

Comparison of Tesseract, PaddleOCR and EasyOCR on a Pharmaceutical Certificate of Quality

This comparison evaluated three OCR engines, Tesseract, PaddleOCR, and EasyOCR, using the same pharmaceutical Certificate of Quality document. Each engine was assessed based on text extraction accuracy, layout handling, ease of use, output quality, and suitability for pharmaceutical document processing. The aim was not to identify a universally superior OCR engine, but to understand the strengths and limitations of each tool when extracting structured information from compliance documents.

Based on your comparison, which OCR tool worked best for your pharmaceutical PDF? Why?

All three OCR engines successfully extracted the key information from the document, including the product name, lot number, product article number, manufacture date, and expiration date. However, PaddleOCR provided the best overall balance between text accuracy and document structure. Its layout-aware approach preserved the organisation of the page more effectively than the other engines, and it produced structured polygon bounding boxes that would be useful for downstream processing. It made a few recognition errors in the middle of the document, but the general output remained well organised and easy to interpret.

Tesseract also performed very well and produced highly accurate text for most of the document. Its main limitation was that it treated the page primarily as continuous text and required additional post-processing to recover document structure.

EasyOCR successfully extracted the important compliance fields and was straightforward to use, but it produced more fragmented output by splitting related information across multiple text segments and introducing several minor OCR errors.

What surprised you?

The biggest surprise was how well Tesseract performed on this document despite being the oldest and simplest OCR engine. It accurately recognised nearly all of the important compliance fields with only minor OCR artefacts. I also found that PaddleOCR was not perfect despite its advanced layout capabilities, as it occasionally misread individual words within the document. EasyOCR performed well on the document but struggled more with certain characters and the grouping of related fields than expected.

Which engine felt best suited to pharmaceutical compliance workflows?

PaddleOCR appeared to be the most suitable for pharmaceutical compliance workflows. Pharmaceutical documents often contain structured sections, tables, labels, and multiple information blocks that benefit from layout-aware OCR. PaddleOCR preserved these relationships more effectively than the other engines while also providing polygon-based bounding boxes that could be integrated into AI document processing pipelines.

Were there any frustrations with install/setup or output structure?

Tesseract was the simplest tool to install and use. PaddleOCR required the most setup and initially produced compatibility issues that required updating the configuration before it would run correctly. It took some time to resolve the compatibility issues, and I also used AI tools to help troubleshoot the configuration. Once configured, however, it produced high-quality structured output. EasyOCR was straightforward to install but required more interpretation because some information was divided into multiple text segments and several alphanumeric fields contained minor recognition errors.

Would you combine tools or pick one for production use?

For a production pharmaceutical document workflow, I would primarily use PaddleOCR because of its layout awareness and structured output. I would consider combining PaddleOCR with Tesseract for validation of critical compliance fields such as lot numbers, product codes, and expiration dates. Using two OCR engines together could improve reliability by allowing important values to be cross-checked before they are passed to downstream AI models or document processing systems.

Conclusion

This comparison demonstrated that no single OCR engine is ideal for every scenario. Tesseract delivered strong text accuracy with minimal setup, PaddleOCR provided the best structured output and layout preservation, and EasyOCR offered a good balance between simplicity and performance. For pharmaceutical compliance documents that require accurate extraction of structured information, PaddleOCR was the strongest overall choice, particularly when combined with appropriate preprocessing and validation techniques.