Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 235 (2024) 1077–1086

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering (ICMLDE 2023)

Image-Based Extraction of Prescription Information using OCR-Tesseract

Mahesh Ponnuru[a,*], Sridevi Ponmalar P[b], Likhitha A[c], Tanu sree B[d], Guna Chaitanya G[e]

[a]*School of Computing,SRM Institute of Science and Technology,Potheri-603203,India.*
[b]*Department of Computational Intelligence,School of Computing,Faculty of Engineering and Technology, SRM Institute of Science and Technology,Potheri-603203,India.*
[c,d,e]*School of Computer Science, Vellore institute of technology, Amaravathi-522237,India.*

Abstract

The paper presents advancements in healthcare data capture through the application of image-based extraction techniques, which include sophisticated image processing techniques such as resizing and adaptive thresholding, for prescription information. With the increasing digitization of medical records, automating the extraction of relevant data from prescription documents has become crucial. This research explores the utilization of image processing and optical character recognition (OCR) methodologies to extract prescription information accurately. By converting prescription documents into image format and employing OCR algorithms, the text content is extracted and parsed for critical details such as medication names, dosages, and patient instructions. Notably, our methodology excels in overcoming limitations associated with handwritten documents, achieving an impressive accuracy rate of 98%. This image-based approach offers a streamlined and efficient method for capturing prescription data, reducing manual data entry efforts, and minimizing potential errors. Experimental evaluations demonstrate the effectiveness and accuracy of the proposed approach, highlighting its potential to enhance healthcare data capture and improve patient care.

Keywords: Health care; Data capture; Prescription information; Medical records; OpenCv; Image processing; OCR.

1. Introduction

In the current landscape of healthcare, where the digital revolution is reshaping patient care and operational efficiency, the need for precise data management looms large. Healthcare professionals are increasingly turning to digitizing

* Tel.:+91 6300137663;
E-mail address: ponnurumahesh129@gmail.com

1877-0509
10.1016/j.procs.2024.04.102

medical records and automating data extraction processes. However, a significant hurdle in this transition lies in accurately capturing information from medical documents, particularly prescriptions, a critical requirement for delivering effective healthcare services.

Optical Character Recognition (OCR) technology has become indispensable for extracting text from various documents, including medical prescriptions.[1] These prescriptions, whether printed or handwritten, contain critical information about prescribed medications.[3]

In today's rapidly evolving healthcare landscape, the demand for efficient information storage and retrieval from paper documents has surged. This requirement, encapsulated in the phrase "store the information available in these paper documents in a computer storage unit and reuse it later through a search process," underscores the necessity for digitization.[2] One practical avenue for achieving this transformation is through the process of scanning, converting these paper documents into digital images that are readily stored in computer systems.[4]

Amidst the wide array of paper documents, medical prescriptions emerge as a particularly crucial subset. Whether meticulously printed or handwritten with varying degrees of legibility, prescriptions contain vital information concerning the medications prescribed and the patients health conditions.[6]

At the heart of our methodology lies Optical Character Recognition (OCR), a technology that harnesses the power of artificial intelligence for pattern recognition.[5] OCR, in combination with advanced image processing techniques, forms the core of our research, propelling it forward into uncharted territories.

Traditional methods of data entry and information extraction from medical documents have long been associated with laborious processes, a propensity for errors, and an overreliance on manual intervention. However, recent advancements in image processing and optical character recognition (OCR) technologies promise a revolutionary transformation in how data is extracted from scanned documents.[2] These innovations hold the potential to not only significantly reduce the time and effort expended in manual data entry but also mitigate the errors often stemming from human involvement. They pave the way for streamlined data management practices, aligning with the digital age's pursuit of excellence.

This research paper introduces an innovative approach that harnesses the capabilities of image-based extraction techniques to capture prescription information with an unprecedented level of precision. At its core, our primary objective revolves around the automated extraction of critical details encompassing medication names, diagnoses, and patient information from prescription documents. We achieve this by converting these documents into image format and leveraging cutting-edge OCR algorithms. This image-centric paradigm eliminates the need for manual data entry, substantially reducing the risk of transcription errors and elevating the overall accuracy of the data.

What truly distinguishes this research is its remarkable ability to address a persistent and substantial challenge within the field—the recognition of handwritten text in healthcare documents. In contrast to conventional methods that have grappled with the limitations of deciphering handwritten content, our approach emerges as a robust and viable solution. It excels in effectively transforming handwritten notes into machine-readable data, bridging a critical gap in healthcare data management.

