IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁝ Peer-reviewed & Refereed journal ⁝ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13484

# Image to Excel Conversion: A Methodology Proposal

**Girish Shewale[1], Nitesh Shinde[2], Jay More[3], Suraj Sahu[4], Prof. Geeta Arwindekar[5]**

Department of Information Technology, Datta Meghe College of Engineering, Airoli, Navi Mumbai

**Abstract:** This paper introduces a novel approach to text extraction and conversion using PyPDF2 technology, aimed at enhancing literacy education. A web application is developed to facilitate the conversion of images to Excel format, with a focus on leveraging PyPDF2 functionalities. The research investigates various methodologies for image-to-text conversion, highlighting the advantages and challenges associated with PyPDF2 compared to traditional OCR techniques. By addressing identified gaps in existing literature, the study presents a comprehensive methodology consisting of capturing, extracting, recognizing, and convert ing phases within the web application. Unlike conventional OCR methods, PyPDF2 offers improved text processing and segmentation algorithms, resulting in enhanced accuracy and efficiency in text extraction. The web application seamlessly converts uploaded images into editable text, making it a valuable resource for both literacy education and teaching staff in diverse educational settings.

**Keywords**: Recognition; PyPDF2: Text Extraction

## I. INTRODUCTION

In today's digitally driven landscape, efficient data extraction from images or PDFs and its presentation in user-friendly formats, such as Excel sheets, is paramount. Our project aims to bridge the gap between image data and organized presentation, lever aging PyPDF2 technology to streamline this pro cess. From scanned documents to photographs, valuable information often remains concealed within images. Recognizing the necessity for a robust solution, our project is dedicated to extracting data from images and presenting it in easily comprehensible formats, with a primary focus on Excel integration. Across various industries, data extraction from images poses a common challenge, whether dealing with invoices, receipts, or other documents containing embedded information. Manual interpretation of such data is not only time-consuming but also prone to errors. Our project endeavours to revolutionize this workflow by automating the extraction pro cess, thereby saving time and mitigating the risk of inaccuracies. Unlike Optical Character Recognition (OCR), which traditionally relies on image-to-text translation, PyPDF2 offers distinct advantages in its approach, facilitating seamless extraction and presentation of data from PDF files. This introduction sets the stage for our exploration into the capabilities and potential of PyPDF2 technology in enhancing data extraction and presentation methodologies.

## II. LITERATURE SURVEY

*1)* Rifiana Arief, Achmad Benny Mutiara, Tubagus Maulana Kusuma, Hustinawaty they have researched on how the hierarchical classification of scanned documents with characteristics content that have unstructured text and special pat terns using convolutional neural network and regular expression method. The research data using digital correspondence documents with format PDF images from Pusat Data Technology dan Information. The document hierarchy covers type of letter, type of manuscript letter, origin of letter and subject of letter. The research method consists of preprocessing, classification, and storage to database. Hierarchical classification uses CNN to classify 5 types of letters and regular expression to classify 4 types of manuscript letter, 15 origins of letter and 25 subjects of letter. The classified documents are stored in the Hive database in Hadoop big data architecture.

*2)* Harshavardhan Seetha, Vimal Tiwari, Kartik Reddy Anugu , Shanthi Makka, Ramesh Karnati they have researched on how a Graphical User Interface (GUI) application for PDF processing tools and file conversion tools. The application provides a user-friendly interface for users to perform various operations on PDF documents, such as splitting, merging, extracting, rotating, and deleting pages. Depending on the user's needs, the user can perform these operations on every page, even pages, odd pages, random specific pages, all pages after some nth page, and between some specific ranges. Nowa days, most universities and schools provide their students with all of their course materials online, therefore many students access these resources via PDF files. Using the PyPDF2 library for PDFs, the Python graphics package Tkinter is used to develop the graphical user interface and UI library Custom Tkinter which provides fresh, modern, and completely customizable widgets. These tools are simple to use and can help users save time and effort.

*3)* Rajan Rayhan , Abu Rayhan , Robert Kinzler they have researched on how In the ever-expanding landscape of data-driven endeavours, the Python programming language has emerged as a stalwart companion, bolstered by a trio of libraries- NumPy, SciPy, and Pandas. This research paper delves into the intricacies of these libraries, unravelling their unique attributes and collective prowess in facilitating data manipulation and analysis. NumPy, the bedrock of numerical computing, provides efficient N-dimensional arrays and a panoply of operations. SciPy, an extension of NumPy, extends

DOI: 10.17148/IJARCCE.2024.13484

the repertoire with specialized submodules for optimization, integration, signal processing, and more. Pandas, on the other hand, introduces a versatile Data Frame structure, revolutionizing data manipulation with its intuitive interface and powerful functionalities.

*4)* Rohit Sahoo, Chinmay Kathale, Milind Kubal, Shaveta Malik In this paper, they have proposed a Machine Learning based system called Auto-Table-Extract. This tool identifies and extracts the tables from PDF documents and dumps the data into excel sheets. It works with all kinds of PDF containing bordered, border less, or partially bordered tables. This system can extract data from both searchable and scanned PDF. The system's performance is commensurate to other table detection and extraction methods, but it over comes limitations of both detecting borderless as well as partially bordered tables and proves to be an efficient solution for the detection of tables from diverse documents

## III. PROPOSED METHODOLOGY

*1)* Data Collection and Flow Analysis: In our project utilizing PyPDF2 technology, the document data is acquired in PDF format. The PDF files may contain various types of content, including text, images, and tables. Careful attention is given to ensure the quality of the PDF files, ensuring they are not corrupted or damaged during acquisition

*2)* Text Extraction: PyPDF2 library is employed to extract text from the PDF documents. This involves parsing the PDF files and extracting textual content present within them. The library allows for efficient extraction of text elements, preserving formatting and structure as much as possible.

