

The aim of this study is to develop an automated and scalable PDF text extraction system using Python, incorporating Optical Character Recognition (OCR) and advanced parsing techniques to accurately process diverse document structures while ensuring efficiency, reliability, and usability.

This study focuses on the design and implementation of an automated system for text extraction from PDF documents using Python-based tools and machine learning–driven approaches. The system is developed to handle both text-based and image-based PDFs by integrating Optical Character Recognition (OCR) and layout-aware parsing techniques. The scope specifically covers the extraction of text from documents containing varied structures such as multi-column layouts, tables, and charts, while ensuring the preservation of document integrity and minimizing common extraction errors.

The dataset for this study will consist of diverse PDF documents drawn from academic, legal, financial, and scanned sources, thereby reflecting real-world variability in formatting and content structures. The system will be evaluated using standard performance metrics, including precision, recall, F1-score, structural similarity, and processing time, in order to determine its accuracy, efficiency, and scalability.

The study is limited to text extraction and does not address advanced downstream tasks such as semantic analysis, document summarization, or natural language understanding. Furthermore, the system is designed for experimental and research purposes; therefore, aspects such as full-scale enterprise deployment, integration with large-scale cloud infrastructures, or multilingual OCR beyond English may be considered outside the scope of this project.