International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Marksheet Parser: A Proposed System for Automated Extraction, Visualization, and Analytics of Educational Results

Ayush Y.G[1], Ruturaj A.V[2], Alark D.M[3], Soumya B.D[4], Sonali N.P[5]

[1,2,3,4,5]Student, Department of Computer Engineering, V. E. S. Polytechnic

Abstract

This paper presents a proposal for the Marksheet Parser, an integrated platform designed to automate the extraction, visualization, and analytics of student results. The system accepts PDF format and templates such as SSC, CBSE, ICSE, and MSBTE, utilizing PDFplumber for digital extraction and OCR technologies such as Tesseract and PaddleOCR for scanned documents. By generating interactive dashboards via Chart.js and providing advanced analytics such as growth-over-time tracking, the system aims to improve academic transparency and reduce manual processing time.

Keywords: Marksheet parsing, OCR, PDFplumber, Educational analytics, Chart.js, Performance monitoring, Data visualization, Academic evaluation

1. Introduction

Result management is considered one of the most data-intensive processes in academic institutions. Annually, thousands of student marksheets are handled by schools and universities, which involves repetitive manual work. This challenge is exacerbated by human entry errors, delays in report generation, and non-standardized templates. Various boards like SSC, CBSE, and ICSE follow distinct structures, necessitating flexible solutions for automated processing.

The Marksheet Parser is proposed to address these issues by combining PDF parsing capabilities through PDFplumber, optical character recognition (OCR), data extraction algorithms, visualization, and analytics into a single system. Beyond digitizing marksheets, it is intended that the proposed tool will provide actionable insights through performance tracking, trend analysis, and growth visualization. It is expected that this approach will enable stakeholders such as students, parents, and administrators to make informed academic decisions.

2. Problem Statement

The following issues are observed in the existing academic result management process. Academic institutions process a large number of student marksheets annually in digital document formats such as Portable Document Format (PDF). These marksheets follow different structural templates depending on the educational board such as SSC, CBSE, or ICSE. Manual extraction and analysis of academic data from such documents is time-consuming and error-prone. Existing OCR-based solutions fail to accurately interpret structured tabular data in many cases. Digital PDF documents often contain

embedded text layers which are not efficiently utilized by traditional OCR-only systems. Most available systems focus only on digitization and do not provide automated computation or performance analytics. Therefore, there is a need for a unified system capable of extracting academic information from PDF-based marksheets and generating meaningful analytical insights.

3. Scope of the Study

The scope of the proposed Marksheet Parser system includes the following

- Processing of student marksheets available exclusively in PDF format.
- Support for multiple educational board templates such as SSC, CBSE, and ICSE.
- Handling of both digitally generated and scanned PDF documents.
- Automatic extraction of student details and subject-wise marks from uploaded PDF files.
- Computation of total marks, percentage, grade, and pass or fail status.
- Structuring of extracted data into a unified schema format.
- Export of processed data into JSON or CSV formats.
- Generation of graphical visualizations such as bar charts, pie charts, and line graphs using Chart.js.

Handwritten marksheets or highly distorted scanned PDF documents may not fall within the operational scope of the proposed system.

4. System Requirements

4.1 Functional Requirements

The system shall

- Allow users to upload student marksheets in PDF format.
- Automatically identify the document template based on the respective educational board.
- Extract personal and academic information using PDF parsing or OCR techniques.
- Compute total marks, percentage, grade, and pass or fail status.
- Structure extracted data into machine-readable formats such as JSON or CSV.
- Generate interactive visualizations representing subject-wise performance.
- Support growth-over-time tracking based on processed marksheet data.