*3)* Data Transformation: Once the text and image data are extracted, they are transformed into a structured format suitable for further analysis and presentation. This may involve converting extracted text into tabular form or organizing image data into a format compatible with the desired output.

*4)* Segmentation: The image after undergoing Thresholding and Noise removal undergoes a process known as Segmentation. The image undergoes vertical and horizontal segmentation. In horizontal segmentation, the lines of char acters are separated from each other in a horizontal manner, i.e., each row is separated. These separated rows are then made to undergo vertical segmentation. During vertical segmentation, the characters in each row are separated from each other. Thus after segmentation, we get individual character images.

*5)* Integration with Excel The transformed data is then integrated into Excel sheets using the Pandas library. This step involves creating Excel worksheets and populating them with the extracted and transformed data. The resulting Excel sheets provide a user-friendly and organized representation of the data extracted from the PDF documents.

## IV. PROPOSED SYSTEM

Our proposed system utilizes PyPDF2 technology to efficiently extract and process text data from PDF documents. It includes the following components: PDF Document Acquisition: Users can upload PDF documents containing printed text to the system for processing. Text Extraction: The PyPDF2 library is utilized to extract text elements from the PDF files. Image Processing (if applicable): Optional image processing techniques may be employed to extract text from images within the PDF documents. Text Optimization: Processed text is optimized for readability and usability. Output Generation: The system generates machine-encoded text extracted from the PDF documents, which can be saved in various formats. User Interface: A user-friendly interface allows seamless interaction with the system, including up loading documents and accessing converted text. Overall, our system aims to streamline text extraction from PDFs, providing a reliable tool for users to convert scanned images of printed text into machine encoded text for further analysis and editing.

*1)* PyPDF2: PyPDF2 is a Python library for working with PDF files. It allows users to manipulate PDF documents by extracting text, merging or splitting pages, adding watermarks, and more. PyPDF2 enables developers to perform various operations on PDF files programmatically, making it a versatile tool for PDF processing tasks.

*2)* Pandas: Pandas is a powerful Python library for data manipulation and analysis. It provides data structures and functions designed to work with structured or tabular data, making it an essential tool for tasks such as data cleaning, trans formation, exploration, and visualization.

*3)* NLP(Natural Language Processing): NLP techniques were used for processing and understanding the extracted text data. This includes tasks such as text preprocessing, tokenization, and semantic analysis.

*4)* Regular Expression: Regular expressions were employed for pattern matching and text manipulation tasks. They are useful for extracting specific information from the text data.

### *A.* The Suggested Method:

*1)* PDF Document Acquisition: Users upload PDF documents containing printed text to the system for processing.

*2)* Text Extraction: Utilizing the PyPDF2 library, the system extracts text elements from the PDF files.

DOI: 10.17148/IJARCCE.2024.13484

*3)* Data Preprocessing: Data preprocessing involves any processing per formed on raw data to prepare it for further pro cessing. We utilize NumPy and OpenCV for preprocessing image and video data, ensuring optimal data quality for subsequent steps.

*4)* Conversion of PDF to Excel: Finally, the extracted text data is converted into an Excel file. A dictionary is created to map the extracted text to the corresponding cells in the table structure. This process involves looping over each cell, extracting the text content, and storing it in the Excel file for further data processing.

*B.* System Diagram:

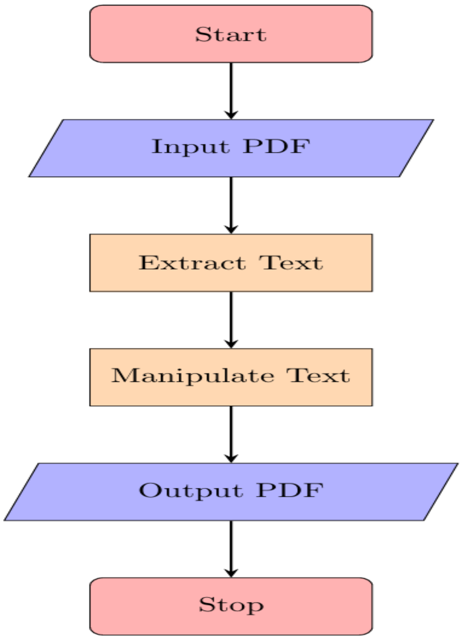
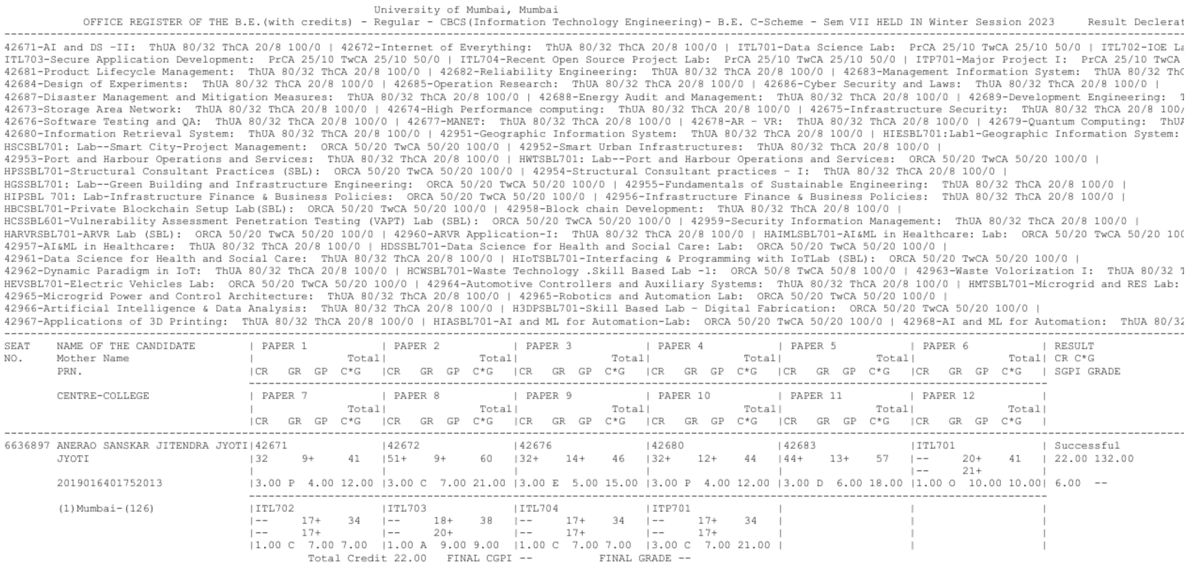


