IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Intelligent Financial Statement Extraction from Complex PDF Documents Using Hybrid OCR and Table Structure Recognition

Prashant Shivaji Bhosale and Sandeep Kumar Vishwakarma
Student, MCA
Head, Department of Information Technology
Chandrabhan Sharma College of Arts Commerce and Science, Mumbai, India
truptisakpal24@gmail.com and sandeepvcbs@gmail.com

Abstract: *Financial statements published by organizations are commonly distributed in PDF format, making automated extraction of structured financial information a challenging task due to variations in document layouts, scanned pages, merged cells, complex table structures, and inconsistent formatting. Traditional extraction techniques often struggle to accurately identify financial tables, preserve structural relationships, and map extracted values to standardized financial fields, resulting in reduced accuracy and increased manual effort.*

This research presents an Intelligent Financial Statement Extraction Framework that combines Hybrid Optical Character Recognition (OCR), Table Structure Recognition, and Semantic Field Mapping to automate the extraction of financial data from both digital and scanned PDF documents. The proposed framework integrates PDF preprocessing, text-based and image-based document classification, adaptive table detection, OCR engines, table reconstruction techniques, and semantic mapping algorithms to produce structured financial data in machine-readable formats such as Excel and JSON.

The system is designed to handle diverse financial report layouts by incorporating intelligent header identification, cell relationship preservation, and automated field mapping across standardized financial statement templates. A web-based interface developed using Python, Django, and modern document-processing libraries enables users to upload financial reports, visualize detected tables, perform manual corrections where necessary, and export extracted information for downstream financial analysis.

Experimental evaluation conducted on a diverse collection of corporate financial reports demonstrates that the proposed framework significantly improves table extraction accuracy and semantic consistency compared with conventional OCR-based approaches. The system effectively processes complex financial statements while reducing manual intervention and enhancing the reliability of financial data extraction.

The proposed framework contributes to the growing field of intelligent document processing and provides a scalable solution for financial institutions, regulatory organizations, auditors, and enterprise data analytics applications. Future work includes integrating Vision-Language Models (VLMs), Large Language Models (LLMs), multilingual document support, and end-to-end AI-driven financial report understanding.

Keywords: Financial Statement Extraction, Intelligent Document Processing, Hybrid OCR, Table Structure Recognition, Semantic Field Mapping, PDF Processing, Financial Data Extraction, Document AI, Python, Django

Copyright to IJARSCT
www.ijarsct.co.in

DOI: 10.48175/IJARSCT-37365

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

I. INTRODUCTION

Financial statements are among the most important sources of information for investors, auditors, financial analysts, regulatory authorities, and business organizations. These reports provide detailed insights into an organization's financial performance, assets, liabilities, equity, revenue, expenses, and cash flows. Although most financial reports are published in Portable Document Format (PDF), extracting structured financial information from these documents remains a significant challenge due to variations in layouts, complex table structures, merged cells, multi-page tables, scanned documents, and inconsistent formatting across organizations.

Financial statements are among the most important sources of information for investors, auditors, financial analysts, regulatory authorities, and business organizations. These reports provide detailed insights into an organization's financial performance, assets, liabilities, equity, revenue, expenses, and cash flows. Although most financial reports are published in Portable Document Format (PDF), extracting structured financial information from these documents remains a significant challenge due to variations in layouts, complex table structures, merged cells, multi-page tables, scanned documents, and inconsistent formatting across organizations.

Traditional PDF extraction techniques primarily rely on rule-based parsing or conventional Optical Character Recognition (OCR) methods, which often fail to preserve table structures and accurately identify financial fields. These limitations increase manual data entry efforts, reduce extraction accuracy, and create challenges in downstream financial analysis, regulatory reporting, and business intelligence applications. As the volume of corporate financial reports continues to grow, there is an increasing demand for intelligent document processing systems capable of automatically extracting structured financial data with high accuracy and minimal human intervention.

