



# Marksheet Analysis Using OCR

<sup>1</sup>Swaraj Kakade, <sup>2</sup>Tushar Patle, <sup>3</sup>Rohan Dable, <sup>4</sup>Yash Nathe, <sup>5</sup>Prof. Pranjali Bahalkar

<sup>1,2,3,4</sup>Student, <sup>5</sup>Professor, Department of Artificial Intelligence and Data Science, D. Y. Patil College of Engineering, Pune, Maharashtra, India

**Abstract :** In this paper, a major advancement in educational data management—Marksheet Analysis with OCR—is presented. The need to quickly extract, store, and analyze marksheet data is growing in today's data-driven educational environment. This is because marksheet data is essential for gauging student accomplishment and supporting administrative choices. Conventional data entry techniques are ineffective and prone to errors. The system reduces human data entry errors and manual labor by automating the extraction of text data from scanned or photographed mark sheets using Optical Character Recognition (OCR) technology. Strong database management for safe storage and retrieval is complemented by the validated data, which guarantees a reliable platform for analysis. Data-driven insights are provided to administrators by Python libraries like Pandas and Matplotlib, which generate reports, statistics, and visualizations. Adoption is made easy with comprehensive documentation and user-friendly interfaces.

**Index Terms:** OCR, marksheet analysis, educational data management, Python, Pandas, Matplotlib, data visualization, predictive analytics.

## I. INTRODUCTION

There has been a lot of interest in the Marksheet Analysis using OCR project since there is a growing need for efficient data management in education. Educational institutions face difficulties with traditional manual mark data entry because it is labor-intensive and prone to errors. By automating the extraction and analysis of academic mark sheets using Optical Character Recognition (OCR) technology, this initiative seeks to revolutionize data management. Colleges have a lot of mark sheets, which makes manual entering ineffective. This procedure is automated by the "Marksheet Analysis using OCR" project, which uses cutting-edge technology to guarantee accuracy and efficiency. This automation facilitates data-driven decision-making, which is essential for developing curricula, helping students, and assessing performance. Additionally, by freeing up critical time and resources, it enables school administrators to concentrate on strategic duties. This idea presents a potential alternative that will increase efficiency for contemporary educational institutions.

## II. PROBLEM STATEMENT

It is extremely difficult for schools to manage student data effectively in the quickly changing educational environment. There is a critical need for creative solutions because the conventional approach of manually entering and processing marksheet data is laborious and prone to mistakes. The enormous amount of mark sheets that schools handle exacerbates this issue by making manual entering error-prone and inefficient. A comprehensive solution like the OCR-based Marksheet Analysis solution, which is intended to ensure accuracy and efficiency and streamline data management procedures, is desperately needed to address these issues. The wide variety of marksheet formats that educational institutions deal with causes confusion and inefficiencies in data processing. Staff decision fatigue is being decreased and the extraction process is becoming simpler with the OCR-based Marksheet Analysis System. Different scanned document quality levels can result in inconsistent data extraction. To ensure smooth data extraction even with fluctuations in scan quality, the system must exhibit a high degree of accuracy in OCR recognition. This dependability is essential to preserving the accuracy of educational judgments and guaranteeing the integrity of student data. The manual research and mark comparison process used in the old method is time-consuming and labour-intensive. From scanned or photographed marksheet documents, the OCR-based technology automatically extracts the necessary data, including student information like names and roll numbers, subject names and the scores they correspond with, grades, and any other pertinent information. The amount of time and effort needed for data entry and processing is greatly decreased by this automation. The process of entering and verifying data is made difficult by the unfriendly interfaces of many of the current solutions. Educational institutions can easily utilize the OCR-based Marksheet Analysis System because it has user-friendly interfaces for input, verification, and result visualization. This feature decreases the possibility of errors in marksheet analysis while also reducing manual labour. To produce detailed individual or aggregated performance reports, cumulative scores, and averages should be computed automatically from the retrieved data. Data-driven decision-making is supported by this real-time processing, which guarantees educational institutions have instant access to current and accurate performance data. By solving these important difficulties, the OCR-based Marksheet Analysis System presents a potential alternative for contemporary educational institutions, greatly increasing administrative efficiency and effectiveness.

### III. LITERATURE SURVEY

1) Title: Text Recognition and Classification in Floor plan Images

Published: 2019

This paper introduces a method for text recognition in floor plan images, addressing the need to accurately locate, read, and categorize text within these images to obtain detailed information about buildings. Convolutional neural networks (CNNs), a recent methodology, are contrasted with traditional image processing-based word identification algorithms. Multiple procedures are combined to improve accuracy, outperforming the original methods. Text regions are categorized into four semantic groups according to their intended use. Two datasets with different quality and size characteristics were used for the experiments. As a reliable approach for extracting building information, the results show how well the suggested strategy works to improve word recognition in floor plan photos.

2) Title: Automatic Number Plate Recognition Using TensorFlow and EasyOCR

Published: 2022

In this paper, an OCR (Optical Character Recognition) system for video processing-based vehicle number plate detection is presented. The system records video, transforms it into picture format, uses different algorithms to extract the text from license plates, and stores the data in a database. It offers a real-time vehicle identification system and is especially helpful for college entrances and other heavily restricted places. The interdisciplinary scientific topic of computer vision makes it possible for computers to comprehend digital photos or films. In computer vision, tasks include gathering, manipulating, deciphering, and comprehending digital images as well as extracting high-dimensional data to generate numerical or symbolic information.

