Journal of Molecular and Pharmaceutical Sciences

Original Research Article

An OCR system optimized for Algerian drug package labels

Boufenara Mohamed Nadjib [a*], Bahi Meriem [a], Naidja Lamri [a]

[a] *Pharmaceutical Sciences Research Center (CRSP), Constantine 25000, Algeria.*

Abstract

The manual transcription of drug labels in Algerian pharmacies, especially within the CNAS, remains slow and vulnerable to error. The diversity of label designs and the uneven print quality make the task even more challenging. To improve this process, we introduce an OCR system tailored to Algerian drug labels. The method builds on PaddleOCR, enhanced with custom parameter tuning and a dedicated post-OCR correction dictionary. A dataset of 300 drug labels was collected and manually annotated. Two hundred images were used to construct the domain-specific post-OCR dictionary and tune system parameters, while 100 images were held out as an independent test set for evaluation. The system reaches a Character Error Rate of 8.5%, yielding a relative reduction of approximately 70% compared to Tesseract (28.6% CER) and 62% compared to EasyOCR (22.3% CER). It also extracts core pharmaceutical information—drug name, dosage, lot number, manufacturer, and reference price-into structured JSON outputs that integrate easily with pharmacy software. This leads to faster price estimation and smoother inventory updates, while reducing the workload of CNAS personnel. Overall, the system constitutes the first OCR solution designed for Algerian drug labels and offers a practical, scalable step toward modernizing pharmaceutical verification workflows.

Keywords:
OCR, Drug Package Labels, PaddleOCR, Automated Verification, Pharmaceutical Data

*Received November 20, 2025; accepted December 20, 2025

*Corresponding author
Email address: medboufenara@gmail.com (Mohamed Nadjib Boufenara)

Cited as: Boufenara M.N, Bahi M, Naidja L. An OCR system optimized for Algerian drug package labels. J. Mol. Pharm. Sci. 04 (02), 2025, 61-70.

1. Introduction

Pharmaceutical packaging labels play a critical role in healthcare, conveying essential information such as drug name, dosage, lot number, manufacturer, and reference price [1]. In Algeria, as in many countries, these labels are still verified manually, a process that is time-consuming, labor-intensive, and prone to error [2]. Each day, pharmacies submit hundreds of labels to the Caisse Nationale des Assurances Sociales (CNAS), where agents must ensure both the accuracy of the information and the correctness of total prices. The high volume and critical nature of this task underscore the need for an automated solution. Optical Character Recognition (OCR) offers a practical approach to streamline label verification [3]. Industrial packaging lines have successfully integrated OCR systems to reduce error rates and improve throughput [4]. Recent deep learning approaches have further enhanced recognition accuracy [5], demonstrating the feasibility of automated extraction even in non-standard layouts [6]. However, no system has been specifically optimized for Algerian drug labels.

These labels pose unique issues, including small font sizes, mixed vertical and horizontal text orientation, printing defects such as ink surplus or smudges, and variability across manufacturers and years[7]. While some general product-label OCR systems use techniques such as SIFT and line detection [8], they do not sufficiently address the challenges inherent to Algerian pharmaceutical packaging. To fill this gap, this study presents an OCR pipeline tailored to Algerian drug package labels. The system employs PaddleOCR as the core recognition engine [9], enhanced with a specialized post-OCR correction step using a custom correction dictionary to correct common misrecognitions. A dataset of 300 labels was collected from multiple pharmacies, years, and manufacturers using various smartphone models and manually annotated for validation. Recognized text is mapped to structured JSON fields—including dosage, laboratory, lot number, and reference price—via simple rule-based logic, enabling seamless integration with pharmacy software and CNAS workflows. By automating structured data extraction, the system reduces manual workload, increases accuracy, and supports scalable processing in both pharmacies and CNAS.

The remainder of this paper is organized as follows. Section 2 reviews related works, highlighting previous approaches to drug label OCR, their advantages, and limitations. Section 3 details the methodology, including dataset collection, preprocessing, OCR pipeline configuration, and post-OCR correction techniques for structured data extraction. Section 4 presents evaluation results and discusses error patterns. Finally, Section 5 concludes and suggests directions for future research.

