. Appl Sci. 2019. https://doi.org/10.3390/app9183658.
40. Martins P, Sa F, Morgado F, Cunha C. Using machine learning for cognitive Robotic Process Automation (RPA). Iber Conf Inf Syst Technol Cist. 2020. https://doi.org/10.23919/CISTI49556.2020.9140440.
41. Laubrock J, Dunst A. Computational approaches to comics analysis. Top Cogn Sci. 2020;12(1):274–310. https://doi.org/10.1111/tops.12476.
42. Francis S, Van Landeghem J, Moens MF. Transfer learning for named entity recognition in financial and biomedical documents. Inf. 2019;10(8):1–17. https://doi.org/10.3390/info10080248.
43. Huang K et al. Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. 2020; 94–100. https://doi.org/10.18653/v1/2020.clinicalnlp-1.11.
44. Weng W, Zhu X. INet: convolutional networks for biomedical image segmentation. IEEE Access. 2021;9:16591–603. https://doi.org/10.1109/ACCESS.2021.3053408.
45. Desai S, Singh A. Optical character recognition using template matching and back propagation algorithm. Proc Int Conf Inven Comput Technol ICICT. 2016;2016:2016. https://doi.org/10.1109/INVENTIVE.2016.7830161.

46. Ye Y et al. A unified scheme of text localization and structured data extraction for joint OCR and data mining. Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018, no. 1. 2019; 2373–2382. https://doi.org/10.1109/BigData.2018.8622129.
47. Mehul G, Ankita P, Namrata D, Rahul G, Sheth S. Text-based image segmentation methodology. Procedia Technol. 2014;14:465–72. https://doi.org/10.1016/j.protcy.2014.08.059.
48. Saba T, Almazyad AS, Rehman A. Language independent rule based classification of printed & handwritten text (Classification of Printed & Handwritten Text). 2015 IEEE Int. Conf. Evol. Adapt. Intell. Syst. EAIS 2015. 2015. https://doi.org/10.1109/EAIS.2015.7368806.
49. Reul C, et al. OCR4all-An open-source tool providing a (semi-)automatic OCR workflow for historical printings. Appl Sci. 2019. https://doi.org/10.3390/app9224853.
50. Tarawneh AS, Hassanat AB, Chetverikov D, Lendak I, Verma C. Invoice classification using deep features and machine learning techniques. 2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2019 - Proc., no. June. 2019; 855–859. https://doi.org/10.1109/JEEIT.2019.8717504.
51. Pitou C, Diatta J. Textual information extraction in document images guided by a concept lattice. Int Conf Concept Lattices Their Appl. 2016;CLA2016:325–36.
52. Sidhwa H, Kulshrestha S, Malhotra S, Virmani S. Text Extraction from Bills and Invoices. Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018. 2018; 564–568. https://doi.org/10.1109/ICACCCN.2018.8748309.
53. Kofax. Five case studies to inspire your intelligent automation strategy. 2019.
54. Šimek D, Šperka R. How Robot/human orchestration can help in an hr department: a case study from a pilot implementation. Organizacija. 2019;52(3):204–17. https://doi.org/10.2478/orga-2019-0013.
55. Chalkidis I, Androutsopoulos I, Michos A. Extracting contract elements. In: ICAIL '17: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. 2017. p. 19–28. https://doi.org/10.1145/3086512.3086515
56. Shah P, Joshi S, Pandey AK. Legal clause extraction from contract using machine learning with heuristics improvement. 2018 4th Int. Conf. Comput. Commun. Autom. ICCCA 2018. 2018; 1–3. https://doi.org/10.1109/CCAA.2018.8777602.
57. Sun Y, Mao X, Hong S, Xu W, Gui G. Template matching-based method for intelligent invoice information identification. IEEE Access. 2019;7:28392–401. https://doi.org/10.1109/ACCESS.2019.2901943.
58. Patel S, Bhatt D. Abstractive Information Extraction from Scanned Invoices (AIESI) using End-to-end Sequential Approach. 2020. http://arxiv.org/abs/2009.05728.
59. Chen Y, Argentinis E, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38(4):688–701. https://doi.org/10.1016/j.clinthera.2015.12.001.
60. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform. 2018;83(April):112–34. https://doi.org/10.1016/j.jbi.2018.04.007.
61. Mezghani A, Slimane F, Kanoun S, Kherallah M. Window-based feature extraction framework for machine-printed/handwritten and Arabic/Latin text discrimination. Proc. - 2016 IEEE 12th Int. Conf. Intell. Comput. Commun. Process. ICCP 2016. 2016; 329–335. https://doi.org/10.1109/ICCP.2016.7737168.
62. Agrawal G, Deng Y, Park J, Liu H, Chen Y-C. Building knowledge graphs from unstructured texts: applications and impact analyses in cybersecurity education. Information. 2022;13(11):526. https://doi.org/10.3390/info13110526.