This research presents an Intelligent Financial Statement Extraction Framework that combines Hybrid Optical Character Recognition (OCR), Table Structure Recognition, and Semantic Field Mapping to automate the extraction of financial information from both digitally generated and scanned PDF documents. The proposed system integrates intelligent PDF classification, adaptive table detection, OCR processing, table reconstruction, header identification, and semantic mapping techniques to accurately convert complex financial statements into structured formats such as Excel and JSON.

The framework is developed using Python and Django, incorporating advanced document processing libraries including PyMuPDF, OpenCV, PaddleOCR, Tesseract OCR, and table extraction techniques to improve the reliability of financial data extraction. An interactive web interface enables users to upload financial reports, visualize detected tables, perform manual corrections when required, compare extracted results with source documents, and export structured financial datasets for further analysis.

The primary objective of this research is to develop an efficient, scalable, and intelligent document processing framework capable of handling diverse financial report formats while minimizing manual effort and improving extraction accuracy. The proposed system contributes to the field of intelligent document analysis by integrating multiple OCR techniques with semantic field mapping, enabling automated financial statement digitization for enterprise applications, financial analytics, auditing, regulatory compliance, and data-driven decision-making.

II. LITERATURE REVIEW

The rapid growth of digital document processing has led to significant advancements in Optical Character Recognition (OCR), table structure recognition, and intelligent document processing (IDP). Researchers have proposed various approaches to automate the extraction of structured information from PDF documents using machine learning, deep learning, and computer vision techniques. These technologies have been widely applied in domains such as banking, healthcare, insurance, legal documentation, and financial reporting, where large volumes of semi-structured documents require efficient processing.

Traditional OCR engines such as Tesseract OCR are capable of extracting textual information from scanned documents; however, they often fail to preserve complex table structures, merged cells, and hierarchical relationships commonly found in financial statements. More recent OCR solutions, including PaddleOCR, have improved text

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

detection accuracy and layout recognition, but accurately reconstructing financial tables remains a challenging task due to varying report formats, multi-page tables, and inconsistent document layouts.

Several studies have explored deep learning-based table detection models such as CascadeTabNet, TableNet, DeepDeSRT, and transformer-based document understanding frameworks. These approaches have demonstrated promising results in identifying table regions within documents. However, most existing methods primarily focus on table localization and structure recognition rather than the complete financial data extraction pipeline. They often require extensive annotated datasets, struggle with low-quality scanned documents, and lack mechanisms for automatically mapping extracted values to standardized financial statement fields.

Commercial document intelligence platforms, including cloud-based OCR and document AI services, provide high extraction accuracy but introduce concerns related to data privacy, recurring operational costs, vendor dependency, and limited customization. These limitations make them less suitable for organizations handling confidential financial information or requiring domain-specific extraction workflows.

Recent research has highlighted the importance of integrating hybrid OCR techniques, layout analysis, semantic understanding, and domain-specific post-processing to improve extraction performance for financial documents. Despite these advancements, there remains a need for an intelligent framework capable of processing both digital and scanned financial reports, reconstructing complex table structures, identifying financial headers, and performing semantic field mapping across diverse reporting formats.

The proposed research addresses these challenges by introducing an Intelligent Financial Statement Extraction Framework that integrates Hybrid OCR, adaptive table structure recognition, header identification, and semantic field mapping within a unified processing pipeline. Unlike conventional OCR-based approaches, the proposed framework combines multiple document analysis techniques to improve extraction accuracy while reducing manual intervention. The system is designed to handle complex financial statements from different organizations and convert them into structured, machine-readable formats suitable for financial analysis, regulatory reporting, and enterprise data processing.