2. Related Works

Optical Character Recognition (OCR) is a well-established technology for extracting text from images [10]. While traditional OCR engines such as Tesseract [11] perform well on regular, high-contrast text, they face significant challenges when applied to drug labels with variable layouts, mixed text orientations, inconsistent print quality, and uncontrolled imaging conditions.

The task of automating text extraction and interpretation from drug labels has been the subject of growing interest in both industrial and academic contexts [12]. Historically, many pharmaceutical packaging lines have utilised OCR/OCV (Optical Character Recognition / Optical Character Verification) systems to improve efficiency and reduce error rates. For example, one case study described how a pharmaceutical packaging line implemented OCR/OCV and reduced false reject rates from ~25% to ~0.5%, thus improving overall equipment effectiveness (OEE) by ~200% [4]. While highly effective in controlled production environments, these solutions often assume standardized layouts, uniform print quality and lighting, which limits their applicability to real-world pharmacies where label quality, orientation, and imaging conditions vary considerably.

Beyond industrial settings, recent academic research has addressed more challenging scenarios involving curved packaging, unstructured text, and uncontrolled capture conditions. For instance, one study proposed a deep learning model to extract medication information from cylindrically distorted pill bottles. By de-

warping, stitching, and segmenting curved surfaces, the system achieved 81-90% accuracy across different information categories [13]. Similarly, the DLI IT system combined scene-text detection, Tesseract OCR, and sentence-embedding similarity to identify drug labels, achieving up to 88% precision on external datasets [14].

Studies evaluating OCR on complex, multilingual packaging reveal persistent challenges. For example, research on food packaging labels in South Africa found that multilingual text, dense layouts, and curved surfaces significantly degrade performance even for state-of-the-art engines such as Tesseract, EasyOCR [15], and PaddleOCR [16]. However, incorporating line-level script identification prior to recognition has been shown to reduce character error rates by approximately 33% [17].

In low- and middle-income countries (LMICs), drug label quality often falls below Good Pharmacy Practice standards, with inconsistent typography, small fonts, and variable layouts [18]. Recent work in Thailand integrated OCR with retrieval-augmented generation (RAG) models, achieving 100% label-naming accuracy under controlled conditions, though OCR quality remained a bottleneck in real-world deployment [19].

In light of these works, some key gaps remain which are crucial for the context of Algerian drug labels:

- Many prior systems assume high image quality, stable lighting, flat surfaces, and standard fonts. However, Algerian labels often present very small fonts, mixed text orientations (vertical/horizontal), printing defects (ink surplus, smudges), and smartphone-captured images under uncontrolled conditions.
- Few, if any, systems have been explicitly developed for African or Arabic-French multilingual drug labels, where script, layout and regulatory conditions differ from those in Western contexts.
- While generic OCR engines exist (Tesseract, EasyOCR, PaddleOCR), their error-rates on domain-specific labels remain high, and very limited work focuses on post-OCR domain-specific correction dictionaries and structured JSON-output mapping tailored for downstream workflows (e.g., verifying total price, lot number, dosage) in high-volume agencies like the CNAS.

Building on these prior works, this paper introduces a hybrid methodology specifically targeted at the Algerian context. This approach combines a parameter-optimised PaddleOCR engine, a domain-specific post-OCR correction dictionary, and rule-based field-mapping logic to produce structured JSON outputs for downstream CNAS workflows. By closing the gap between generic OCR pipelines and the domain-specific requirements of Algerian drug label processing, this work aims to deliver a scalable, accurate, and efficient solution for both the pharmacy point-of-sale and CNAS high-volume batch processing environments.

3. Materials and Methods

This section describes the dataset, preprocessing, OCR pipeline, post-OCR correction, structured output generation, and the validation workflow used in the development of an Algerian drug label recognition system. The methodology was designed to address the unique challenges posed by real-world drug labels, including small fonts, vertical and horizontal text orientations, printing defects, and variable smartphone image acquisition conditions. To provide a clear overview of the complete pipeline, Figure 1 illustrates the sequence of steps from dataset collection and image preprocessing to the final validation.

