ClinTex AI Medical Prescription

Professor Dr. Neeraj Gupta[1] Ankush Kumar Poddar[1,2], Pragya Verma[3],

[1] Department of Information Technology, Panipat instute of engineering and technology

Abstract :

The ClinTex Medical Prescription Project focuses on digitizing and organizing medical prescription data to enhance healthcare accessibility and efficiency. This system is designed to process both handwritten and printed prescriptions, accurately extracting key information such as patient details, doctor identification, medicine names, dosage instructions, frequency, and duration. By applying structured data extraction techniques and integrating a medical database, ClinTex not only streamlines record-keeping but also suggests affordable generic alternatives along with pricing information. This helps patients make informed choices while reducing medication costs. The project ultimately aims to bridge the gap between manual prescriptions and smart digital healthcare solutions.

1 Introduction

In today's healthcare environment, managing prescriptions efficiently is crucial for both patients and medical professionals. Prescriptions are often handwritten and difficult to read or organize, leading to confusion, errors, or loss of important information. The ClinTex AI project was developed to address this challenge by creating a system that can automatically extract and structure key details from medical prescriptions.

This project focuses on converting scanned or photographed prescriptions into digital text, identifying relevant information such as patient details, doctor's name, prescribed medicines, dosage instructions, and more. By organizing this information into a clean and accessible format like Excel or CSV, ClinTex AI helps users maintain medical records with ease. Additionally, it can suggest affordable generic alternatives to branded medicines, making treatment more cost-effective.

The aim is to support digital transformation in the healthcare sector, reduce human error, and improve access to essential medical data using a practical and user-friendly solution.

1.1 Contribution

The ClinTex AI AI project contributes to healthcare digitization by offering a system that can automatically extract and organize information from medical prescriptions. The main contributions of this project are:

- Developed a tool that converts handwritten or scanned prescriptions into structured digital data.
- Enabled accurate extraction of key information such as patient name, doctor details, medicines, dosages, and frequency.
- Implemented a suggestion system for generic medicine alternatives to promote cost-effective treatment.
- Provided a downloadable format (CSV/Excel) for easy storage and access to prescription records.
- Created a foundation for integrating smart healthcare applications using extracted medical data.

1.2 Paper Organisation

This report is divided into several sections. The Introduction outlines the project purpose. The Background explains the need for prescription digitization. The Methodology describes the steps followed in system development. System Architecture presents the technical workflow. Features and Implementation detail key functions. Results and Discussion showcase the output and challenges, and the Conclusion highlights the final outcomes and future scope.

2 Background

Medical prescriptions are one of the most common and critical documents in the healthcare system. They contain important information about the patient's treatment plan, such as the name and dosage of medicines, frequency, and duration. However, most prescriptions are handwritten, which often makes them hard to read and difficult to store or analyze digitally. This not only causes inconvenience but can also lead to errors in medication dispensing.

With the rapid growth of digital healthcare solutions, there is a growing need to convert handwritten or scanned prescriptions into a structured digital format. Doing this manually is time-consuming and prone to mistakes. Therefore, an automated approach is necessary to handle large volumes of prescriptions efficiently.

The ClinTex AI project was initiated to tackle this issue by developing a tool that can read prescriptions from images, extract all relevant details, and convert them into a usable digital format. The idea is to simplify record-keeping for patients, improve accuracy for pharmacists, and create a foundation for data-driven healthcare services such as medicine recommendations, cost comparisons, and long-term health tracking.

This project bridges the gap between traditional paper-based prescriptions and modern digital healthcare by combining practical tools into one streamlined solution.

2.1 Literature Review

Digitizing healthcare processes has been an active area of research, especially in the field of prescription analysis. Several studies and technologies have been developed to address the challenges of reading, interpreting, and managing medical prescriptions, yet most systems still struggle with the complexity of handwritten notes and varying formats.

Previous research has shown that Optical Character Recognition (OCR) techniques like Tesseract and Google Vision API are effective in extracting text from printed and scanned medical documents. However, their accuracy drops significantly when dealing with cursive or poorly written handwriting, which is common

in prescriptions. To improve this, some approaches combine OCR with natural language processing (NLP) to filter noise and identify key fields such as medicine names, dosages, and instructions.

Studies have also explored the integration of medical knowledge bases to enhance the understanding of drug names and their compositions. For instance, certain models use pre-built drug dictionaries or APIs to validate and suggest corrections for extracted medicine names. This becomes particularly useful when providing generic alternatives to branded medications, making healthcare more affordable.