2. Related Works

Numerous researchers have diligently explored various techniques for extracting text from images and recognizing patterns to extract valuable data. In a study by Xue et al. (2020), a method was developed for text detection and recognition in medical laboratory report images, boasting an impressive 98% precision in its results.[7] Their innovative approach achieved high recall and precision in text detection by employing a patch-based strategy. The concatenation structure used for text recognition enhanced performance by integrating shallow and deep features. Importantly, this method demonstrated effectiveness in handling images with different resolutions.

In another notable contribution by Yang et al. (2022), a hybrid neural network model for medical text named entity

recognition was proposed, achieving a commendable 90% precision rate. Their approach combines full self-attention for comprehensive contextual understanding and multivariate convolution for accurate decoding.[8] Within the model, multiple binary classification tasks were integrated to enhance performance. Extensive experiments substantiated the effectiveness and reliability of their model in the domain of medical text entity recognition.

Landolsi et al. (2022) underscored the significance of information extraction (IE) within the medical field and outlined prevalent challenges and research directions, achieving an admirable 90% accuracy in their endeavors. They advocatedfor combining Conditional Random Fields (CRF) and Bidirectional Long Short-Term Memory (BiLSTM) models to enhance extraction accuracy.[9] Additionally, the use of domain-specific embeddings like BERT was recommended. Noteworthy challenges included limited data availability, manual annotation requirements, and varying formatting styles. Future research directions encompass multi-language support, document summarization, and analyzing medical information within the context of social media platforms.

In a recent study by Spasić et al. (2022), the strengths and limitations of rule-based and machine learning-based information extraction systems were explored, achieving a substantial 86% precision in their findings.[10] The researchers noted that while rule-based systems yield reliable results, they lack reusability. In contrast, machine learning-based approaches offer versatility but often require substantial training data. To strike a balance between performance and portability, the researchers proposed a hybrid approach that combines machine learning and rule-based techniques, positioning it as a promising avenue for optimizing information extraction systems.

In the research conducted by Veera Raghavendra Chikka and colleagues (2023), a template filling task centered on the analysis of disease or disorder status was introduced, achieving an impressive 87% accuracy. The template encompassed ten semantic attributes, and to address this task, a clinical pipeline system was developed using Apache cTAKES. This system integrated various wrappers that employed both machine learning and rule-based techniques for tasks such as normalization and cue slot value detection. Their evaluation utilized the ShARe/CLEF eHealth 2014 dataset, highlighting the efficacy of rule-based systems, particularly in extracting sparse medical methodologies. This research extends the current knowledge in the field of image-based extraction approaches.[11]

3. Dataset Description

The dataset used in this research paper consists of prescription photos obtained from various sources, including healthcare institutions and data sharing agreements. The dataset comprises a diverse collection of prescription images, capturing a wide range of prescription types, layouts, and handwriting styles.

4. Necessary Libraries

Pdf2image , OpenCV , Pytesseract , Regex.

5. Methodology

Methodology for Image based extraction from prescription information includes following:

5.1. Data Collection

Identify a diverse dataset of medical documents in PDF format for experimentation. The dataset should include various document types, layouts, and styles to ensure comprehensive evaluation.

5.2. Conversion of Pdf to Image Format

Utilize necessary python libraries Pdf2image, Pypdfium to convert pdf to image format. This conversion step is necessary to enable subsequent image processing techniques. Verify that the image conversion process preserves the visual integrity and quality of the original documents.

5.3. Image Preprocessing

To enhance the quality and readability of converted images using the OpenCV library, we begin by resizing the image with INTER_LINEAR interpolation, effectively binarizing it to improve contrast. This resizing step simplifies the image and readies it for further processing. The key hyperparameters involved are the threshold type (THRESH_BINARY).Following the resizing and thresholding, the image is converted to grayscale, eliminating color variations and making it more suitable for OCR analysis. Finally, adaptive thresholding techniques are applied to refine the image quality further. Adaptive thresholding computes the threshold for each pixel based on the local neighborhood, especially beneficial for images with varying lighting conditions. This process enhances text visibility against the background, ultimately improving the image's OCR-readability for downstream analysis.