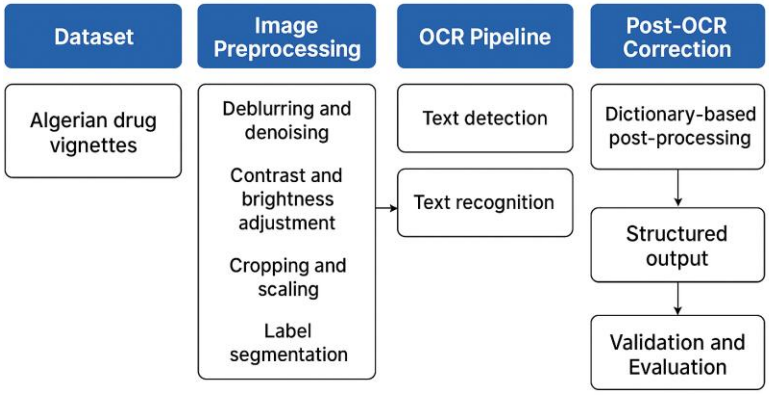

Figure 1 : Methodology Overview for the Recognition of Algerian Drug labels

3.1. Dataset Collection

A total of 300 drug labels were collected from pharmacies across Constantine and manually annotated. Of these, 200 images were used to construct the domain-specific post-OCR correction dictionary and to tune recognition hyperparameters, while the remaining 100 images were reserved as an independent test set for final evaluation. No neural-network weights were retrained; all improvements were achieved via hyperparameter optimization and domain-specific post-OCR correction. Each label was annotated with all the fields typically found on Algerian pharmaceutical reimbursement labels, including:

- Laboratory name
- Brand or commercial name of the medication
- Dosage (e.g., 500 mg, 20 mg/5 mL)
- Pharmaceutical form (e.g., tablet, capsule, syrup, injectable, cream)
- Formulation / active ingredients
- Lot number (batch number)
- Expiration date (DDE – Date of Expiration)
- Manufacturing date (DDF – Date of Fabrication)
- Barcode (linear or DataMatrix)
- Reference price (TR – Tarif de Référence)
- Reimbursement code / CNAS code
- PPA – Public Price in Algeri
- SHP / Price outside pharmacy (if present)
- Reimbursement category (e.g., 100%, 80%)
- Packaging / Conditioning (e.g., box of 30 tablets)
- Additional identifiers printed by the pharmacy (e.g., date, stamp, or internal code, if present)

This annotation process is carried out to allow a detailed evaluation of the performance of the OCR engine on heterogeneous visual and textual components of the labels.

3.2. Image Preprocessing

Images underwent several preprocessing steps to enhance OCR accuracy [20] :

- Deblurring and denoising: Gaussian and median filters were applied to reduce the effect of camera shake and printing artifacts.
- Contrast and brightness adjustment: Histogram equalization was used to normalize the intensity distribution across images.
- Cropping and scaling: Labels were cropped from larger CNAS sheets when necessary and resized to a standard resolution compatible with the OCR model.
- Segmentation of individual labels: For CNAS sheets containing multiple labels (typically 30 per sheet), rectangular bounding boxes were used to segment each label. This allowed the OCR pipeline to operate on a single label at a time, reducing errors caused by neighboring text.

With the images preprocessed and segmented, the next step is text recognition and structured data extraction using the OCR pipeline.

3.3. OCR Pipeline and Automated Structured Data Extraction

The OCR pipeline was designed to handle the specific challenges of Algerian drug labels, including small fonts, mixed text orientations, and printing defects, while providing structured outputs ready for integration into CNAS systems and pharmacy management software.

The system was implemented using PaddleOCR with the following configuration: detection model ch_PP-OCRv4 (DBNet-based architecture), French language recognition model (lang='fr'), and angle classification enabled (use_angle_cls=True) to handle mixed vertical and horizontal text orientations. The ch_PP-OCRv4 detection model, while originally designed for Chinese text, employs a character-agnostic region detection mechanism that performs effectively on Latin script.

The process begins with text recognition using PaddleOCR [21], selected for its flexibility and strong performance on complex layouts. Text regions are first detected with PaddleOCR's DBNet model, then extracted using CRNN-based recognition where detection thresholds and confidence cutoffs to improve accuracy for small fonts and unusual characters. Orientation correction is applied to adjust for rotation or skew, ensuring that vertical and horizontal text can be reliably recognized.