Fig. 1 Project Flowchart

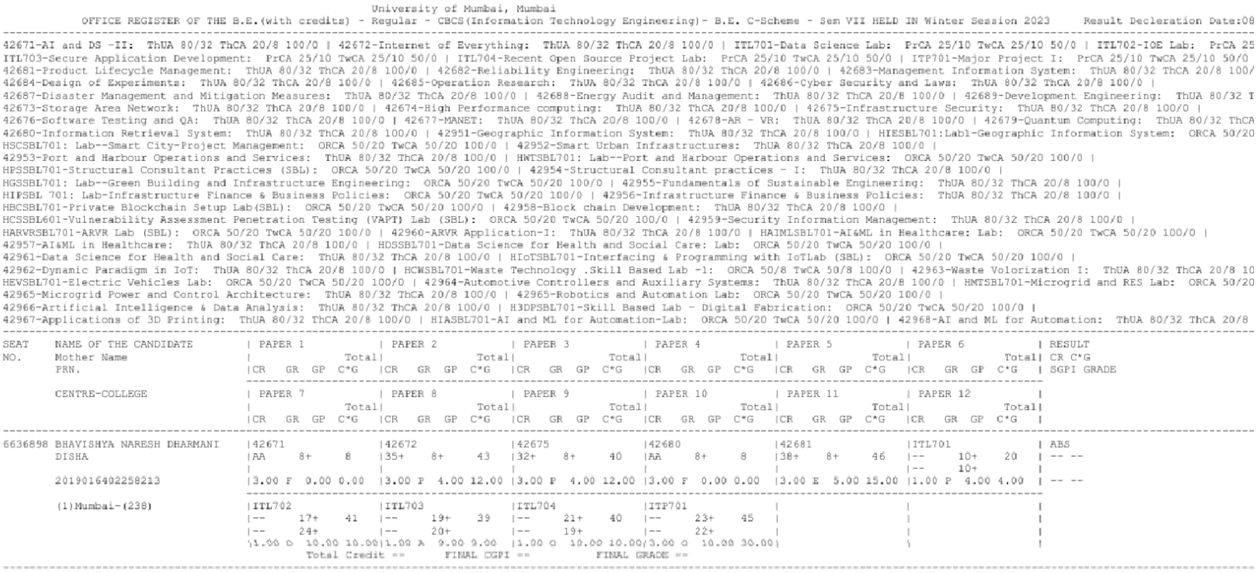
## V. IMPLEMENTATION



```
                                                 University of Mumbai, Mumbai
          OFFICE REGISTER OF THE B.E.(with credits) - Regular - CBCS(Information Technology Engineering)- B.E. C-Scheme - Sem VII HELD IN Winter Session 2023     Result Declerat
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
42671-AI and DS -II:  ThUA 80/32 ThCA 20/8 100/0 | 42672-Internet of Everything:  ThUA 80/32 ThCA 20/8 100/0 | ITL701-Data Science Lab:  PrCA 25/10 TwCA 25/10 50/0 | ITL702-IOE La
ITL703-Secure Application Development:  PrCA 25/10 TwCA 25/10 50/0 | ITL704-Recent Open Source Project Lab:  PrCA 25/10 TwCA 25/10 50/0 | ITP701-Major Project I:  PrCA 25/10 TwCA
42681-Product Lifecycle Management:  ThUA 80/32 ThCA 20/8 100/0 | 42682-Reliability Engineering:  ThUA 80/32 ThCA 20/8 100/0 | 42683-Management Information System:  ThUA 80/32 ThC
42684-Design of Experiments:  ThUA 80/32 ThCA 20/8 100/0 | 42685-Operation Research:  ThUA 80/32 ThCA 20/8 100/0 | 42686-Cyber Security and Laws:  ThUA 80/32 ThCA 20/8 100/0 |
42687-Disaster Management and Mitigation Measures:  ThUA 80/32 ThCA 20/8 100/0 | 42688-Energy Audit and Management:  ThUA 80/32 ThCA 20/8 100/0 | 42689-Development Engineering:  T
42673-Storage Area Network:  ThUA 80/32 ThCA 20/8 100/0 | 42674-High Performance computing:  ThUA 80/32 ThCA 20/8 100/0 | 42675-Infrastructure Security:  ThUA 80/32 ThCA 20/8 100/
42676-Software Testing and QA:  ThUA 80/32 ThCA 20/8 100/0 | 42677-MANET:  ThUA 80/32 ThCA 20/8 100/0 | 42678-AR - VR:  ThUA 80/32 ThCA 20/8 100/0 | 42679-Quantum Computing:  ThUA
42680-Information Retrieval System:  ThUA 80/32 ThCA 20/8 100/0 | 42951-Geographic Information System:  ThUA 80/32 ThCA 20/8 100/0 | HIESBL701:Lab1-Geographic Information System:
HSCSBL701: Lab--Smart City-Project Management:  ORCA 50/20 TwCA 50/20 100/0 | 42952-Smart Urban Infrastructures:  ThUA 80/32 ThCA 20/8 100/0 |
42953-Port and Harbour Operations and Services:  ThUA 80/32 ThCA 20/8 100/0 | HWTSBL701: Lab--Port and Harbour Operations and Services:  ORCA 50/20 TwCA 50/20 100/0 |
HPSSBL701-Structural Consultant Practices (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42954-Structural Consultant practices - I:  ThUA 80/32 ThCA 20/8 100/0 |
HGSSBL701: Lab--Green Building and Infrastructure Engineering:  ORCA 50/20 TwCA 50/20 100/0 | 42955-Fundamentals of Sustainable Engineering:  ThUA 80/32 ThCA 20/8 100/0 |
HIPSBL 701: Lab-Infrastructure Finance & Business Policies:  ORCA 50/20 TwCA 50/20 100/0 | 42956-Infrastructure Finance & Business Policies:  ThUA 80/32 ThCA 20/8 100/0 |
HBCSBL701-Private Blockchain Setup Lab(SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42958-Block chain Development:  ThUA 80/32 ThCA 20/8 100/0 |
HCSSBL601-Vulnerability Assessment Penetration Testing (VAPT) Lab (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42959-Security Information Management:  ThUA 80/32 ThCA 20/8 100/0 |
HARVRSBL701-ARVR Lab (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42960-ARVR Application-I:  ThUA 80/32 ThCA 20/8 100/0 | HAIMLSBL701-AI&ML in Healthcare: Lab:  ORCA 50/20 TwCA 50/20 10
42957-AI&ML in Healthcare:  ThUA 80/32 ThCA 20/8 100/0 | HDSSBL701-Data Science for Health and Social Care: Lab:  ORCA 50/20 TwCA 50/20 100/0 |
42961-Data Science for Health and Social Care:  ThUA 80/32 ThCA 20/8 100/0 | HIoTSBL701-Interfacing & Programming with IoTLab (SBL):  ORCA 50/20 TwCA 50/20 100/0 |