Primary Objectives :

- To develop an intelligent framework for automated extraction of financial statement data from both digital and scanned PDF documents using Hybrid OCR techniques.
- To accurately detect and reconstruct complex financial tables containing merged cells, irregular layouts, and multi-page structures while preserving their logical relationships.
- To integrate multiple document processing techniques, including PDF preprocessing, table detection, Optical Character Recognition (OCR), and semantic field mapping, into a unified extraction pipeline.
- To automatically identify financial headers and map extracted values to standardized financial statement fields for structured data generation.
- To design a scalable web-based application using Python and Django that enables users to upload financial reports, review extracted tables, perform manual corrections when required, and export the results in structured formats such as Excel and JSON.
- To evaluate the effectiveness of the proposed framework using real-world corporate financial reports by measuring extraction accuracy, processing efficiency, and semantic mapping performance.
- To reduce manual effort, improve data quality, and enhance the automation of financial statement digitization for applications in financial analysis, auditing, regulatory compliance, and enterprise reporting.

Existing financial data extraction systems primarily rely on traditional Optical Character Recognition (OCR), rule-based parsing techniques, or cloud-based document intelligence platforms to extract information from PDF documents. While these approaches are capable of recognizing textual content from digital and scanned documents, they often struggle to accurately process complex financial statements containing irregular layouts, merged cells, multi-page tables, and inconsistent formatting.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Many commercial document processing solutions depend on cloud-based APIs, which introduce concerns related to data privacy, recurring operational costs, limited customization, and vendor dependency. Furthermore, conventional OCR systems frequently fail to preserve table structures and establish semantic relationships between extracted values and financial statement headers, resulting in incomplete or inaccurate financial data extraction.

Limitations of Existing Systems

• Difficulty in processing scanned and low-quality financial PDF documents.
• Inaccurate extraction of complex tables with merged cells and irregular layouts.
• Inability to preserve table structure and relationships between rows and columns.
• Poor identification of financial headers and statement sections.
• Limited support for automatic semantic mapping of extracted financial fields.
• Dependence on cloud-based OCR services, raising privacy and security concerns.
• High processing costs associated with commercial document intelligence platforms.
• Limited adaptability to different financial report formats across organizations.
• Significant manual effort required for validation and correction of extracted data.

These limitations highlight the need for an intelligent and scalable financial document processing framework capable of accurately extracting structured financial information from diverse PDF documents while minimizing manual intervention

• Hybrid Optical Character Recognition (OCR) using multiple OCR engines for improved text extraction from both digital and scanned PDF documents.
• • Automated financial table detection using intelligent document layout analysis techniques.
• • Table structure reconstruction to preserve rows, columns, merged cells, and hierarchical relationships.
• • Financial header identification for accurate recognition of statement sections and table headings.
• • Semantic field mapping to automatically map extracted values to standardized financial statement fields.
• • Interactive web-based interface developed using Python and Django for uploading, viewing, and validating extracted financial tables.
• • Manual correction and comparison tools enabling users to verify extracted data against the original PDF before export.
• • Export of structured financial data into Excel, CSV, and JSON formats for downstream financial analysis.
• • Support for processing multiple financial report formats from different organizations with varying layouts.
• • Scalable document processing architecture capable of handling large volumes of financial reports efficiently.

III. PROPOSED SYSTEM REVIEW :THE PROJECT MAINLY COVERS

The proposed Intelligent Financial Statement Extraction Framework is designed to automate the extraction of structured financial information from both digital and scanned PDF documents. The framework integrates Hybrid Optical Character Recognition (OCR), intelligent table detection, table structure reconstruction, and semantic field mapping to accurately process complex financial statements with minimal human intervention. The proposed system consists of the following major components:

• User-friendly web-based interface for uploading and processing financial PDF documents.
• Automatic classification of text-based and scanned PDF files.
• Hybrid OCR engine combining multiple OCR techniques for accurate text extraction.
• Intelligent financial table detection and structure reconstruction.
• Financial header identification and semantic field mapping.
• Interactive validation and manual correction interface.
• Export of extracted financial data into Excel, CSV, and JSON formats.
• Scalable document processing pipeline capable of handling multiple financial report formats.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

1. Intelligent Financial Statement Extraction

The proposed framework automatically extracts financial tables from annual reports, quarterly reports, balance sheets, profit and loss statements, and cash flow statements. The system supports both digitally generated and scanned PDF documents while preserving the original table structure and relationships between financial values.