4.2 Non-Functional Requirements

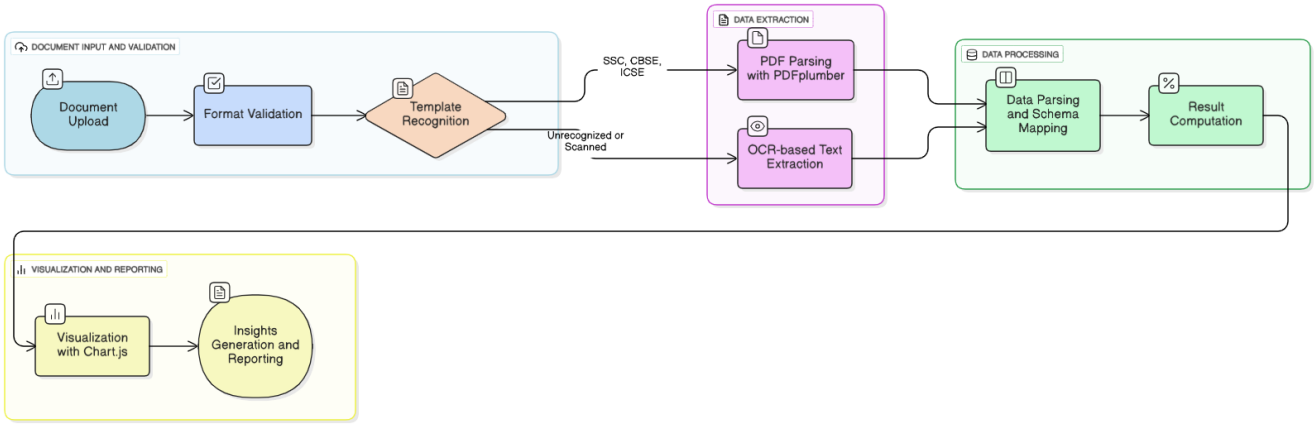
The system shall:

- Ensure accurate data extraction by selecting between direct PDF text parsing and OCR-based recognition depending on the PDF type.
- Maintain consistent processing performance across different marksheet templates.
- Provide confidence scoring for extracted academic data.
- Support scalability for processing multiple PDF documents in batch mode.
- Provide responsive dashboards for students and administrators.
- Maintain integrity of generated reports and processed data.

5. Proposed System Design and Workflow

The proposed system will be designed as a modular pipeline as shown in Figure 1.

Figure 1: Proposed Workflow Of The Marksheet Parser System



The system workflow includes the following stages

- Document upload and PDF format validation
- Template recognition for SSC, CBSE, and ICSE
- PDF parsing using PDFplumber and OCR-based text extraction
- Data parsing and schema mapping
- Result computation including totals, percentages, and grades
- Visualization using Chart.js
- Insights generation and reporting

5.1 Upload and Format Support

- The proposed parser will support student marksheets in PDF format only.
- Pre-validation will ensure file integrity before processing.
- Board-specific layouts will be identified using template heuristics.
- The system will attempt direct text extraction using PDFplumber for digital PDFs.
- OCR techniques will be used for scanned PDF documents when necessary.

5.2 PDF Parsing and OCR Integration

The system will employ a dual-approach strategy for text extraction

- PDFplumber will be utilized for digital PDFs containing selectable text.
- OCR engines such as Tesseract and PaddleOCR will be used for scanned PDF documents.
- Automatic detection of PDF type will determine the appropriate extraction method.
- Table structure recognition will be used for structured data extraction.
- Keyword spotting will assist in subject identification.
- Positional heuristics will map marks to relevant categories.
- Rule-based error correction will improve extraction accuracy.
- Confidence scoring will determine extraction reliability.

6. System Functionalities and Visualization

The proposed system will provide the following functionalities

- Upload and processing of student marksheets in PDF format
- Automatic extraction of personal and academic data
- Hybrid text extraction using PDFplumber and OCR technologies
- Computation of total marks, percentage, grade, and pass or fail status
- Export of processed data into structured formats such as JSON or CSV
- Error handling and correction suggestions
- Quality assessment and confidence scoring for extracted data

Visualization will assist in transforming extracted academic data into meaningful insights using Chart.js, including:

- Bar charts for subject-wise marks
- Pie and doughnut charts for distribution analysis
- Line charts for growth-over-time tracking
- Interactive dashboards with drill-down capabilities

7. Data Model and Growth Analysis

A unified schema will ensure consistency across outputs regardless of the extraction method used. The system will employ a hierarchical JSON structure that captures student information, examination details, and analytics data.