While many hospital systems have electronic health record (EHR) platforms, they often lack the ability to process external paper prescriptions. Research on document automation and smart health data extraction highlights the need for user-friendly tools that can convert physical documents into usable formats like CSV or Excel.

The ClinTex AI project builds upon these ideas by combining OCR, rule-based NLP, and drug mapping to form an end-to-end solution. Unlike earlier systems that focus only on recognition, ClinTex AI adds value by structuring the output, suggesting generic alternatives, and allowing for easy export of information, making it a practical tool for both patients and healthcare providers.

3 Methodology

The development of the ClinTex AI project followed a step-by-step approach that involved image processing, text extraction, information structuring, and data recommendation. The aim was to build a reliable system that can take a prescription image as input and return organized medical data in a digital format.

Step 1: Image Input

Users upload an image or scanned copy of a medical prescription through a simple interface. The system accepts various image formats like JPG, PNG, and PDF.

Step 2: Text Extraction (with Mathematical Insight)

Once the prescription image is uploaded, it is processed using an Optical Character Recognition (OCR) engine such as **Tesseract** or **EasyOCR**. These tools extract textual information from images using a combination of image processing, mathematical techniques, and deep learning models.

1. Image Preprocessing The image is first converted to grayscale to simplify analysis. The grayscale value is calculated using a weighted sum of RGB values:

$$\text{Gray} = 0.299R + 0.587G + 0.114B$$

Next, binarization is performed using Otsu's thresholding method, which finds the optimal threshold that minimizes intra-class variance:

$$\sigma_w^2(T) = q_1(T)\sigma_1^2(T) + q_2(T)\sigma_2^2(T)$$

where q_1, q_2 are the class probabilities and σ_1^2, σ_2^2 are the variances of the two classes (foreground and background).

2. Segmentation The binarized image is segmented into individual characters or words using projection profiles or connected component analysis. This involves summing pixel intensities across horizontal and vertical lines to detect character boundaries.

3. Feature Extraction Character shapes are transformed into numerical feature vectors using methods such as zoning (counting pixel density in grid zones) and image moments:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y)$$

where $f(x, y)$ is the intensity of the pixel at position (x, y).

4. Character Recognition with LSTM Tesseract uses Long Short-Term Memory (LSTM) neural networks for character recognition, allowing it to understand sequences and context in handwriting. The LSTM unit is defined as:

$$h_t = o_t \cdot \tanh(C_t)$$

with internal computations:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad \text{and} \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

5. Post-processing To improve accuracy, OCR output is passed through spelling correction algorithms using Levenshtein distance:

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \text{cost} \end{cases}$$

This measures the number of edits required to convert one word into another, helping correct OCR mistakes.

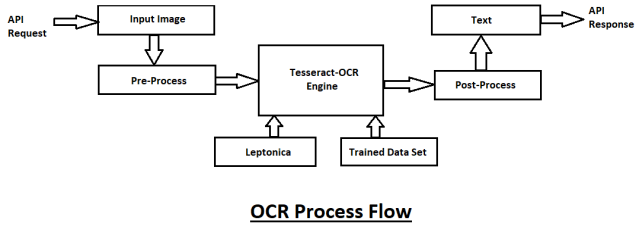

Figure 1. OCR processing workflow used in the ClinTex AI project. An input image is preprocessed and passed to the Tesseract-OCR engine, which, with the help of Leptonica and trained data, extracts text. The result is refined through post-processing and returned via API.

Step 3: Data Cleaning and Preprocessing

The extracted text often contains noise, such as extra spaces, symbols, and irrelevant content. This step aims to clean and organize the data to ensure it is accurate and readable. Special characters, punctuation, and unwanted symbols are removed, while spacing issues are corrected to make the text more coherent. Additionally, the text is standardized, such as converting abbreviations (e.g., "mg" to "milligrams") and ensuring consistent date and dosage formats. The result is cleaner, more reliable text that can be further processed in the next steps.

Step 4: Key Information Identification

This is a crucial step where the cleaned text is analyzed to identify and extract meaningful information such as:

Patient Name
Doctor's Name
Date of Prescription
List of Medicines
Dosage and Duration
Frequency of Intake

Text processing techniques and rule-based matching are used to locate and label these fields correctly.