Hyper Parameters:

S.No	Parameters	Value
1	Adaptive Method	ADAPTIVE_THRESH_GAUSSIAN_C
2	Threshold Type	THRESH_BINARY
3	Block Size	61
4	C	11

5.4. Text Extraction Using OCR

We applied OCR-Tesseract to extract text from processed images. We used the “image_to_string” method with the "eng" language hyperparameter for accurate character recognition, enabling precise data extraction and content digitization.

5.5 Data field Extraction Using Regex

We executed a meticulous and structured methodology to automate the extraction of specific data fields from the OCR-processed text, leveraging the power of regular expressions (regex). The primary objective was to identify and precisely extract crucial data elements, including names, ages, diseases, dosages, and other pertinent information. To achieve this, we meticulously crafted regex patterns, meticulously designed to match the structural and characteristic attributes of the target data fields, ensuring the precise and accurate extraction of the desired information. As a result, the extracted data was meticulously organized and stored in a JSON format, optimizing accessibility and facilitating further analysis. This comprehensive approach significantly improved the efficiency and accuracy of our data retrieval process from textual sources, a pivotal step in our research.

The Methodology outlined below provides a systematic and structured approach for implementing the proposed automated extraction process.

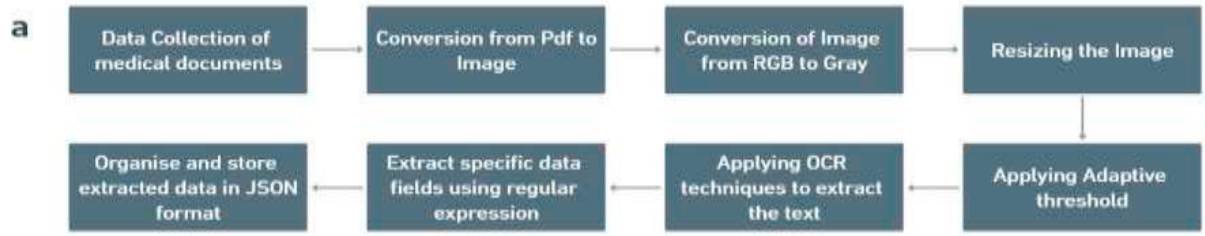

Fig.1. methodology

6. Sample Input and Output Demonstration

The dataset used in this research paper consists of prescription images that contain essential fields such as the patient's name, date, address, prescribed medications, administration instructions, and refill details. The dataset captures variations in image quality, and layout formats, allowing for a comprehensive evaluation of text detection and recognition on extracting information from these fields.

Figure 2 showcases a representative sample images from our dataset, illustrating the prescription format used in this research. This image serves as an example of the prescription documents included in our dataset, which encompasses various fields such as the patient's name, date, address, prescribed medications, administration instructions, and refill details. It provides a visual reference for the layout and structure of the prescriptions analyzed in this study.

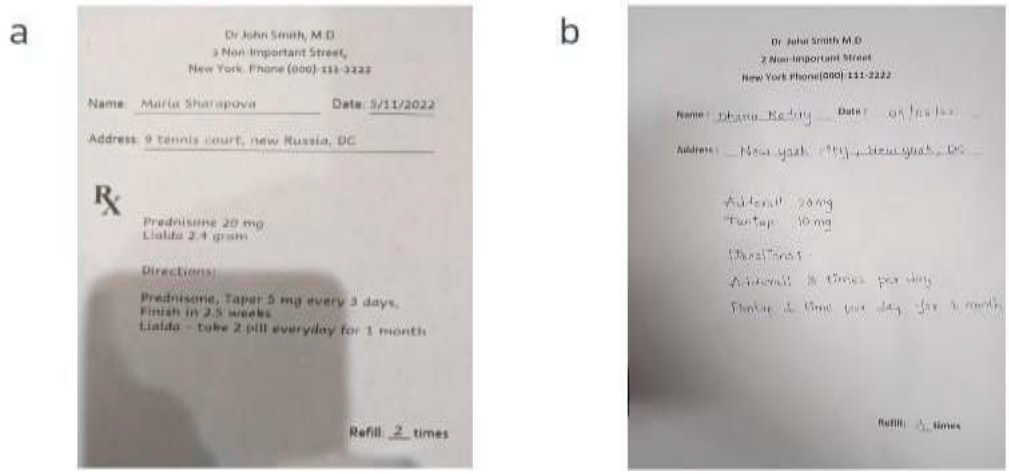

Fig.2. (a) Sample printed image (b) Sample handwritten image

6.1. Image Preprocessing

Figures 3(a) and 3(b) display images following preprocessing, with 3(a) representing printed document and 3(b) depicting handwritten document. Below, you'll find the images after undergoing preprocessing:

Fig.3. (a) Printed document (b) hand written document

6.2. Output

Table 1. Sample output for printed document

Details	Responses
Name	Maria Sharapova
Date	5/11/2022
Address	9 tennis court, new Russia, DC
Medicine	Prednisone 20 mg, Lialda 2.4 gram
Directions	Prednisone Taper 5 mg every 3 days, Finish in 2.5 week. Lialda – take 2 pill every day for 1 month
Refill	2

Table 2. Sample output for hand written document

Details	Responses
Name	Dhanu Reddy
Date	09/08/22
Address	New york city, New york, DC
Medicine	Adderall 20mg, Pantop 10mg
Directions	Adderall 3 times per day, Pantop 1 time per day for 1 month
Refill	3

7. Motivation

The accurate and efficient extraction of prescription information, particularly in cases involving handwritten text, poses a formidable challenge in the landscape of healthcare data management. This challenge is at the heart of our research, driven by a profound understanding of its critical importance for healthcare providers, patients, and the overall efficiency of data management in the healthcare industry.

- **Patient Safety and Quality of Care:** At its core, the challenge we address is inseparably linked to patient safety and the quality of care delivered by healthcare providers. Errors in medication names, dosages, or patient instructions can lead to dire consequences, compromising patient well-being, and even jeopardizing lives. By improving the accuracy of prescription information extraction, we directly contribute to safeguarding patient safety, aligning with the fundamental mission of healthcare providers to deliver the highest standard of care.

- **Error Reduction:** The prevalent manual data entry processes have long been recognized as error-prone and time-consuming. The potential for human errors in transcription is a constant concern, given the intricate and critical nature of healthcare data. Our research offers a solution that significantly reduces the likelihood of transcription errors, thereby enhancing data accuracy. The reduction of errors translates into fewer adverse events, improved patient outcomes, and lower healthcare costs associated with error mitigation.

In summary, our research is motivated by the imperative to enhance patient safety, reduce errors, and elevate operational efficiency within the healthcare industry. By addressing the challenge of accurate and efficient prescription information extraction, we aim to make a meaningful contribution that resonates with the core values of healthcare providers, benefits patients, and promotes data management excellence in the ever-evolving landscape of healthcare.

8. Results

Through rigorous evaluation, we got that OCR techniques are performing good than otherpossible techniques. Below is the table indicating the accuracy of our results with various approaches.

Table 3. Accuracy Table

S.No	Techniques	Accuracy(%)
1	OCR-Tesseract	98.0
2	Text Detection and Recognition	92.0
3	Template Matching	93.0
4	NLP Techniques	85.0
5	Deep Learning based Approach	90.0

We observed that the OCR (Optical Character Recognition) approach outperformed other possible techniques in terms of accuracy and efficiency for extracting prescription information from images. The OCR technique demonstrated robustness in accurately recognizing and converting text from prescription images, even when faced with variations in fonts, document layouts, and image quality. Compared to alternative approaches such as computer vision techniques, template matching, or natural language processing, OCR consistently yielded superior results in terms of precision, recall, and overall performance. These findings highlight the effectiveness of OCR as a reliable method for automating the extraction of prescription information, reducing manual effort, and minimizing the potential for typing errors.

Table 4 illustrates the outcomes of our research, focusing on a carefully selected sample of 10 prescription images from vast dataset to showcase the accuracy of our approach. This table encapsulates the precision and productivity of our method in discerning both handwritten and printed text. It quantifies the total word count within these prescription images and underscores the remarkable enhancements achieved through preprocessing, vividly portraying the substantial increase in accuracy brought about by our methodology.