42962-Dynamic Paradigm in IoT:  ThUA 80/32 ThCA 20/8 100/0 | HCWSBL701-Waste Technology .Skill Based Lab -1:  ORCA 50/8 TwCA 50/8 100/0 | 42963-Waste Volorization I:  ThUA 80/32 T
HEVSBL701-Electric Vehicles Lab:  ORCA 50/20 TwCA 50/20 100/0 | 42964-Automotive Controllers and Auxiliary Systems:  ThUA 80/32 ThCA 20/8 100/0 | HMTSBL701-Microgrid and RES Lab:
42965-Microgrid Power and Control Architecture:  ThUA 80/32 ThCA 20/8 100/0 | 42965-Robotics and Automation Lab:  ORCA 50/20 TwCA 50/20 100/0 |
42966-Artificial Intelligence & Data Analysis:  ThUA 80/32 ThCA 20/8 100/0 | H3DPSBL701-Skill Based Lab - Digital Fabrication:  ORCA 50/20 TwCA 50/20 100/0 |
42967-Applications of 3D Printing:  ThUA 80/32 ThCA 20/8 100/0 | HIASBL701-AI and ML for Automation-Lab:  ORCA 50/20 TwCA 50/20 100/0 | 42968-AI and ML for Automation:  ThUA 80/32
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
SEAT    NAME OF THE CANDIDATE           | PAPER 1            | PAPER 2            | PAPER 3            | PAPER 4            | PAPER 5            | PAPER 6            | RESULT
NO.     Mother Name                     |               Total|               Total|               Total|               Total|               Total|               Total| CR C*G
        PRN.                            |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   | SGPI GRADE
                                        -------------------------------------------------------------------------------------------------------------------------
        CENTRE-COLLEGE                  | PAPER 7            | PAPER 8            | PAPER 9            | PAPER 10           | PAPER 11           | PAPER 12           |
                                        |               Total|               Total|               Total|               Total|               Total|               Total|
                                        |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
6636897 ANERAO SANSKAR JITENDRA JYOTI|42671              |42672              |42676              |42680              |42683              |ITL701             | Successful
        JYOTI                        |32      9+     41  |51+     9+     60  |32+     14+    46  |32+     12+    44  |44+     13+    57  |--      20+    41  | 22.00 132.00
                                     |                   |                   |                   |                   |                   |--      21+        |
        2019016401752013             |3.00 P  4.00 12.00 |3.00 C  7.00 21.00 |3.00 E  5.00 15.00 |3.00 P  4.00 12.00 |3.00 D  6.00 18.00 |1.00 O  10.00 10.00| 6.00   --
                                     -------------------------------------------------------------------------------------------------------------------------
        (1)Mumbai-(126)              |ITL702             |ITL703             |ITL704             |ITP701             |                   |                   |
                                     |--      17+    34  |--      18+    38  |--      17+    34  |--      17+    34  |                   |                   |
                                     |--      17+        |--      20+        |--      17+        |--      17+        |                   |                   |
                                     |1.00 C  7.00 7.00  |1.00 A  9.00 9.00  |1.00 C  7.00 7.00  |3.00 C  7.00 21.00 |                   |                   |
                                             Total Credit 22.00   FINAL CGPI --          FINAL GRADE --
```

