Issue Report on PaddleOCR Malfunction Post-Update and Workaround Implementation

Document Version: 1.0

Date: 09 June 2025

Author: Abir Chakraborty

1. Background

This project involves extracting structured medicine-related data from prescription and packaging images using object detection (YOLOv11) and text recognition (PaddleOCR). The OCR component is critical for identifying manufacturing date, expiry date, price, and tablet count.

The previously stable pipeline—last confirmed working on **16 May 2025**—was dependent on PaddleOCR, which at the time provided accurate and structured text extraction across diverse medicine box layouts.

2. Problem Statement

As of **early June 2025**, following the updates in the PaddleOCR repository (per the latest documentation), the OCR output has degraded significantly, especially on medicine packaging.

3. Observed Issues

3.1 OCR Output Degradation

From comparative results:

Date	OCR Engine	Result
On and Before 16 May 2025	PaddleOCR	Accurate parsing of MFG, EXP, PRICE, TABLETS
07 June 2025	PaddleOCR (latest)	Incorrect word segmentation, missed key entities (e.g., expiry dates merged with batch numbers, spacing issues, broken lines)
08 June 2025	EasyOCR	Consistent, structured output restored

3.2 Specific Issues Noted

- Text lines like MFG: JUL.2024. EXP. JUN.2026 are poorly segmented.
- Line joining errors: INCL.OF.ALL.TAXES becomes a single unrecognizable token.
- Date formats not consistently recognized anymore.
- Output in "ocr_result.boxes" now lacks the expected structure and contains empty strings or misplaced characters.

4. Root Cause Hypothesis

The issues are correlated with a structural or API-level update in the PaddleOCR core library or model weights after **16 May 2025**. The latest version prioritizes multilingual features and layout detection enhancements, which may have affected simple English-only structured document parsing.

5. Mitigation Attempted

To ensure continuity and reliable results for the final project evaluation, the following mitigation was adopted:

Switched to EasyOCR

- **Reason**: Simple API, robust English support, no observed regressions.
- **Result**: Correct parsing of medicine label formats. Structured extraction restored.

6. Recommendation

Until a stable version of PaddleOCR restores the earlier parsing fidelity for medicine labels:

- **Preferred OCR Engine**: EasyOCR (for English text)
- **Future Option**: Consider reverting to specific PaddleOCR commit hashes (pre-17 May 2025) if dependency lock is necessary
- **Alternative**: Test Tesseract with custom preprocessing (binarization, line segmentation)

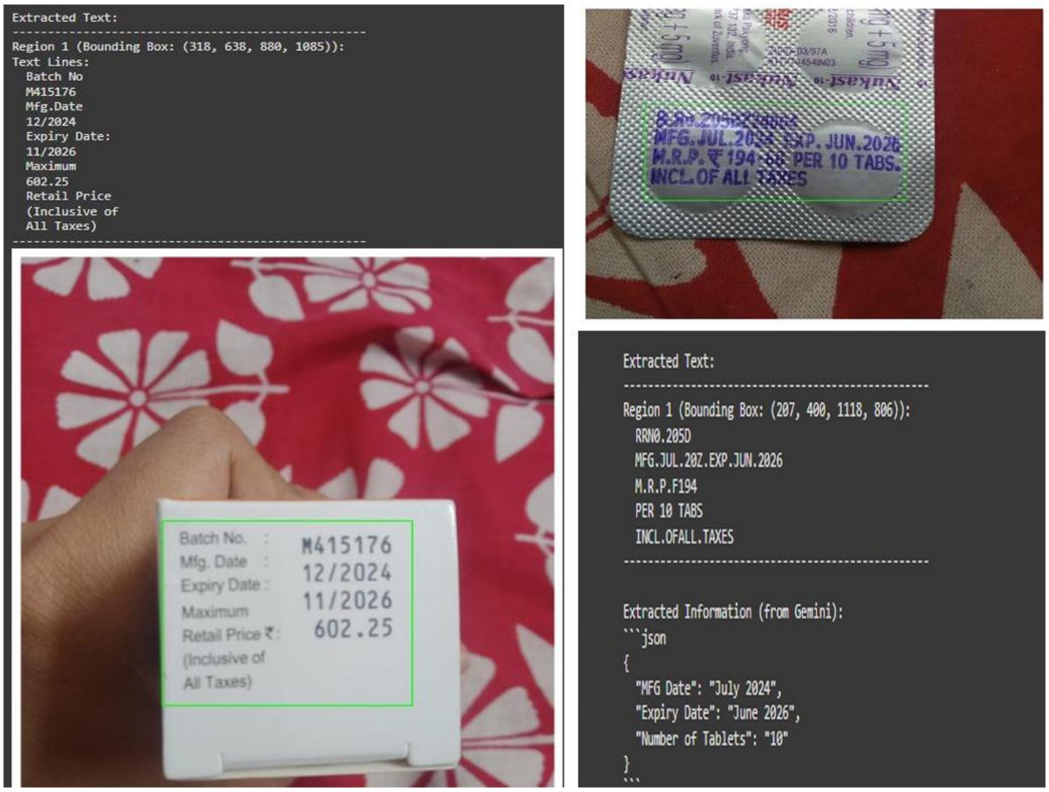
7. Conclusion

The regression in PaddleOCR performance, post-latest update, disrupts consistent text extraction. EasyOCR presents itself as a temporary yet effective solution to sustain project milestones. Further stability testing and version control of dependencies are strongly advised.