Following recognition, a dictionary-based post-OCR correction module addresses frequent misrecognitions caused by printing inconsistencies and domain-specific terminology. The correction dictionary was systematically constructed by processing the 200 training labels with raw PaddleOCR and cataloguing all observed misrecognitions against ground truth annotations.

Error patterns were categorized into four principal types:

- Compound term spacing errors where tight character spacing causes word concatenation (e.g., "tarifderef" → "Tarif de Ref")
- Typographic substitutions in pharmaceutical vocabulary (e.g., "bolte" → "Boite", "ineuline" → "Insuline", "cornprime" → "comprimé")
- Character-level confusions between visually similar characters due to print quality (O/0, l/1/I, S/5, B/8)
- Dosage formatting inconsistencies (e.g., "500rng" → "500 mg", "2Smg" → "25 mg").

The correction system operates in two tiers: an exact-match dictionary containing high-frequency pharmaceutical-specific corrections, and regular expression patterns for structural transformations such as spacing normalization around units and lot number formatting. The compile_patterns() function applies these corrections sequentially—exact matches first, then regex patterns—to prevent cascading errors. High-frequency corrections observed in the training-set include:

- "tarifderef" → "Tarif de Ref" (45% of labels)
- "bolte" → "Boite" (32%)
- "ineuline" → "Insuline" (28%)
- Numeric character confusions like "5OO" → "500" (19%)

Numeric values such as dosages and prices are validated and corrected using contextual rules: the presence of "mg", "mL", or "g" immediately following a number indicates dosage assignment, while "DA" or the symbol "%" implies formulation or reimbursement rate. This hybrid approach combining deep learning OCR with domain-specific corrections significantly improves field-level accuracy and consistency, reducing Character Error Rate from 15.2% (raw PaddleOCR) to 8.5% after correction—a 44% relative improvement.The dictionary was constructed exclusively from the 200 training images and frozen before evaluation on the test set to prevent data leakage.

Once text fields have been corrected, they are exported in a standardized JSON format, providing structured data suitable for automatic computation of total prices and integration into CNAS workflows or pharmacy management systems. For example, a single label may be represented as follows:

{
 "laboratory": "DAR AL DAWA ALG",
 "drug_name": "Amoxicilline",
 "dosage": "250 mg",
 "pharmaceutical_form": "Comprimé pelliculé sécable",
 "formulation": "Amoxicilline trihydratée",
 "lot_number": "12345",
 "expiration_date": "2025-08",
 "manufacturing_date": "2023-08",
 "tarif_de_reference": 1200,
 "ppa": 1450,
 "reimbursement_rate": "100%",
 "barcode": "6130009876543"
}

This structured output allows rapid and accurate verification, eliminates the need for manual data entry, and supports high-throughput processing of drug labels, making the system practical for both pharmacy point-of-sale applications and batch verification at CNAS.

3.4. Validation and Evaluation

To facilitate validation and correction by users, a lightweight graphical interface was developed. Each detected text box is highlighted directly on the label image, and the recognized text for each region is displayed alongside the image. Users can manually correct any misrecognized fields, and these corrections are fed back into the post-OCR correction dictionary, enabling iterative improvement of the OCR rules and increasing overall accuracy. Figure 2 illustrates the interface in action: each detected text region on a drug label is marked with a red rectangle, and the recognized text is shown next to each field. This visualization allows users to quickly verify and correct entries, ensuring high-quality structured data and providing immediate feedback to enhance the OCR system's performance.

The system implementation in Python utilizes PaddleOCR for core OCR functionality, OpenCV for image preprocessing and visualization, and NumPy for numerical operations. The modular architecture separates concerns: the OCRProcessor class handles detection and recognition, the compile_patterns() function manages post-OCR corrections, and field extraction logic maps corrected text to structured outputs.

The performance of the system was evaluated against the manually annotated ground truth using three complementary metrics. Character Error Rate [22] measures the proportion of misrecognized characters,

field-level accuracy quantifies the percentage of correctly recognized fields after post-OCR correction, and end-to-end workflow accuracy assesses whether all fields within a label were correctly recognized and structured in JSON format.