63. Stauffer M, Fischer A, Riesen K. a novel graph database for handwritten word images. 2016; 3: 553–563. https://doi.org/10.1007/978-3-319-49055-7.
64. Peng R, Liu K, Yang P, Yuan Z, Li S. Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data. 2023http://arxiv.org/abs/2308.03107.
65. Gehrmann S, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS ONE. 2018;13(2):1–19. https://doi.org/10.1371/journal.pone.0192360.
66. Cheng L, Bing L, He R, Yu Q, Zhang Y, Si L. IAM: a comprehensive and large-scale dataset for integrated argument mining tasks. 2022; 1:2277–2287. https://doi.org/10.18653/v1/2022.acl-long.162.
67. Marti UV, Bunke H. The IAM-database: an English sentence database for offline handwriting recognition. Int J Doc Anal Recognit. 2003;5(1):39–46. https://doi.org/10.1007/s100320200071.
68. Cohen G, Afshar S, Tapson J, Van Schaik A. EMNIST: Extending MNIST to handwritten letters. Proc. Int. Jt. Conf. Neural Networks. 2017; 2017-May: 2921–2926. https://doi.org/10.1109/IJCNN.2017.7966217.
69. Bataineh B. A printed PAW image database of Arabic language for document analysis and recognition. J ICT Res Appl. 2017;11(2):199–211. https://doi.org/10.5614/itbj.ict.res.appl.2017.11.2.6.
70. Zhang Y, Baldridge J, He L. PAWS: Paraphrase adversaries from word scrambling. NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 2019; 1(2): 1298–1308.
71. Rexit A, Muhammat M, Xu X, Kang W, Aysa A, Ubul K. Multilingual handwritten signature recognition based on high-dimensional feature fusion. Information. 2022. https://doi.org/10.3390/info13100496.
72. Ajij M, Pratihar S, Nayak SR, Hanne T, Roy DS. Off-line signature verification using elementary combinations of directional codes from boundary pixels. Neural Comput Appl. 2021. https://doi.org/10.1007/s00521-021-05854-6.
73. Schenck EJ, Hoffman KL, Cusick M, Kabariti J, Sholle ET, Campion TR. Critical carE Database for Advanced Research (CEDAR): an automated method to support intensive care units with electronic health record data. J Biomed Inform. 2021;118:103789. https://doi.org/10.1016/j.jbi.2021.103789.
74. Yang W, Jin L, Liu M. Chinese character-level writer identification using path signature feature, DropStroke and deep CNN. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2015; 2015-Novem: 546–550. https://doi.org/10.1109/ICDAR.2015.7333821.
75. Li Y et al. Sentence-level Online Handwritten Chinese Character Recognition, vol. 1, no. 1. Association for Computing Machinery, 2021.
76. Zhang H, Guo J, Chen G, Li C. HCL2000 - A large-scale handwritten Chinese character database for handwritten character recognition. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2009; 286–290. https://doi.org/10.1109/ICDAR.2009.15.

77. Clausner C, Antonacopoulos A, Pletschacher S. Efficient and effective OCR engine training. Int J Doc Anal Recognit. 2020;23(1):73–88. https://doi.org/10.1007/s10032-019-00347-8.
78. Todoran L, Worring M, Smeulders AW. The UvA color document dataset. Int J Doc Anal Recognit. 2005;7(4):228–40. https://doi.org/10.1007/s10032-004-0135-2.
79. Tkaczyk D, Szostek P, Fedoryszak M, Dendek PJ, Bolikowski Ł. CERMINE: automatic extraction of structured metadata from scientific literature. Int J Doc Anal Recognit. 2015;18(4):317–35. https://doi.org/10.1007/s10032-015-0249-8.
80. Yi F, et al. Dual model medical invoices recognition. Sensors. 2019. https://doi.org/10.3390/s19204370.
81. Christou D. Feature extraction using Latent Dirichlet Allocation and Neural Networks: A case study on movie synopses. 2016. http://arxiv.org/abs/1604.01272.
82. Krishnan P, Jawahar CV. Generating synthetic data for text recognition. 2016. http://arxiv.org/abs/1608.04224.
83. Kassim MN, Jali SHM, Maarof MA, Zainal A, Wahab AA. Enhanced text stemmer with noisy text normalization for malay texts. Singapore: Springer Singapore; 2020.
84. Zhao X, Niu E, Wu Z, Wang X. CUTIE: learning to understand documents with convolutional universal text information extractor. 2019. http://arxiv.org/abs/1903.12363.
85. Pramanik R, Bag S. Shape decomposition-based handwritten compound character recognition for Bangla OCR. J Vis Commun Image Represent. 2018;50:123–34. https://doi.org/10.1016/j.jvcir.2017.11.016.
