Medical Prescription Data Extraction Using OCR and LLM for Import into the Takere Platform

Cauã Miranda[1], Pedro Jacinto[1], William Niemiec[1], Érika Cota[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

{caua.miranda, pedro.jacinto, william.niemiec, erika}@inf.ufrgs.br

Abstract. *This paper presents an automated solution for extracting data from medical prescription images and importing it into the Takere care plan platform. The objective is to make the data entry process more agile and reliable, replacing the manual procedure. The methodology employs Image Processing and Optical Character Recognition for text extraction, combined with Large Language Models to correct and semantically structure the information. Preliminary results indicate promising preliminary performance, achieving an F1-Score of 96.8% in the OCR extraction stage and 99.5% in component structuring. We conclude that this approach is potentially viable for mitigating errors and enhancing usability in healthcare management systems.*

1. Introduction

The advancement of digital health technologies, such as the Takere system[1] [Niemiec 2022], plays a fundamental role by centralizing the control of medication routines, assisting in patient adherence to treatment. However, the data entry process is often performed manually, limiting the acceptance of these solutions.

To mitigate this problem, this work proposes and evaluates the development of an automated medical prescription ingestion solution integrated into the Takere platform. The proposed solution is capable of processing medication entry from photographs of medical prescriptions. To this end, it employs Image Processing techniques through Optical Character Recognition (OCR) and Prompt Engineering to convert raw information into a structured format compatible with the system. This process seeks to contribute to enhancing the user experience and increasing the reliability of the records.

2. Related Works

Automated prescription information extraction has been investigated using both image-processing/OCR techniques and language models. [Jeevidha et al. 2026] and [Ponnuru et al. 2024], for example, investigate image preprocessing and OCR-based prescription extraction. Other studies have explored LLMs for extracting and structuring clinical information [Fornasiere et al. 2024, Richter-Pechanski et al. 2025].

The present work combines these approaches in an end-to-end pipeline specifically designed for integration with the Takere mHealth platform. Its contribution is therefore primarily architectural and applied: the extracted information is transformed into the native representation required by a care-plan management system.

[1]https://www.takere.com.br/

3. Methodology

The operational pipeline of the system's first stage was designed to enable extraction through images (scanned or photographs). The process consists of 4 sequential parts: visual pre-processing, text extraction via OCR, text trimming, and semantic structuring using LLM, as illustrated by Figure 1.

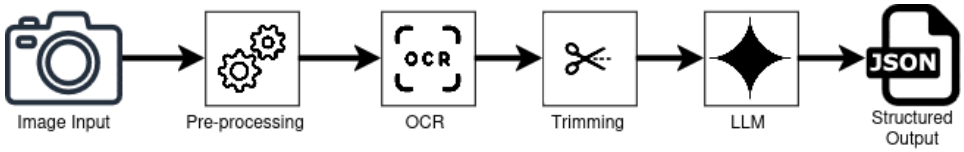


Figure 1. OCR operational pipeline

3.1. Visual Pre-processing and Extraction via OCR

In order to ensure the legibility of the characters present in the image, we applied a visual pre-processing pipeline. This process involves converting the image to grayscale, image resizing (upscaling), and binarization via thresholding. The text transcription is performed using the open-source EasyOCR[2] library, which returns the identified characters, their spatial coordinates (bounding boxes), and a confidence metric.

3.2. Trimming and Semantic Structuring

Before structuring the output in JSON, we crop the prescription to regions of interest. This is done using Regex and the coordinates of the bounding boxes. The text is then sent to the *Gemini 3.5 Flash* model[3], where text correction and the assembly of the structured output into 10 distinct fields for each medication will be performed.

3.3. Structuring the Prescription into Takere Components

In this stage, the system receives the medical prescription in JSON format, containing free-text data or generic structuring (Figure 2). The input file is the result of the image processing pipeline (Figure 1). This file is parsed and converted into a dictionary structure. This approach preliminarily organizes the data, facilitating the separation of the different medications contained in the prescription. Then, the dictionary is converted into text and incorporated into a prompt submitted to an LLM (Gemini 3.5 Flash-Lite model[4]).

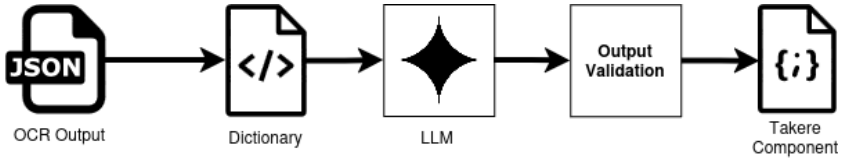


Figure 2. Processing flow for Takere component

The output structuring is controlled using the Pydantic library[5], which defines a JSON schema responsible for detailing the types and descriptions of each field in the

[2]https://www.jaided.ai/easyocr/
[3]https://openrouter.ai/google/gemini-3.5-flash
[4]https://openrouter.ai/google/gemini-3.5-flash-lite
[5]https://pydantic.dev/

output component. This schema consists of 26 structural fields, which represent the following aspects of the prescription: medication name, active ingredients, dosage, quantity, route of administration, relation to meals, treatment duration, frequency, and schedules. Finally, the response generated by the LLM undergoes automated validation against the predefined schema, ensuring that the resulting file conforms to the structural requirements expected by the Takere system.