Appendix

(Sample Output Screenshots)

- **Figure A**:



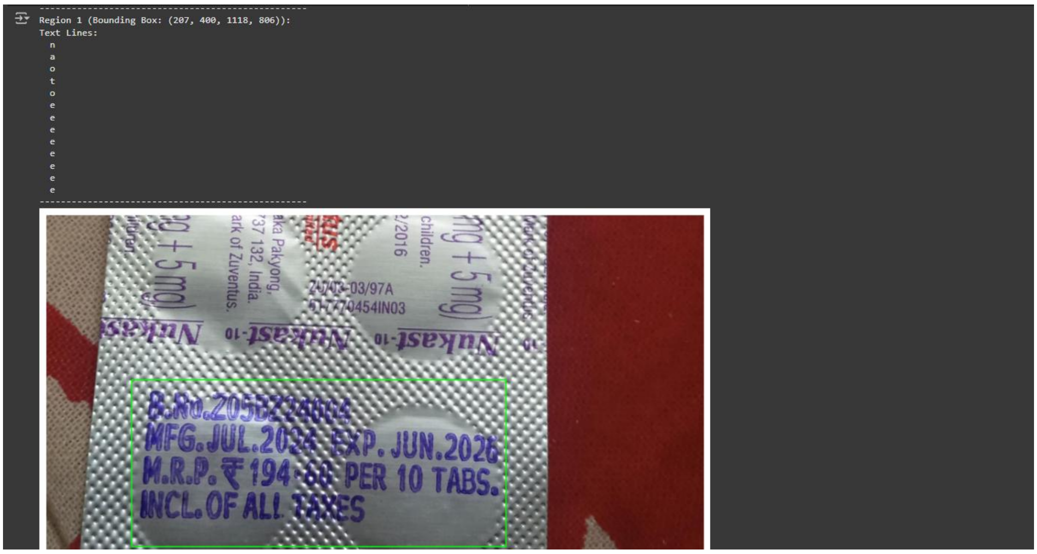
```
Extracted Text:
--------------------------------------------------
Region 1 (Bounding Box: (318, 638, 880, 1085)):
Text Lines:
  Batch No
  M415176
  Mfg.Date
  12/2024
  Expiry Date:
  11/2026
  Maximum
  602.25
  Retail Price
  (Inclusive of
  All Taxes)
--------------------------------------------------
```

~~~
Extracted Text:
--------------------------------------------------
Region 1 (Bounding Box: (207, 400, 1118, 806)):
  RRN0.205D
  MFG.JUL.20Z.EXP.JUN.2026
  M.R.P.F194
  PER 10 TABS
  INCL.OFALL.TAXES
--------------------------------------------------

Extracted Information (from Gemini):
```json
{
  "MFG Date": "July 2024",
  "Expiry Date": "June 2026",
  "Number of Tablets": "10"
}
```
~~~

[paddleocr_before.jpg] – Accurate OCR (16 May 2025)

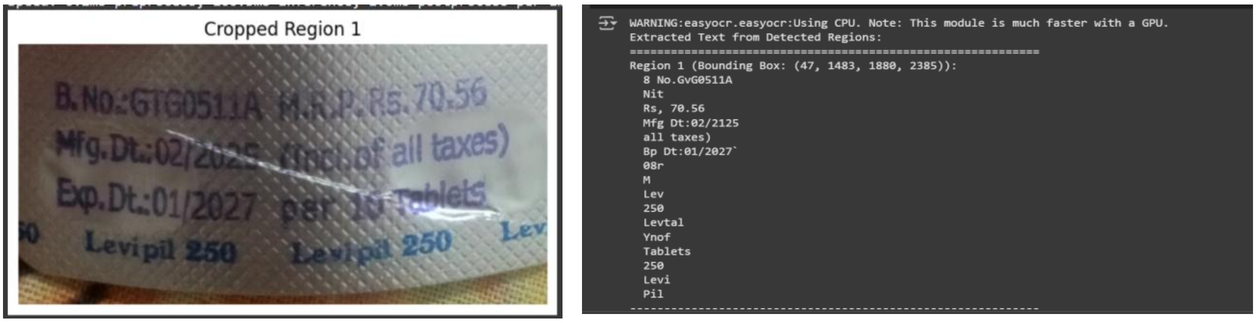
- **Figure B**:



```
--------------------------------------------------
Region 1 (Bounding Box: (207, 400, 1118, 806)):
Text Lines:
  n
  a
  o
  t
  o
  e
  e
  e
  e
  e
  e
  e
  e
--------------------------------------------------
```

[New Screenshot] – Broken formatting (09 June 2025)

- **Figure C**:


Cropped Region 1
B.No.:GTG0511A M.R.P.Rs.70.56
Mfg.Dt.:02/2025 (Incl.of all taxes)
Exp.Dt.:01/2027 per 10 Tablets
Levipil 250 Levipil 250 Lev

```
WARNING:easyocr.easyocr:Using CPU. Note: This module is much faster with a GPU.
Extracted Text from Detected Regions:
============================================================
Region 1 (Bounding Box: (47, 1483, 1880, 2385)):
  8 No.GvG0511A
  Nit
  Rs, 70.56
  Mfg Dt:02/2125
  all taxes)
  Bp Dt:01/2027`
  08r
  M
  Lev
  250
  Levtal
  Ynof
  Tablets
  250
  Levi
  Pil
------------------------------------------------------------
```

[EasyOCR Output] – Restored structure with EasyOCR