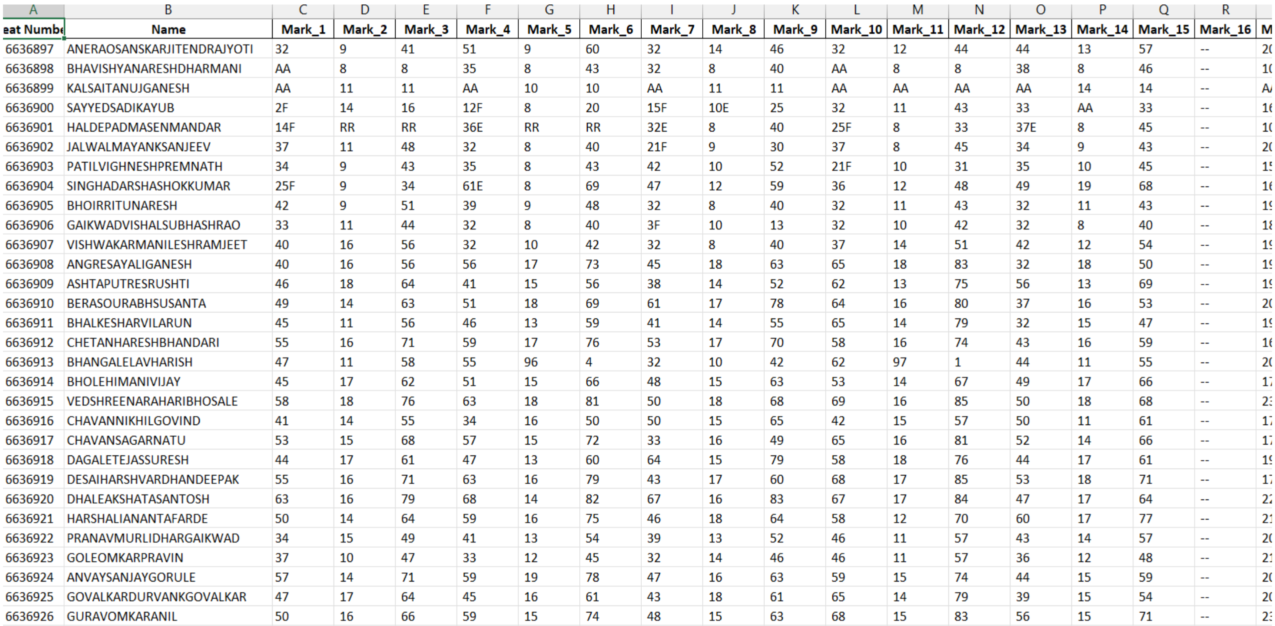
Fig. 2.1 Project Input 1

Fig. 2.2 Input 2



```
                                                  University of Mumbai, Mumbai
          OFFICE REGISTER OF THE B.E.(with credits) - Regular - CBCS(Information Technology Engineering)- B.E. C-Scheme - Sem VII HELD IN Winter Session 2023     Result Decleration Date:08
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
42671-AI and DS -II:  ThUA 80/32 ThCA 20/8 100/0 | 42672-Internet of Everything:  ThUA 80/32 ThCA 20/8 100/0 | ITL701-Data Science Lab:  PrCA 25/10 TwCA 25/10 50/0 | ITL702-IOE Lab:  PrCA 25
ITL703-Secure Application Development:  PrCA 25/10 TwCA 25/10 50/0 | ITL704-Recent Open Source Project Lab:  PrCA 25/10 TwCA 25/10 50/0 | ITP701-Major Project I:  PrCA 25/10 TwCA 25/10 50/0
42681-Product Lifecycle Management:  ThUA 80/32 ThCA 20/8 100/0 | 42682-Reliability Engineering:  ThUA 80/32 ThCA 20/8 100/0 | 42683-Management Information System:  ThUA 80/32 ThCA 20/8 100/
42684-Design of Experiments:  ThUA 80/32 ThCA 20/8 100/0 | 42685-Operation Research:  ThUA 80/32 ThCA 20/8 100/0 | 42686-Cyber Security and Laws:  ThUA 80/32 ThCA 20/8 100/0 |
42687-Disaster Management and Mitigation Measures:  ThUA 80/32 ThCA 20/8 100/0 | 42688-Energy Audit and Management:  ThUA 80/32 ThCA 20/8 100/0 | 42689-Development Engineering:  ThUA 80/32 T
42673-Storage Area Network:  ThUA 80/32 ThCA 20/8 100/0 | 42674-High Performance computing:  ThUA 80/32 ThCA 20/8 100/0 | 42675-Infrastructure Security:  ThUA 80/32 ThCA 20/8 100/0 |
42676-Software Testing and QA:  ThUA 80/32 ThCA 20/8 100/0 | 42677-MANET:  ThUA 80/32 ThCA 20/8 100/0 | 42678-AR - VR:  ThUA 80/32 ThCA 20/8 100/0 | 42679-Quantum Computing:  ThUA 80/32 ThCA
42680-Information Retrieval System:  ThUA 80/32 ThCA 20/8 100/0 | 42951-Geographic Information System:  ThUA 80/32 ThCA 20/8 100/0 | HIESBL701:Lab1-Geographic Information System:  ORCA 50/20
HSCSBL701: Lab--Smart City-Project Management:  ORCA 50/20 TwCA 50/20 100/0 | 42952-Smart Urban Infrastructures:  ThUA 80/32 ThCA 20/8 100/0 |
42953-Port and Harbour Operations and Services:  ThUA 80/32 ThCA 20/8 100/0 | HWTSBL701: Lab--Port and Harbour Operations and Services:  ORCA 50/20 TwCA 50/20 100/0 |
HPSSBL701-Structural Consultant Practices (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42954-Structural Consultant practices - I:  ThUA 80/32 ThCA 20/8 100/0 |
HGSSBL701: Lab--Green Building and Infrastructure Engineering:  ORCA 50/20 TwCA 50/20 100/0 | 42955-Fundamentals of Sustainable Engineering:  ThUA 80/32 ThCA 20/8 100/0 |
HIPSBL 701: Lab-Infrastructure Finance & Business Policies:  ORCA 50/20 TwCA 50/20 100/0 | 42956-Infrastructure Finance & Business Policies:  ThUA 80/32 ThCA 20/8 100/0 |
HBCSBL701-Private Blockchain Setup Lab(SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42958-Block chain Development:  ThUA 80/32 ThCA 20/8 100/0 |
HCSSBL601-Vulnerability Assessment Penetration Testing (VAPT) Lab (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42959-Security Information Management:  ThUA 80/32 ThCA 20/8 100/0 |
HARVRSBL701-ARVR Lab (SBL):  ORCA 50/20 TwCA 50/20 100/0 | 42960-ARVR Application-I:  ThUA 80/32 ThCA 20/8 100/0 | HAIMLSBL701-AI&ML in Healthcare: Lab:  ORCA 50/20 TwCA 50/20 100/0 |
42957-AI&ML in Healthcare:  ThUA 80/32 ThCA 20/8 100/0 | HDSSBL701-Data Science for Health and Social Care: Lab:  ORCA 50/20 TwCA 50/20 100/0 |
42961-Data Science for Health and Social Care:  ThUA 80/32 ThCA 20/8 100/0 | HIoTSBL701-Interfacing & Programming with IoTLab (SBL):  ORCA 50/20 TwCA 50/20 100/0 |
42962-Dynamic Paradigm in IoT:  ThUA 80/32 ThCA 20/8 100/0 | HCWSBL701-Waste Technology .Skill Based Lab -1:  ORCA 50/8 TwCA 50/8 100/0 | 42963-Waste Volorization I:  ThUA 80/32 ThCA 20/8 10