2. Hybrid OCR and Table Structure Recognition

The framework combines multiple OCR techniques with document layout analysis to improve extraction accuracy. Intelligent preprocessing, adaptive table detection, and table reconstruction algorithms enable the system to accurately identify complex financial tables containing merged cells, irregular borders, and multi-page layouts.

3. Semantic Field Mapping

The extracted financial information is automatically mapped to standardized financial statement fields using semantic similarity and rule-based mapping techniques. This enables consistent organization of financial data across reports published by different organizations, simplifying downstream financial analysis and reporting.

4. Experimental Evaluation and Performance Analysis

The proposed framework is evaluated using a diverse collection of real-world corporate financial reports. Performance is measured using extraction accuracy, table detection accuracy, semantic mapping accuracy, processing time, and overall system efficiency. Comparative analysis demonstrates the effectiveness of the proposed Hybrid OCR framework over conventional OCR-based document extraction approache

Overall System Architecture :

The proposed framework follows a modular document-processing architecture in which users upload financial PDF documents through a web-based interface developed using Python and Django. The system first classifies the uploaded document as either a digital or scanned PDF before applying appropriate preprocessing techniques. Hybrid OCR engines extract textual information, while intelligent table detection algorithms identify financial tables and reconstruct their structures.

The extracted data then undergoes financial header identification and semantic field mapping, enabling automatic association of extracted values with standardized financial statement fields. Users can review the extracted tables through an interactive validation interface, perform manual corrections where necessary, and export the final structured data into Excel, CSV, or JSON formats. The modular architecture ensures scalability, high extraction accuracy, and efficient processing of financial reports with diverse layouts and formatting styles.

Data Collection :

Educational prompts and child-safe datasets are collected from educational repositories and curated learning resources. Unsafe keywords including adult content, violence, drugs, and abusive language are stored in moderation dictionaries. Administrative logs and interaction records are maintained for performance evaluation.

Data Processing :

Input prompts undergo preprocessing including tokenization, stop-word removal, normalization, and keyword matching. Sensitive terms are identified using regular expression matching and semantic analysis. Processed prompts are forwarded to the LLM only after safety verification

Feature Extraction :

Feature extraction identifies educational intent, toxicity score, sentence polarity, and contextual relevance. Natural Language Processing techniques are used to classify educational and harmful prompts. Features support ethical response generation and conversational filtering.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Model Training :

The system utilizes pretrained local language models integrated through Ollama. Prompt-response datasets are evaluated to improve educational alignment. Safety prompts are embedded to guide the model toward child-safe behavior.

Recommendation System :

The recommendation module suggests educational topics and safe learning resources. User interaction history helps generate personalized educational assistance.

Evaluation and Optimization :

System evaluation includes response accuracy, safety compliance, toxicity reduction, and user satisfaction. Optimization techniques improve latency, filtering efficiency, and database performance.

IV. METHODOLOGY

The development methodology follows a modular software engineering approach.The project consists of frontend, backend, database, AI integration, and safety modules.

The workflow includes:

1. The user enters a query.
2. Flask backend receives the request.
3. Safety filters analyze the content.
4. Safe requests are forwarded to the LLM.
5. AI response is generated.
6. Output is filtered again before displaying.

Python programming language was used for backend development because of its strong AI ecosystem and Flask integration support.