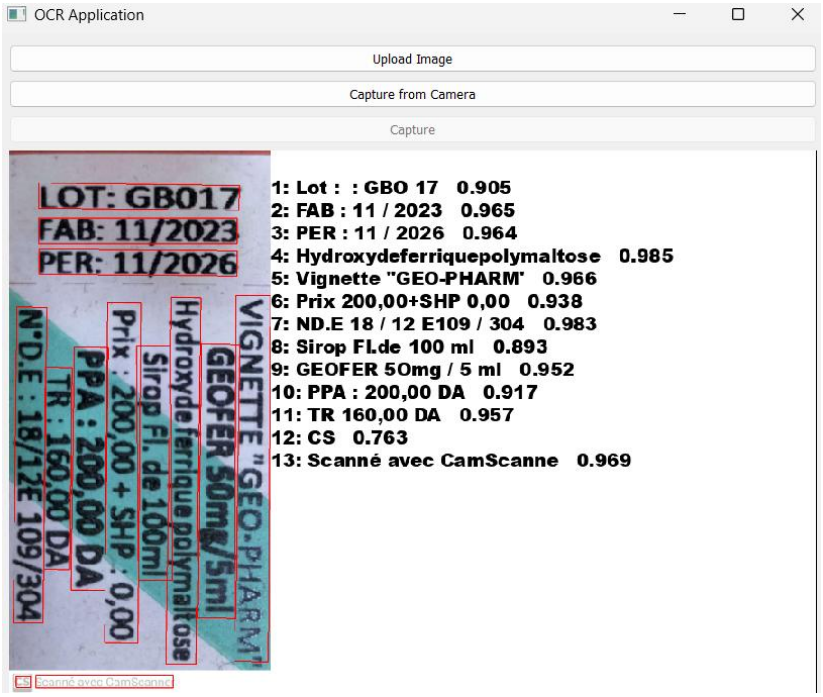

Figure 2: Screenshot of the graphical interface showing detected text regions for a drug label

This integrated interface and comprehensive evaluation framework, designed with careful consideration of **real-world design constraints**, thus provides a **robust and representative basis** for the performance analysis presented in the following section.

4. Results and Discussion

This section presents the evaluation of the proposed OCR system on Algerian drug labels. Performance metrics include CER, Field-level Accuracy, and End-to-End Accuracy. Comparative analysis against two popular OCR engines, Tesseract and EasyOCR, is also provided to highlight the improvements introduced by the optimized PaddleOCR pipeline combined with post-OCR correction and rule-based field mapping. In addition to accuracy, system speed and computational efficiency were evaluated to assess the feasibility of real-time integration in pharmacies and CNAS batch processing workflows. On a standard workstation (Intel Core i5-1135G7, 8 GB RAM), the optimized PaddleOCR pipeline achieved an average inference time of 142 ms per label (≈7 labels/s).

4.1 Evaluation Metrics and performance on test dataset

- **Character Error Rate (CER):**
  Measures the proportion of incorrectly recognized characters in each label, calculated as:

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

  Where S is the number of substitutions, D deletions, I insertions, and N the total number of characters in the ground truth.
- **Field-level Accuracy:**
  Measures the percentage of correctly recognized fields (laboratory, dosage, lot, tariff, formulation) after post-OCR correction.
- **End-to-End Accuracy:**
  Measures whether all fields in a label are correctly recognized and correctly structured in JSON format.

To rigorously demonstrate the necessity of our hybrid approach, which combines an optimized PaddleOCR engine with domain-specific post-OCR correction logic, we first contextualize its performance by comparing it against three widely used OCR engines on the same test dataset: Tesseract, EasyOCR, and a raw PaddleOCR implementation. The results, reflecting performance on the heterogeneous test dataset, are compiled in Table 1.

Table 1: Comparative performance of OCR systems on Algerian drug labels

OCR System	CER(%)	Field-level Accuracy (%)	End-to-End Accuracy (%)
Tesseract	28.6	63	42
EasyOCR	22.3	71	51
PaddleOCR (raw)	15.2	81	65
Proposed PaddleOCR + Post-OCR correction	8.5	96	90

As shown, the proposed system significantly outperforms Tesseract and EasyOCR across all metrics. CER was reduced by approximately 70% compared to Tesseract and 62% compared to EasyOCR, demonstrating the effectiveness of the dictionary-based post-correction and field-mapping rules.