HEVSBL701-Electric Vehicles Lab:  ORCA 50/20 TwCA 50/20 100/0 | 42964-Automotive Controllers and Auxiliary Systems:  ThUA 80/32 ThCA 20/8 100/0 | HMTSBL701-Microgrid and RES Lab:  ORCA 50/20
42965-Microgrid Power and Control Architecture:  ThUA 80/32 ThCA 20/8 100/0 | 42965-Robotics and Automation Lab:  ORCA 50/20 TwCA 50/20 100/0 |
42966-Artificial Intelligence & Data Analysis:  ThUA 80/32 ThCA 20/8 100/0 | H3DPSBL701-Skill Based Lab - Digital Fabrication:  ORCA 50/20 TwCA 50/20 100/0 |
42967-Applications of 3D Printing:  ThUA 80/32 ThCA 20/8 100/0 | HIASBL701-AI and ML for Automation-Lab:  ORCA 50/20 TwCA 50/20 100/0 | 42968-AI and ML for Automation:  ThUA 80/32 ThCA 20/8
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
SEAT    NAME OF THE CANDIDATE         | PAPER 1             | PAPER 2             | PAPER 3             | PAPER 4             | PAPER 5             | PAPER 6             | RESULT
NO.     Mother Name                   |               Total|               Total|               Total|               Total|               Total|               Total| CR C*G
        PRN.                          |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   | SGPI GRADE
                                      ------------------------------------------------------------------------------------------------------------------------------
        CENTRE-COLLEGE                | PAPER 7             | PAPER 8             | PAPER 9             | PAPER 10            | PAPER 11            | PAPER 12            |
                                      |               Total|               Total|               Total|               Total|               Total|               Total|
                                      |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |CR   GR  GP  C*G   |
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
6636898 BHAVISHYA NARESH DHARMANI     |42671               |42672               |42675               |42680               |42681               |ITL701              | ABS
        DISHA                         |AA     8+      8    |35+    8+      43   |32+    8+      40   |AA     8+      8    |38+    8+      46   |--     10+     20   | -- --
                                      |                    |                    |                    |                    |                    |--     10+          |
        2019016402258213              |3.00 F  0.00 0.00   |3.00 P  4.00 12.00  |3.00 P  4.00 12.00  |3.00 F  0.00 0.00   |3.00 E  5.00 15.00  |1.00 P  4.00 4.00   | -- --
                                      ------------------------------------------------------------------------------------------------------------------------------
         (1)Mumbai-(238)              |ITL702              |ITL703              |ITL704              |ITP701              |                    |                    |
                                      |--     17+     41   |--     19+     39   |--     21+     40   |--     23+     45   |                    |                    |
                                      |--     24+          |--     20+          |--     19+          |--     22+          |                    |                    |
                                      |1.00 O  10.00 10.00|1.00 A  9.00 9.00   |1.00 O  10.00 10.00|3.00 O  10.00 30.00|                    |                    |
                                              Total Credit ==       FINAL CGPI ==          FINAL GRADE ==
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Fig. 2.2 Project Input 2



| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| eat Numbe | Name | Mark_1 | Mark_2 | Mark_3 | Mark_4 | Mark_5 | Mark_6 | Mark_7 | Mark_8 | Mark_9 | Mark_10 | Mark_11 | Mark_12 | Mark_13 | Mark_14 | Mark_15 | Mark_16 | M |
| 6636897 | ANERAOSANSKARJITENDRAJYOTI | 32 | 9 | 41 | 51 | 9 | 60 | 32 | 14 | 46 | 32 | 12 | 44 | 44 | 13 | 57 | -- | 20 |