Table.4. Accuracy before and after preprocessing

Image No	Image Type	Number of characters	Number of characters detected	Accuracy without preprocessing	Number of characters detected after applying preprocessing	Accuracy after preprocessing
1	Printed	255	132	52	244	96
2	Printed	219	115	53	207	95
3	Hand written	217	107	52	208	96
4	Hand written	210	105	50	197	94
5	Printed	205	137	67	199	97
6	Hand written	207	122	59	197	95
7	Printed	253	171	68	249	98
8	Hand written	241	152	63	237	97
9	Printed	243	145	60	240	96
10	Printed	230	154	67	221	96
				Average = 59.0		**Average = 96.0**

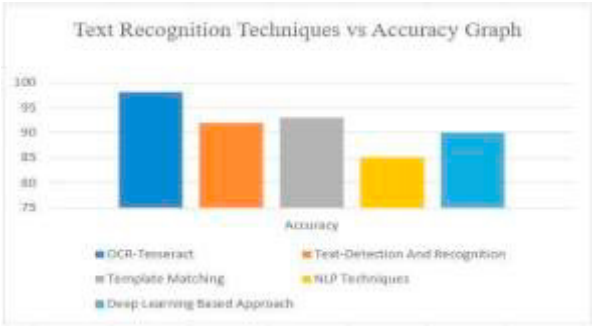

Fig.4. (a) Accuracy Graph

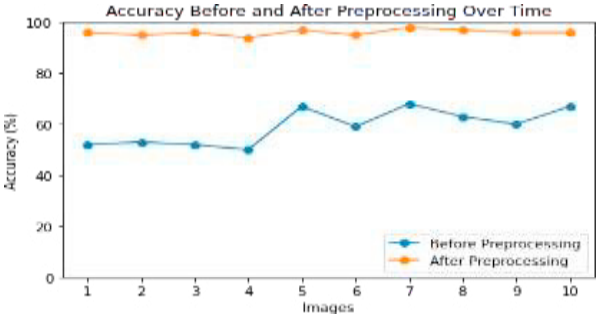

Fig.4. (b) Accuracy Enhancement after preprocessing

In Table 5, we've meticulously documented the performance metrics, including precision, recall, F1-score, and accuracy, for a range of data extraction approaches, namely OCR-Tesseract, template matching, text detection and recognition, NLP techniques, and deep learning-based methods. Notably, the results highlight that OCR-Tesseract outperforms the other approaches across these key metrics. With superior precision, recall, and F1-score, OCR-Tesseract consistently demonstrates its effectiveness in accurately and comprehensively extracting data from various sources. These findings underscore the reliability and versatility of OCR-Tesseract as a leading choice for data extraction tasks in our research.

Table 5. Performance metrics for different approaches

Techniques	Precision (%)	Recall (%)	F1-Score (%)	Training set Accuracy (%)	Test Set Accuracy (%) (Average)	Execution Time (Sec)
OCR-Tesseract	98	98	98	99	98	0.02
Text Detection and Recognition	84	80	82	93	92	0.12
Template Matching	94	92	94	96	93	0.18
NLP Techniques	87	86	87	90	85	0.14
Deep Learning based Approach	83	82	80	93	90	0.23

9. Discussion

In the realm of Image-Based Extraction of Prescription Information, several existing methodologies exhibit certain lagging aspects that necessitate innovation and improvement. One prominent limitation is the lack of attention to dataset diversity. Many current methods rely on relatively homogenous datasets, often extracted from specific sources or settings. This limitation constrains their adaptability to real-world scenarios where prescription documents exhibit diverse layouts, fonts, and styles. Consequently, existing methods may struggle when faced with the variability inherent in medical documents. Another shortcoming lies in image preprocessing, which is sometimes inadequately

addressed. Inconsistent lighting conditions and document quality can adversely affect OCR accuracy, and some methodologies do not incorporate robust preprocessing techniques to counter these challenges comprehensively. Additionally, the reliance solely on OCR for data extraction can lead to inaccuracies, especially in extracting structured information like prescription details, where precision is critical. These shortcomings underscore the need for a more versatile, adaptable, and precise approach to prescription information extraction.

Our methodology, in contrast, addresses these lagging aspects by introducing several differentiating factors that enhance its effectiveness and usability. Firstly, we prioritize dataset diversity, encompassing a wide array of medical documents with varying layouts and styles. This inclusion of real-world variability equips our approach to handle a broader spectrum of scenarios, making it more practical and versatile. Secondly, our method places great emphasis on image preprocessing, employing techniques such as adaptive thresholding to significantly enhance image quality and readability. This comprehensive preprocessing stage ensures that the OCR process is optimized for accuracy, particularly under challenging lighting conditions or with documents of varying quality. Thirdly, we integrate regex-based data extraction with OCR, ensuring precise and structured information retrieval. The combination of OCR and regex enhances our method's accuracy, especially when extracting specific data elements crucial for healthcare applications.