4.2 Comparative Evaluation

The OCR system demonstrated strong overall performance, yet certain errors persisted, particularly in labels with extremely small fonts below six points, severe ink smudges, or partially obscured text. In evaluating the remaining failure cases, it is also essential to distinguish between realistic degradations and conditions that do not occur in operational workflows. Severely damaged labels are not admissible in pharmacies or CNAS workflows and are therefore outside the scope of the system. Only minor imperfections (small smudges, folds, contrast variations) occur in real reimbursable labels, and these were included in the dataset. The post-OCR dictionary corrected most common misrecognitions, but novel or unusual misprints occasionally required manual correction. Ambiguities in numerical strings, such as “25” which could represent either a dosage or a percentage, were resolved through context-based rules, where the presence of “mg” always indicates dosage and the symbol “%” implies formulation. Segmenting multiple labels on a single CNAS sheet was crucial for reducing errors caused by overlapping bounding boxes, interference from neighboring

text, or sheet folds. The custom post-OCR dictionary proved highly effective, while reducing substitution errors for dosage and laboratory fields by 55 to 60 percent. This process ensured consistent structured JSON output, enabling seamless integration with pharmacy management systems and CNAS verification workflows.

The comparative analysis rigorously demonstrated the inadequacy of generic, state-of-the-art OCR engines in handling the heterogeneous and challenging characteristics of Algerian drug labels. Importantly, the system remains compatible with newly introduced medications. Because recognition is performed at the character level rather than through a fixed dictionary of drug names, unfamiliar commercial names or novel formulations are processed normally. The post-OCR correction module only intervenes when a known misrecognition pattern is detected, ensuring that new drugs are handled without requiring updates to the OCR engine. The performance benchmarked against Tesseract and EasyOCR highlights the necessity of a domain-specific, hybrid approach to achieve production-level accuracy.

Tesseract exhibited the poorest performance due to difficulties in robustly segmenting and recognizing text in small font sizes (below six points) and handling mixed text orientations, resulting in a high frequency of substitution and deletion errors. EasyOCR provided marginal improvements, demonstrating greater ability to cope with mixed vertical and horizontal text orientations. However, its overall error rate remained unacceptably high due to lack of adaptation to domain-specific vocabulary and characteristic printing inconsistencies. Even the raw PaddleOCR implementation, a sophisticated deep learning framework, underscored that advanced deep learning alone is insufficient. A substantial gap exists between generic recognition and the required field-level consistency for structured data extraction. The proposed system successfully bridged this performance gap by combining an optimized PaddleOCR pipeline with two critical rule-based components: a domain-specific post-OCR dictionary and rule-based field mapping. These results validate that integrating deep learning with domain-specific knowledge is paramount for deploying accurate and scalable OCR solutions in complex, real-world regulatory environments.

5. Conclusion and Future Work

In this paper, we have optimized an OCR system for Algerian drug labels. This proposal is designed to overcome the challenges of small font size, variable text orientations, and printing defects. The hybrid approach automatically extracts the necessary information, including the manufacturer, dosage, batch number, price, and formulation, and produces appropriate structured JSON data. Evaluation was conducted on a 100-image independent test set, which was held out from the total collection of 300 manually annotated labels. The results demonstrated strong performance, with a character error rate (CER) of 8.5%, field-level accuracy of 96%, and end-to-end accuracy of 90%. Comparative analysis with Tesseract and EasyOCR confirmed that generic OCR engines are insufficient for Algerian drug labels, highlighting the importance of domain-specific corrections. The proposed system offers significant practical advantages. It automates batch verification at the CNAS (National Social Security Fund), eliminates manual calculation of total prices, reduces human error, and enables real-time recognition of labels in pharmacies without manual data entry or barcode generation. Furthermore, the framework is scalable and adaptable, allowing integration with other health information systems and expansion to multilingual labels. Despite its robustness, the system has several avenues for future improvement. Integrating Arabic OCR capabilities is necessary because many Algerian labels contain bilingual text. Advanced image enhancement and dewarping techniques could reduce errors in challenging conditions, such as curved, folded, or damaged labels. Incorporating user-corrected misrecognitions into a dynamic dictionary would enable continuous learning and improve accuracy as new laboratories and formulations appear. Automated detection and segmentation of multiple labels on a single CNAS sheet would further streamline high-volume processing, while AI-powered validation could verify contextual consistency, such as matching dosage with standard drug formulations, to enhance quality control. In conclusion, this study provides a practical, high-accuracy, and scalable OCR solution for Algerian drug labels, addressing a critical gap in the current health informatics landscape. The methodology and system

design offer a blueprint for similar applications in other low- and middle-income countries or contexts with non-standardized drug labeling, paving the way for automated, reliable, and high-throughput healthcare data management.