{
"processing_info": {
"total_files": 2,
"processed_at": "2026-02-08 20:25:14",
"input_folder": "marksheets"
},
"students": [
{
"filename": "example.pdf",
"student_info": {
"name": "STUDENT NAME",
"enrollment_no": "123456789",
"seat_no": "123456",
"examination": "SUMMER 2025",
"semester": "FOURTH SEMESTER",
"course": "Diploma In Computer Engineering"
},
"subjects": [
{
"subject": "SUBJECT NAME",
"fa_th_max": "030",
"fa_th_obt": "025",
"sa_th_max": "070",

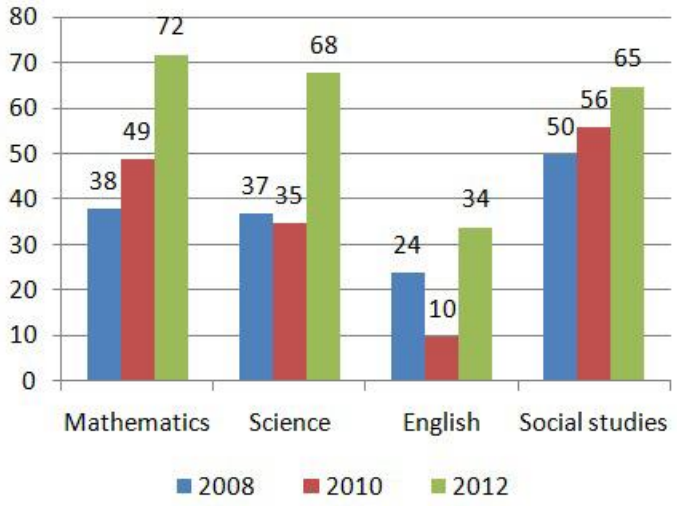
"sa_th_obt": "060",
"total_max": "100",
"total_obt": "085"
}
],
"summary": {
"total_max_marks": "850",
"total_marks_obtained": "746",
"percentage": "87.77",
"result": "FIRST CLASS WITH DISTINCTION"
}
}
]
}

The proposed parser will include a growth-tracking module capable of

- Aggregating multiple PDF marksheets for the same student
- Aligning subjects across different templates
- Visualizing progression across terms or semesters
- Identifying consistent strengths and weaknesses
- Providing historical benchmarks and predictive insights

Growth charts, as shown in Figure 2, will enable both students and administrators to monitor academic trajectories.

Figure 1: Proposed Growth Over Time Chart Interface For A Student



8. Analytics and User Interfaces

The system will integrate analytical features including

- Detection of best and weakest subjects
- Highlighting of marks below defined thresholds
- Automated performance summaries
- Semester-to-semester comparisons
- Batch-level analysis for institutional planning

Students will be able to access personal dashboards to view marks, interactive charts, and growth trends. Administrators will be able to batch-upload PDF marksheets, generate toppers lists, view average marks, and analyze institutional performance through filtering based on percentage, roll number, or pass or fail status.

9. Expected Results and Conclusion

It is anticipated that the proposed parser will achieve significant improvements in academic data processing. Direct PDF text extraction using PDFplumber is expected to achieve high accuracy for digital documents, while OCR methods will assist in processing scanned PDF marksheets.

The hybrid extraction approach ensures optimal accuracy across different document types while maintaining processing efficiency. Visualization modules are expected to be responsive across browsers, while administrative dashboards are expected to reduce processing time compared to manual data entry.

This system provides an integrated approach for extracting, analyzing, and visualizing student performance from PDF-based marksheets, thereby enabling automated performance evaluation and data-driven decision-making in academic institutions.

10. Acknowledgement

The authors would like to express their sincere gratitude to V. E. S. Polytechnic for providing the opportunity and platform to develop the proposed system. The authors also thank Vishwa International for supporting this project, which contributed to the successful completion of this work.

References

1. Eraser.io, Flowchart Generation Tool. https://www.eraser.io
2. Patil S., Marksheet Analysis Using OCR, Journal of Emerging Technologies and Innovative Research, 2023.
3. Sharma R., A Novel Implementation of Marksheet Parser Using PaddleOCR, arXiv, 2024.
4. PDFplumber Contributors, PDFplumber Documentation and Best Practices for Educational Document Processing. https://github.com/jsvine/pdfplumber