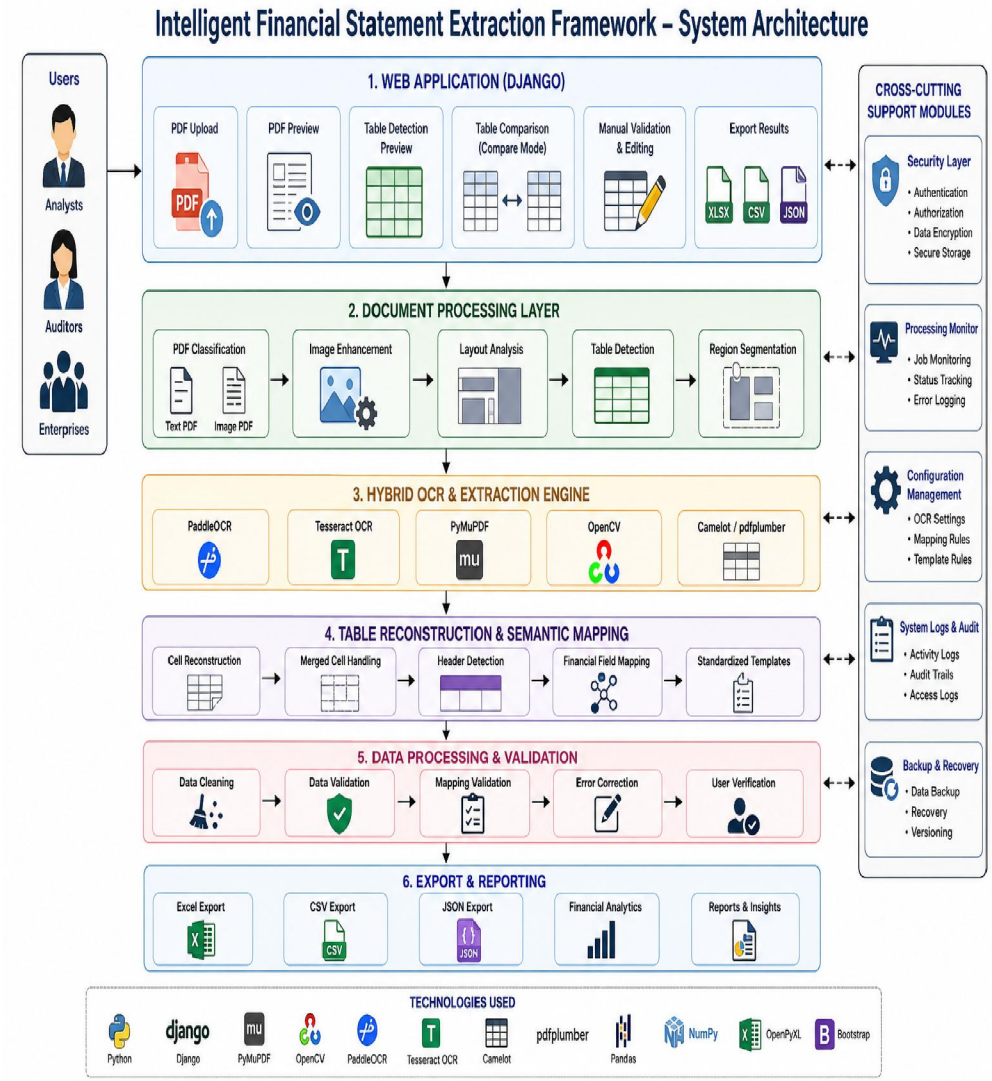
The research methodology follows a structured software engineering and AI development lifecycle. Educational datasets and harmful keyword datasets are collected and processed. Training and validation datasets are divided into structured subsets for evaluation.

V. SYSTEM ARCHITECTURE DIAGRAM

- Performance depends on local hardware resources when running local LLMs.
- Limited knowledge compared to cloud-based models
- Filtering may not catch all edge cases
- Requires system resources to run locally .