References

[1] Shrank, William H., et al. "State generic substitution laws can lower drug outlays under Medicaid." *Health affairs* 29.7 (2010): 1383-1390.

[2] Goldmann, D. "Medication errors and adverse drug events in pediatric inpatients." *Journal of the American Medical Association* 285.16 (2001): 2114-2120.

[3] Smith, Ray. "An overview of the Tesseract OCR engine." *Ninth international conference on document analysis* and recognition (ICDAR 2007). Vol. 2. IEEE, 2007.

[4] OCR/OCV vision systems improve reject rates and quality yield on pharmaceutical line, Vision Systems Design, Aug. 26, 2011.

[5] Shi, Baoguang, Xiang Bai, and Cong Yao. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE transactions on pattern analysis and machine intelligence 39.11 (2016): 2298-2304.

[6] Jaderberg, Max, et al. "Synthetic data and artificial neural networks for natural scene text recognition." *arXiv preprint arXiv:1406.2227* (2014).

[7] Athuraliya, N., et al. "Assessing medication packaging and labelling appropriateness in Sri Lanka." *Journal of pharmaceutical policy and practice* 9.1 (2016): 38.

[8] Lowe, David G. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision* 60.2 (2004): 91-110.

[9] Du, Yuning, et al. "Pp-ocr: A practical ultra lightweight ocr system." *arXiv preprint arXiv:2009.09941* (2020).

[10] Althobaiti, Hassan, and Chao Lu. "A survey on Arabic optical character recognition and an isolated handwritten Arabic character recognition algorithm using encoded freeman chain code." *2017 51st Annual conference on information sciences and systems (CISS)*. IEEE, 2017.

[11] Smith, Ray. "An overview of the Tesseract OCR engine." *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE, 2007.

[12] Koponen, Jarmo, Keijo Haataja, and Pekka Toivanen. "Recent advancements in machine vision methods for product code recognition: A systematic review." *F1000Research* 11 (2022): 1099.

[13] Gromova, Kseniia, and Vinayak Elangovan. "Automatic extraction of medication information from cylindrically distorted pill bottle labels." *Machine Learning and Knowledge Extraction* 4.4 (2022): 852-864.

[14] Liu, Xiangwen, et al. "DLI-IT: a deep learning approach to drug label identification through image and text embedding." *BMC Medical Informatics and Decision Making* 20.1 (2020): 68.

[15] Nazeem, Meharuniza, R. Anitha, and S. Navaneeth. "Open-source OCR libraries: A comprehensive study for low resource language." *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. 2024.

[16] Nagayi, Mayimunah, et al. "Evaluating OCR performance on food packaging labels in South Africa." *arXiv preprint arXiv:* (2025). *2510.03570*

[17] Fujii, Yasuhisa, et al. "Sequence-to-label script identification for multilingual ocr." *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE, 2017.

[18] Athuraliya, N., et al. "Assessing medication packaging and labelling appropriateness in Sri Lanka." *Journal of pharmaceutical policy and practice* 9.1 (2016): 38.

[19] Thetbanthad, Parinya, Benjaporn Sathanarugsawait, and Prasong Praneetpolgrang. "Application of Generative Artificial Intelligence Models for Accurate Prescription Label Identification and Information Retrieval for the Elderly in Northern East of Thailand." *Journal of Imaging* 11.1 (2025): 11.

[20] Li, Dong-Lin, Shih-Kai Lee, and Yin-Ting Liu. "Printed document layout analysis and optical character recognition system based on deep learning." Scientific Reports 15.1 (2025): 23761.

[21] Du, Yuning, et al. "Pp-ocr: A practical ultra lightweight ocr system." arXiv preprint arXiv:2009.09941 (2020).

[22] Guan, Shuhao, and Derek Greene. "Advancing post-OCR correction: A comparative study of synthetic data." arXiv preprint arXiv:2408.02253 (2024).