Step 5: Data Structuring

After extracting the relevant information, the system organizes the data into structured tables. Each row represents a medication, with columns for details such as medicine name, dosage, frequency, and duration. This structured format ensures clarity and makes the data easy to access and share. The system saves this data in widely used formats like CSV or Excel, which are convenient for further analysis or sharing with healthcare professionals.

Step 6: Generic Medicine Suggestion

To make medications more affordable, the system compares the extracted branded medicine names with an internal database to identify and suggest generic alternatives. Generic medicines often cost less while offering the same therapeutic benefits, thus helping reduce the overall cost of prescriptions for patients.

Step 7: Output Generation

Once the data is structured and suggestions are made, the system generates a CSV or Excel file containing all the extracted and organized information, including any suggested generic alternatives. This file is downloadable for easy access and sharing. Additionally, users can view the results directly on the interface for immediate review and action.

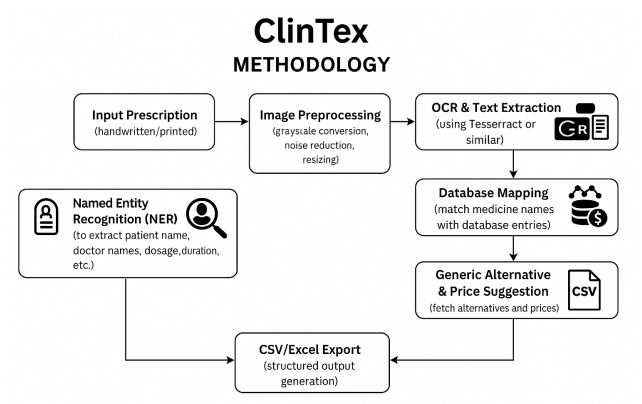

Figure 2. Overview of the ClinTex system architecture, highlighting the key components involved in prescription digitization, including OCR text extraction, Named Entity Recognition (NER) for data categorization, and the drug database mapping module for suggesting generic alternatives.

4 Results

The ClinTex Medical Prescription Project successfully extracted data from prescriptions and suggested cost-effective generic alternatives. Below are the medicine suggestions made by the system:

- **Suggestion 1:**
  - **Medicine Name:** Ascoril D Plus Syrup (Sugar Free)
  - **Composition:** Phenylephrine 5 mg
  - **Pack Size:** Bottle of 100 ml Syrup
  - **Manufacturer:** Glenmark Pharmaceuticals Ltd.
  - **Type:** Allopathy
  - **Price:** 129.0
- **Suggestion 2:**
  - **Medicine Name:** Alprax 0.25 Tablet
  - **Composition:** Alprazolam 0.25 mg
  - **Pack Size:** Strip of 15 Tablets
  - **Manufacturer:** Torrent Pharmaceuticals Ltd.
  - **Type:** Allopathy
  - **Price:** 29.0
- **Suggestion 3:**
  - **Medicine Name:** Alprax 0.5mg Tablet
  - **Composition:** Alprazolam 0.5 mg
  - **Pack Size:** Strip of 15 Tablets
  - **Manufacturer:** Torrent Pharmaceuticals Ltd.
  - **Type:** Allopathy
  - **Price:** 66.9
- **Suggestion 4:**
  - **Medicine Name:** Amlokind 5 Tablet
  - **Composition:** Amlodipine 5 mg
  - **Pack Size:** Strip of 15 Tablets
  - **Manufacturer:** Mankind Pharma Ltd.
  - **Type:** Allopathy
  - **Price:** 22.15
- **Suggestion 5:**
  - **Medicine Name:** Akt 4 Kit
  - **Composition:** Isoniazid 300 mg
  - **Pack Size:** Packet of 1 Kit
  - **Manufacturer:** Lupin Ltd.
  - **Type:** Allopathy
  - **Price:** 24.0

The table below provides a structured view of these suggestions:

Medicine Name	Composition	Pack Size	Manufacturer	Type	Price ()
Ascoril D Plus	Phenylephrine 5 mg	100 ml Syrup	Glenmark Pharma	Allopathy	129.0
Alprax 0.25	Alprazolam 0.25 mg	15 Tablets	Torrent Pharma	Allopathy	29.0
Alprax 0.5	Alprazolam 0.5 mg	15 Tablets	Torrent Pharma	Allopathy	66.9
Amlokind 5	Amlodipine 5 mg	15 Tablets	Mankind Pharma	Allopathy	22.15
Akt 4 Kit	Isoniazid 300 mg	1 Kit	Lupin Ltd.	Allopathy	24.0

Table 1
Medicine Suggestions with Generic Alternatives and Pricing

The system provided accurate data extraction and recommendations based on available prescriptions, enabling better decision-making for patients and healthcare providers.