| 6636898 | BHAVISHYANARESHDHARMANI | AA | 8 | 8 | 35 | 8 | 43 | 32 | 8 | 40 | AA | 8 | 8 | 38 | 8 | 46 | -- | 10 |
| 6636899 | KALSAITANUJGANESH | AA | 11 | 11 | AA | 10 | 10 | AA | 11 | 11 | AA | AA | AA | AA | 14 | 14 | -- | AA |
| 6636900 | SAYYEDSADIKAYUB | 2F | 14 | 16 | 12F | 8 | 20 | 15F | 10E | 25 | 32 | 11 | 43 | 33 | AA | 33 | -- | 16 |
| 6636901 | HALDEPADMASENMANDAR | 14F | RR | RR | 36E | RR | RR | 32E | 8 | 40 | 25F | 8 | 33 | 37E | 8 | 45 | -- | 10 |
| 6636902 | JALWALMAYANKSANJEEV | 37 | 11 | 48 | 32 | 8 | 40 | 21F | 9 | 30 | 37 | 8 | 45 | 34 | 9 | 43 | -- | 20 |
| 6636903 | PATILVIGHNESHPREMNATH | 34 | 9 | 43 | 35 | 8 | 43 | 42 | 10 | 52 | 21F | 10 | 31 | 35 | 10 | 45 | -- | 15 |
| 6636904 | SINGHADARSHASHOKKUMAR | 25F | 9 | 34 | 61E | 8 | 69 | 47 | 12 | 59 | 36 | 12 | 48 | 49 | 19 | 68 | -- | 16 |
| 6636905 | BHOIRRITUNARESH | 42 | 9 | 51 | 39 | 9 | 48 | 32 | 8 | 40 | 32 | 11 | 43 | 32 | 11 | 43 | -- | 19 |
| 6636906 | GAIKWADVISHALSUBHASHRAO | 33 | 11 | 44 | 32 | 8 | 40 | 3F | 10 | 13 | 32 | 10 | 42 | 32 | 8 | 40 | -- | 18 |
| 6636907 | VISHWAKARMANILESHRAMJEET | 40 | 16 | 56 | 32 | 10 | 42 | 32 | 8 | 40 | 37 | 14 | 51 | 42 | 12 | 54 | -- | 19 |
| 6636908 | ANGRESAYALIGANESH | 40 | 16 | 56 | 56 | 17 | 73 | 45 | 18 | 63 | 65 | 18 | 83 | 32 | 18 | 50 | -- | 19 |
| 6636909 | ASHTAPUTRESRUSHTI | 46 | 18 | 64 | 41 | 15 | 56 | 38 | 14 | 52 | 62 | 13 | 75 | 56 | 13 | 69 | -- | 19 |
| 6636910 | BERASOURABHSUSANTA | 49 | 14 | 63 | 51 | 18 | 69 | 61 | 17 | 78 | 64 | 16 | 80 | 37 | 16 | 53 | -- | 20 |
| 6636911 | BHALKESHARVILARUN | 45 | 11 | 56 | 46 | 13 | 59 | 41 | 14 | 55 | 65 | 14 | 79 | 32 | 15 | 47 | -- | 19 |
| 6636912 | CHETANHARESHBHANDARI | 55 | 16 | 71 | 59 | 17 | 76 | 53 | 17 | 70 | 58 | 16 | 74 | 43 | 16 | 59 | -- | 16 |
| 6636913 | BHANGALELAVHARISH | 47 | 11 | 58 | 55 | 96 | 4 | 32 | 10 | 42 | 62 | 97 | 1 | 44 | 11 | 55 | -- | 20 |
| 6636914 | BHOLEHIMANIVIJAY | 45 | 17 | 62 | 51 | 15 | 66 | 48 | 15 | 63 | 53 | 14 | 67 | 49 | 17 | 66 | -- | 17 |
| 6636915 | VEDSHREENARAHARIBHOSALE | 58 | 18 | 76 | 63 | 18 | 81 | 50 | 18 | 68 | 69 | 16 | 85 | 50 | 18 | 68 | -- | 23 |
| 6636916 | CHAVANNIKHILGOVIND | 41 | 14 | 55 | 34 | 16 | 50 | 50 | 15 | 65 | 42 | 15 | 57 | 50 | 11 | 61 | -- | 17 |
| 6636917 | CHAVANSAGARNATU | 53 | 15 | 68 | 57 | 15 | 72 | 33 | 16 | 49 | 65 | 16 | 81 | 52 | 14 | 66 | -- | 17 |
| 6636918 | DAGALETEJASSURESH | 44 | 17 | 61 | 47 | 13 | 60 | 64 | 15 | 79 | 58 | 18 | 76 | 44 | 17 | 61 | -- | 19 |
| 6636919 | DESAIHARSHVARDHANDEEPAK | 55 | 16 | 71 | 63 | 16 | 79 | 43 | 17 | 60 | 68 | 17 | 85 | 53 | 18 | 71 | -- | 17 |
| 6636920 | DHALEAKSHATASANTOSH | 63 | 16 | 79 | 68 | 14 | 82 | 67 | 16 | 83 | 67 | 17 | 84 | 47 | 17 | 64 | -- | 22 |
| 6636921 | HARSHALIANANTAFARDE | 50 | 14 | 64 | 59 | 16 | 75 | 46 | 18 | 64 | 58 | 12 | 70 | 60 | 17 | 77 | -- | 21 |
| 6636922 | PRANAVMURLIDHARGAIKWAD | 34 | 15 | 49 | 41 | 13 | 54 | 39 | 13 | 52 | 46 | 11 | 57 | 43 | 14 | 57 | -- | 20 |
| 6636923 | GOLEOMKARPRAVIN | 37 | 10 | 47 | 33 | 12 | 45 | 32 | 14 | 46 | 46 | 11 | 57 | 36 | 12 | 48 | -- | 21 |
| 6636924 | ANVAYSANJAYGORULE | 57 | 14 | 71 | 59 | 19 | 78 | 47 | 16 | 63 | 59 | 15 | 74 | 44 | 15 | 59 | -- | 20 |
| 6636925 | GOVALKARDURVANKGOVALKAR | 47 | 17 | 64 | 45 | 16 | 61 | 43 | 18 | 61 | 65 | 14 | 79 | 39 | 15 | 54 | -- | 20 |
| 6636926 | GURAVOMKARANIL | 50 | 16 | 66 | 59 | 15 | 74 | 48 | 15 | 63 | 68 | 15 | 83 | 56 | 15 | 71 | -- | 23 |

‘

Fig. 2.3 Output of Fig. 2.1 and Fig. 2.2