86. Lu Y. Handwritten capital letter recognition based on OpenCV. MATEC Web Conf. 2019;277:02030. https://doi.org/10.1051/matecconf/201927702030.
87. Vaidya R, Trivedi D, Satra S, Pimpale PM. Handwritten character recognition using deep-learning. Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018. 2018; 772–775. https://doi.org/10.1109/ICICCT.2018.8473291.
88. Agrawal N, Kaur A. An Algorithmic Approach for Text Recognition from Printed/Typed Text Images. Proc. 8th Int. Conf. Conflu. 2018 Cloud Comput. Data Sci. Eng. Conflu. 2018; 876–879. https://doi.org/10.1109/CONFLUENCE.2018.8442875.
89. Jindal A, Amir M. Automatic classification of handwritten and printed text in ICR boxes. Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014. 2014; 1028–1032. https://doi.org/10.1109/IAdCC.2014.6779466.
90. Najadat HM, Alshboul AA, Alabed AF. Arabic Handwritten Characters Recognition using Convolutional Neural Network. 2019 10th Int. Conf. Inf. Commun. Syst. ICICS 2019, no. September 2020. 2019; 147–151. https://doi.org/10.1109/IACS.2019.8809122.
91. Palm RB, Winther O, Laws F. CloudScan - a configuration-free invoice analysis system using recurrent neural networks. Proc. Int. Conf. Doc. Anal. Recognition, ICDAR. 2017;1: 406–413. https://doi.org/10.1109/ICDAR.2017.74.
92. Kang YS, Kayaalp M. Extracting laboratory test information from biomedical text. J Pathol Inform. 2013;4(1):23. https://doi.org/10.4103/2153-3539.117450.
93. Nicolaieff L, Kandi MM, Zegaoui Y, Bortolaso C. Intelligent document processing with small and relevant training dataset. 2022 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2022. 2022; 1–7. https://doi.org/10.1109/ISCV54655.2022.9806100.
94. Jose B, Abraham S. Intelligent processing of unstructured textual data in document based NoSQL databases. Mater Today Proc. 2023;80:1777–85. https://doi.org/10.1016/j.matpr.2021.05.605.
95. Cho S, Moon J, Bae J, Kang J, Lee S. A framework for understanding unstructured financial documents using RPA and multimodal approach. Electron. 2023;12(4):1–17. https://doi.org/10.3390/electronics12040939.
96. Altwaijry N, Al-Turaiki I. Arabic handwriting recognition system using convolutional neural network. Neural Comput Appl. 2021;33(7):2249–61. https://doi.org/10.1007/s00521-020-05070-8.
97. Ullah Z, Jamjoom M. An intelligent approach for Arabic handwritten letter recognition using convolutional neural network. PeerJ Comput Sci. 2022. https://doi.org/10.7717/peerj-cs.995.
98. Alheraki M, Al-matham R, Al-khalifa H.Handwritten Arabic Character Recognition for Children Writ- ing Using Convolutional Neural Network and Stroke Identification.
99. Albattah W. Applied sciences Standalone and Hybrid CNN Architectures. 2022.
100. ParEunjeongk L, Cho S, Kang P. Supervised paragraph vector: distributed representations of words, documents and class labels. IEEE Access. 2019;7:29051–64. https://doi.org/10.1109/ACCESS.2019.2901933.
101. Tej MS, Saradhi TV, Spandana M, Savya V. Hand witten text recognition using deep learning. Int J Res Appl Sci Eng Technol. 2022;10(4):84–9. https://doi.org/10.22214/ijraset.2022.41156.
102. Huang Z et al. ICDAR2019 competition on scanned receipt OCR and information extraction. Proc. Int. Conf. Doc. Anal. Recognition. ICDAR. 2019; 1516–1520. https://doi.org/10.1109/ICDAR.2019.00244.
103. Arlazarov VV, Bulatov KB, Chernov TS, Arlazarov VL. MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream. Comput Opt. 2019;43(5):818–24. https://doi.org/10.18287/2412-6179-2019-43-5-818-824.
104. Christian R, Christoph W, Maximilian N, Andreas B, Maximilian W, Uwe S. Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning, vol. 1, no. 1. Association for Computing Machinery, 2021.
105. Nikolaidis A, Strouthopoulos C. Robust text extraction in mixed-type binary documents. Proc. 2008 IEEE 10th Work. Multimed. Signal Process. MMSP 2008. 2008; 393–398. https://doi.org/10.1109/MMSP.2008.4665110.
106. Wang, Jiapeng, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. arXiv preprint. 2022. arXiv:2202.13669.
107. Xu Y et al. Layoutlm: Pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.
108. Kim G, et al. Ocr-free document understanding transformer. European Conference on Computer Vision. Cham: Springer Nature; 2022.
109. Liao H et al. Doctr: Document transformer for structured information extraction in documents. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Publisher's Note