

An Example Scanned Page

The rapid advancement of document understanding systems has enabled automated extraction of text, tables and mathematical expressions from scientific literature. In this paper we present a hybrid pipeline that combines fast native text extraction with neural layout analysis and optical character recognition for scanned material. We evaluate the approach on a diverse corpus of born-digital and rasterised documents and report favourable results across all routing modes.

Equation recognition is performed by a lightweight ONNX model that transcribes cropped formula regions into LaTeX source.

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

Table regions are converted to GitHub-flavoured markdown using a structure-recognition pass, while figures are staged as assets for downstream embedding and retrieval pipelines.