In our study, we utilized PyPDF2 technology to convert attendance register images into Excel files through text data extraction. We implemented stringent criteria to ensure image quality and legible hand writing, which contributed to the accuracy of our con versions. We observed that higher image resolution significantly enhanced the accuracy of text extraction and conversion process. Looking ahead, we identified several areas for future enhancements. One such area is the implementation of cursive handwriting recognition, which would broaden the applicability of our system. Addition ally, integrating our solution with mobile platforms would improve accessibility and usability, enabling users to capture attendance register images directly from their devices.

DOI: 10.17148/IJARCCE.2024.13484

## VI. CONCLUSIONS

In conclusion, this paper has outlined a method for converting images or PDF files into Excel sheets, facilitating the extraction and presentation of data contained within these documents. Leveraging PyPDF2 technology, we proposed a web application that al lows users to upload PDF files or sets of images containing tabular data, which are then converted into Excel files. Strict rules were enforced to ensure the quality of the input data, including requirements for file format (PDF), resolution, and adherence to specific image quality standards. These criteria were crucial for ensuring accurate text extraction and conversion by the PyPDF2 library. In summary, our research highlights the importance of adhering to strict standards for input data quality when utilizing PyPDF2 technology for text extraction and conversion. By following these guide lines, users can achieve optimal results and effectively convert images or PDF files into Excel sheets for further analysis and utilization.

## VII. FUTURE SCOPE

The future of PyPDF2 technology holds great promise for enhancing text extraction and document processing capabilities. One key area of focus involves improving the accuracy of text extraction, which may entail refining algorithms and techniques to better handle complex document layouts and improve char acter recognition. Additionally, there are opportunities to expand the functionalities of PyPDF2 beyond basic text extraction, potentially incorporating features such as image handling, annotation recognition, and metadata extraction. Optimizing real-time processing capabilities is another important aspect, as faster processing speeds and improved efficiency can enable quicker decision making and analysis. Integration with mobile plat forms could enhance accessibility, allowing users to perform document processing tasks directly from their smartphones or tablets. Furthermore, enhancing customization and flexibility in PyPDF2 work f lows can empower users to tailor the software to their specific needs and preferences. Advanced security features may also be incorporated into PyPDF2 to address concerns surrounding data privacy and integrity. This could include encryption capabilities, digital signature verification, and secure handling of sensitive information. Lastly, advancements in document understanding techniques, such as natural language processing and machine learning, can enable deeper analysis and

## REFERENCES

[1] Pooja Jain, Dr. Kavita Taneja, Dr. Harmunish Taneja. (Year). ”Title of the paper.” Journal Name, Volume(Issue), page numbers. DOI or URL.
[2] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. (Year). ”Title of the paper.” Journal Name, Volume(Issue), page numbers. DOI or URL.
[3] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, Antoine Doucet. (Year). ”Title of the paper.” Journal Name, Volume(Issue), page numbers. DOI or URL.
[4] Chirag Patel, Atul Patel, Dharmendra Patel. (Year). ”Title of the paper.” Journal Name, Vol ume(Issue), page numbers. DOI or URL.
[5] The Regular Expression Inference Challenge (2308.07899.pdf (arxiv.org))
[6] Design & implementation of a PDF to Excel conversion tool (P2X)
[7] Penny, LaToyia DeVonne. Oklahoma State University ProQuest Dissertations Publishing, 2008. 1467427.