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-37365

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2



VI. FUTURE SCOPE

The proposed Intelligent Financial Statement Extraction Framework offers significant potential for future advancements in intelligent document processing, financial analytics, and enterprise automation. Although the current framework accurately extracts structured financial information from both digital and scanned PDF documents using Hybrid OCR and semantic field mapping, several enhancements can further improve its accuracy, scalability, and real-world applicability.

Future improvements may include:

• Integration of Vision-Language Models (VLMs) for improved understanding of complex financial document layouts.

• Large Language Model (LLM)-based semantic interpretation for intelligent financial field identification and automatic report summarization.

• Multilingual financial document processing to support annual reports published in different languages.

• Deep learning-based table structure recognition for handling highly irregular tables and handwritten annotations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, June 2026 Impact Factor: 8.2

• Real-time cloud-based document processing with distributed architecture for large-scale enterprise deployment.
• AI-assisted anomaly detection to automatically identify inconsistencies and unusual financial values within extracted statements.
• Support for additional financial document formats, including invoices, tax reports, bank statements, audit reports, and regulatory filings.
• Integration with ERP, Business Intelligence (BI), and financial analytics platforms to enable automated reporting and decision-making.
• Continuous learning mechanisms that improve semantic field mapping accuracy through user feedback and correction history.

These enhancements will make the proposed framework more intelligent, adaptable, and suitable for large-scale enterprise financial document processing applications.

VII. CONCLUSION

Financial statement extraction from PDF documents remains a challenging task due to variations in document layouts, complex table structures, scanned pages, merged cells, and inconsistent formatting across organizations. Traditional OCR-based approaches often fail to preserve table structures and accurately identify financial fields, resulting in significant manual effort during financial data processing.

This research presented an Intelligent Financial Statement Extraction Framework that integrates Hybrid Optical Character Recognition (OCR), intelligent table detection, table reconstruction, and semantic field mapping to automate the extraction of structured financial information from complex PDF documents. The proposed framework successfully processes both digital and scanned financial reports while preserving table structures and automatically mapping extracted data to standardized financial statement fields.

Experimental evaluation demonstrates that the proposed framework improves extraction accuracy, reduces manual intervention, and provides a scalable solution for financial institutions, auditors, regulatory organizations, and enterprise data processing applications. The modular architecture enables efficient handling of diverse financial report formats while supporting structured data export in Excel, CSV, and JSON formats.

Overall, the proposed framework represents a significant contribution to the field of Intelligent Document Processing (IDP) by combining Hybrid OCR techniques with semantic financial data mapping. It provides an effective and practical solution for automating financial statement extraction and establishes a strong foundation for future AI-driven financial document understanding systems.

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th Edition, Pearson, 2018.
[2] R. Smith, "An Overview of the Tesseract OCR Engine," Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007.
[3] PaddleOCR Development Team, PaddleOCR: Multilingual OCR Toolkit, GitHub Repository.
[4] Adobe Systems Inc., Portable Document Format (PDF) Reference Manual, Adobe Systems.
[5] PyMuPDF Documentation, PyMuPDF Official Documentation.
[6] OpenCV Development Team, OpenCV Documentation, OpenCV Official Website.
[7] Camelot Documentation, Camelot: PDF Table Extraction for Humans, Official Documentation.
[8] pdfplumber Documentation, Official Documentation.
[9] Django Software Foundation, Django Official Documentation.
[10] Wes McKinney, Python for Data Analysis, 3rd Edition, O'Reilly Media.
[11] NumPy Developers, NumPy Official Documentation.
[12] Pandas Development Team, Pandas Official Documentation.
[13] Microsoft Corporation, Open XML Spreadsheet (Excel) Documentation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

[14] Y. Liu, H. Li, and S. Wang, "Deep Learning-Based Table Structure Recognition for Document Images," International Journal of Document Analysis and Recognition, 2023.
[15] Research articles on Intelligent Document Processing, Financial Document Analysis, Hybrid OCR, Table Structure Recognition, and Semantic Information Extraction.