4. Partial Results and Discussion

For the development and preliminary evaluation of the methodology, we used a dataset[6] containing 25 real patient prescription images processed in compliance with the Brazilian General Data Protection Law (LGPD) regulations. The prescriptions present structural variability and different numbers of medications. Ground-truth representations were manually prepared from the source prescriptions using the same fields evaluated by the proposed pipeline, and they were validated by expert healthcare professionals. Due to the limited number of prescriptions, the results should be interpreted as preliminary and do not establish generalization to broader prescription populations.

The quantitative evaluation was conducted using a comparison script that iterates, field by field, through the JSONs generated by the system and aligns them with the reference ground truth. For each field, the generated JSONs were compared field by field with the reference ground truth. A true positive (TP) corresponds to a correctly generated non-empty field, whereas a true negative (TN) corresponds to a correctly empty field. A false positive (FP) occurs when an empty field is populated or when the generated value is incorrect. A false negative (FN) occurs when an expected field is omitted or its generated value differs from the reference. Thus, an incorrect non-empty value is counted simultaneously as FP and FN, penalizing both precision and recall. Metrics were calculated using micro-averaging over all evaluated fields. Table 1 presents the obtained results.

Table 1. Global evaluation results using F1 Score

	Precision	Recall	F1 Score	TP	FP	FN	TN
OCR	98.4%	95.3%	96.8%	756	12	37	278
Structuring	99.5%	99.4%	99.5%	15285	70	85	5039

The overall F1 Score of 96.8% demonstrates robust performance by the system in the OCR stage. Of the 1,083 evaluated fields, 756 were extracted correctly (TP) and 278 fields were correctly identified as absent (TN), amounting to an overall accuracy rate of 95.5%. The asymmetry between Precision (98.4%) and Recall (95.3%) reveals a conservative behavior: the system prioritizes extraction correctness over completeness. When it extracts a field, it is correct in 98.4% of cases, although it occasionally omits fields present on the prescription. The low number of false positives (12) compared to the 768 generated fields confirms that hallucinations and incorrect substitutions are rare events, occurring in only 1.6% of extractions. The 37 false negatives are primarily concentrated in fields with missing data.

[6]https://doi.org/10.5281/zenodo.22314162

For the Takere structuring stage, the 25 prescriptions were submitted to the workflow 10 times each, resulting in 250 generated files. These outputs do not represent 250 independent prescriptions; rather, the experiment evaluates the consistency of the structuring process across repeated executions of the same 25 inputs. The generated outputs were compared with the corresponding reference ground truths using the same field-level evaluation procedure.

The repeated executions provided an additional assessment of the stability of the generated outputs across repeated runs of the same inputs. The minimum F1-Score recorded in the experiment was 97.3%, indicating that the occurrence of hallucinations or omissions was limited. Most inconsistencies were concentrated in formatting discrepancies within medication nomenclature, which did not impede their correct identification, preserving the integrity of critical fields such as active ingredient, dosage, and route of administration.

5. Conclusion and Future Works

This paper presented an automated pipeline for extracting and structuring medical prescription information for integration into the Takere care plan platform. The approach combines image preprocessing, OCR, LLM-based correction and semantic structuring, and schema-based validation.

Future work will test the processing workflow on broader datasets, benchmark the performance against other state-of-the-art OCR engines, and investigate the implementation of different LLMs to improve system accuracy, particularly regarding the inference of complex fields and implicit data within prescriptions.

Acknowledgments

This study was financed in part by the Rede Nacional de Ensino e Pesquisa (RNP) (PPI SAUDE DIGITAL Program – REDI-SUS Cluster) and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (BIC UFRGS Program).

References

Fornasiere, R., Brunello, N., Scotti, V., and Carman, M. (2024). Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 456–466.

Jeevidha, S., Yogeshwaran, P., Kannathasan, K., and Anand, A. P. (2026). An advanced electronic health record for prescription analysis using yolo and ocr with spell correction. In *2026 9th International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–6. IEEE.

Niemiec, W. (2022). Takere: a no-code platform for the development of mhealth applications based on care plans. Trabalho de Conclusão de Curso (Graduação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS).

Ponnuru, M., Likhitha, A., et al. (2024). Image-based extraction of prescription information using ocr-tesseract. *Procedia Computer Science*, 235:1077–1086.

Richter-Pechanski, P., Seiferling, M., Kiriakou, C., Schwab, D. M., Geis, N. A., Dieterich, C., and Frank, A. (2025). Medication information extraction using local large language models. *Journal of biomedical informatics*, page 104898.