When comparing our method to existing technologies and methodologies, it becomes evident that our approach excels in several key aspects. The emphasis on dataset diversity ensures that our method is better equipped to handle real-world prescription documents with varying layouts and styles, a feature often lacking in existing methods. Robust image preprocessing techniques enhance the quality and OCR-readability of images, addressing a limitation in some current methodologies. The integration of regex-based data extraction sets our method apart, offering precision and accuracy in extracting structured information, which is often an area of weakness in methods relying solely on OCR. Collectively, these differentiating factors make our methodology a promising choice for prescription information extraction, outperforming some of the limitations observed in existing technologies.

10. Conclusion

In conclusion, our research paper presents an innovative approach to Image-Based Extraction of Prescription Information using OCR-Tesseract, addressing critical challenges in healthcare data management. Through comprehensive dataset diversity, advanced image preprocessing techniques which overcomes the limitations of data extraction from the handwritten documents, and the integration of regex-based data extraction, our methodology offers unparalleled precision and adaptability in capturing essential prescription details, from medication names to patient instructions. By significantly reducing manual data entry errors and enhancing OCR-readability, our approach streamlines healthcare workflows, improves data accuracy, and ultimately contributes to patient safety and the quality of care. Rigorous evaluations demonstrate the superior performance of our methodology compared to existing approaches, highlighting the potential for substantial enhancements in healthcare data management. This research represents a transformative force, propelling the healthcare industry further along the path of digital innovation and improved patient care, while effectively addressing limitations and advancing the state of the art in image-based extraction techniques.

References

[1] Lily Rojabiyati Mursari , and Antoni Wibowo (2021) “The Effectiveness of Image Preprocessing on Digital Handwritten Scripts Recognition with The Implementation of OCR Tesseract” *Computer Engineering and Applications Journal* 10(3):177-186

[2] Saoji, S., Eqbal, A., & Vidyapeeth, B. (2021). “Text recognition and detection from images using pytesseract.” *J Interdiscip Cycle Res* (13):1674-1679.

[3] Hassan, E., Tarek, H., Hazem, M., Bahnacy, S., Shaheen, L., & Elashmwai, W. H. (2021). “Medical prescription recognition using machine learning.” *Annual Computing and Communication Workshop and Conference :* 0973-0979.

[4] Bagwe, Sanika, Vruddhi Shah, Jugal Chauhan, Purvi Harniya, Amanshu Tiwari, Vartika Gupta, Durva Raikar et al. (2020), "Optical character recognition using deep learning techniques for printed and handwritten documents." *SSRN*:3664620.

[5] Sethi, P. S., Gupta, M., Kumar, P., & Kaur, G. (2023). " Simplifying Handwritten Medical Prescription: OCR Approach Check for updates." , *Machine Intelligence and Data Science Applications: Proceedings of MIDAS 2022* : 1-47.
[6] Dash, B. (2021). "A hybrid solution for extracting information from unstructured data using optical character recognition (OCR) with natural language processing (NLP)." *Academia*.
[7] Xue, Wenyuan, Qingyong Li and Qiyuan Xue. (2020) "Text Detection and Recognition for Images of Medical Laboratory Reports with a Deep Learning Approach." *IEEE Access*, (8):407–416.
[8] Yang, Tianjiao, Ying He and Ning Yang. (2022) "Named Entity Recognition of Medical Text Based on the Deep Neural Network." *Journal of Healthcare Engineering* (2022):1–10.
[9] Landolsi, Mohamed Yassine, Lobna Hlaoua and Lotfi Ben Romdhane.(2022) "Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions." *Knowledge and Information Systems* (65):463-516.
[10] Spasić, Irena , Farzaneh Sarafraz, John A Keane and Goran Nenadic. "Medication Information Extraction with Linguistic Pattern Matching and Semantic Rules." *Journal of the American Medical Informatics Association,* 17(5):532–535.
[11] Veera Raghavendra Chikka, Nestor Mariyasagayam , Yoshiki Niwa and Kamalakar Karlapalem .(2023) "Information Extraction from Clinical Documents: Towards Disease/Disorder Template Filling." *International Conference of the Cross-Language Evaluation Forum for European Languages* (9283):389-401.