5 Discussion

The ClinTex AI project demonstrates a practical approach to digitizing and structuring handwritten and printed medical prescriptions using OCR and intelligent post-processing. The integration of Tesseract OCR with tailored preprocessing steps significantly improved the accuracy of text extraction, especially in noisy or low-quality images often found in handwritten prescriptions. Furthermore, the implementation of Named Entity Recognition (NER) played a crucial role in identifying and categorizing key medical information, such as patient and doctor names, dosage details, and medicine frequencies.

One of the standout features of this project is the inclusion of a drug database mapping module, which enables the system to compare extracted medicine names with a curated dataset. This facilitated the accurate retrieval of drug information and made it possible to suggest cost-effective generic alternatives. Such functionality is essential in real-world medical and pharmaceutical applications where affordability and availability of medicine can be a critical concern.

While the system performed well in controlled testing, certain challenges were identified. Variability in handwriting styles, overlapping text, and prescription formats occasionally led to misclassification or missed entities. These limitations highlight the importance of continuous training and expansion of both the OCR engine and NER model to accommodate a wider range of input conditions. Incorporating feedback loops or human-in-the-loop verification could further enhance system reliability.

Overall, the ClinTex AI system offers a strong foundation for scalable digital prescription analysis. With further refinements, it has the potential to support healthcare professionals, pharmacists, and patients by reducing manual errors, improving prescription legibility, and promoting informed medicine usage through intelligent suggestions.

6 Conclusion

The ClinTex AI project aimed to develop a robust solution for extracting and categorizing medical information from prescriptions. Through the process of extracting relevant details such as patient name, doctor information, medicine names, dosage, and timings, this project has effectively demonstrated the potential of automating prescription management.

By organizing the extracted data into structured formats like CSV or Excel files, ClinTex AI offers a practical solution for medical professionals, pharmacies, and healthcare administrators to easily track prescriptions and ensure accuracy in medication distribution. Furthermore, the project's ability to suggest medicine alternatives and dosages based on past data adds a layer of efficiency, reducing human error and improving patient care.

This project also serves as a step forward in leveraging technology to enhance healthcare workflows, providing a scalable solution that can be integrated into existing systems to streamline the prescription process. ClinTex AI represents a significant advancement in healthcare automation, with the potential to improve the quality of medical services, reduce operational costs, and make prescriptions more accessible to both patients and healthcare providers.

Future work could involve expanding the dataset, improving the accuracy of the extraction model, and integrating real-time prescription validation, which will further enhance the system's usability in clinical environments.

References

1. Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 2, pp. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991
2. Baek, J., Lee, B., Han, D., Yun, S., Lee, H. (2019). Character Region Awareness for Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9365–9374. https://doi.org/10.1109/CVPR.2019.00959
3. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 369–376. https://doi.org/10.1145/1143844.1143891
4. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint* arXiv:1711.05225. https://arxiv.org/abs/1711.05225
5. EasyOCR GitHub Repository. https://github.com/JaidedAI/EasyOCR
6. CLINTEX_GEN_AI GitHub Repository (Ankush Kumar Poddar). https://github.com/1213kush/CLINTEX_GEN_AI.git
7. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
8. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
9. Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.
10. Jurafsky, D., Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Prentice Hall. [Online Draft] https://web.stanford.edu/~jurafsky/slp3/

Acknowledgements

I would like to sincerely thank everyone who contributed to the successful completion of the ClinTex Medical Prescription Project.

I am deeply thankful to Professor Dr. Neeraj Gupta, Ph.D., for his invaluable guidance, encouragement, and constructive feedback throughout the course of this project. Undertaken as part of the curriculum at the Department of Information Technology, Panipat Institute of Engineering and Technology, this work greatly benefited from his mentorship, which helped shape both the direction and depth of the research.

I also extend my appreciation to my peers, friends, and family, whose constant support and thoughtful suggestions helped me overcome various challenges during the development process. Additionally, I am grateful to all developers and open-source communities whose resources and tools contributed significantly to the implementation of this project.