3) Title: A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images

Published: 1988

This paper introduces an automated system for document analysis, aimed at efficiently segmenting digitized images that contain a mixture of text and graphics. The algorithm created is independent of changes in text font style, size, and orientation and focuses on automated text string separation. The algorithm does not specifically distinguish individual characters, in contrast to conventional techniques. It groups components into logical character strings by using the Hough transform and connected component creation. The resulting character strings are then extracted from the visuals. Two output images are generated by this method, one containing text strings and the other containing visuals. Systems for recognizing characters and pictures can process these outputs further. Several test photos were used to assess the algorithm's efficacy and computing efficiency, and the results showed that it performed better than alternative methods. The outcomes demonstrate the suggested algorithm's accuracy and resilience in document analysis applications.

4) Title: Automatic Number Plate Recognition Idea Development using AI-based ANNs

Published: 2018

This study uses OpenCV (Open-Source Computer Vision Library), a library of programming functions primarily aimed at real-time computer vision, to provide a thorough understanding of how to implement an automatic number plate recognition system using artificial neural networks, particularly in crowded areas and during the passing of automobiles along with the tolls in expressways and other locations. This facilitates faster transit by reducing the number of cars crammed around the tollways.

### IV. PROPOSED SOLUTION

The problems associated with maintaining student data in educational institutions can be fully resolved with the help of the Marksheet Analysis using OCR technology. Manual marksheet data entering by hand takes a lot of time and is prone to mistakes. By automating the extraction and analysis of marksheet data, this solution improves accuracy and efficiency. The extraction procedure is made simpler by the system's use of a single set of templates to handle a variety of marksheet formats. The accuracy of data extraction from different scan qualities is ensured by advanced OCR technology, protecting the integrity of student records. Labor-intensive manual data entry and comparison are involved. By automating the extraction of crucial data from scanned marksheets, such as student information, subject titles, scores, and grades, the OCR-based system greatly reduces processing time and effort. The system is straightforward to use for educational institutions, minimizing manual work and potential errors, thanks to user-friendly interfaces for input, verification, and result visualization. The system automatically processes the retrieved data in order to provide detailed performance reports, averages, and cumulative scores. Real-time processing facilitates data-driven decision-making by giving educational institutions instant access to precise and current performance data. The OCR-based Marksheet Analysis system improves administrative procedures by providing accurate and up-to-date information via an easy-to-use interface. This enables educational institutions to make well-informed decisions and enhance the administration of student data.

### V. METHODOLOGY

The development of the OCR-based Marksheet Analysis system involves several key steps to ensure efficient data extraction, storage, processing, and visualization. The process is designed to make use of Seaborn and Matplotlib for data visualization, Fast API for integration capabilities, Flask for dynamic website creation, and SQL for data storage. Teachers will also receive Excel sheets with the data generated by the system.

1) Data Collection and Preprocessing: Standardize marksheet templates to ensure consistency across different educational institutions. Preprocess scanned images using techniques such as noise reduction, contrast adjustment, and alignment correction to

enhance OCR accuracy and improve the quality of extracted data. Additionally, implement image enhancement algorithms to mitigate common issues such as shadows and smudges, ensuring optimal OCR performance.

2) OCR Implementation: To extract textual information from marksheets, such as student information, subject names, scores, and grades, use OCR technology. To manage differences in font sizes, styles, and orientations, apply machine learning techniques and sophisticated OCR algorithms. To improve overall data quality and dependability, implement error detection and correction procedures to address inaccuracies in the extracted data.

3) Data Storage: Create a SQL relational database schema to hold the marksheet data that is structured. Create processes or scripts that automatically enter OCR-extracted data into the database, guaranteeing consistency and integrity of the data. Use database optimization strategies to reduce storage cost and enhance query performance, such as indexing and normalization.

4) Dynamic Website Development: Build a dynamic website with user-friendly interfaces using Flask so that users can upload marksheets, confirm extracted data, and see analysis results. Use responsive design principles to make sure that your work will work on a range of screens and devices. Interactive features that improve user experience include real-time data validation and drag-and-drop file uploading.

5) API Integration: FastAPI can be used to create APIs that allow for easy integration with current academic systems. Establish RESTful API endpoints to enable data transfer between external apps and the OCR-based Marksheet Analysis system. Token-based authentication and role-based access control secure APIs that enforce data security and privacy policies.

6) Data Visualization: Utilize Seaborn and Matplotlib to create visual representations of the marksheet data, including histograms, bar charts, and line graphs. Incorporate interactive visualization libraries such as Plotly to enable users to explore and interact with the data dynamically. Implement customization options for visualizations, allowing users to adjust parameters such as color schemes, data aggregation levels, and chart types.

7) Report Generation: Implement functionality to export processed data into Excel sheets, formatted to provide clear and concise reports for teachers. Include summaries of student performance, detailed marks, attendance records, and any additional relevant information. Integrate customizable report templates to accommodate specific reporting requirements and preferences of educational institutions.

8) Testing and Validation: Conduct thorough testing of all system components, including OCR functionality, data insertion scripts, web application features, APIs, and visualizations. Employ automated testing frameworks and tools to validate system behavior under various scenarios and edge cases. Perform user acceptance testing with stakeholders to gather feedback and identify areas for improvement, ensuring the system meets the needs and expectations of educational institutions.

## VI. WORKING

**Data Collection:** The OCR-Based Marksheet Analysis System gathers marksheet data from educational institutions, accepting scanned or photographed marksheets in various formats. It ensures the collection of comprehensive student information, including names, roll numbers, subject names, scores, and grades, from multiple sources.

**User Interface:** The system provides an intuitive web interface developed using Flask, facilitating easy interaction for users. With a user-friendly design and layout, users can navigate the platform effortlessly, enhancing user experience and boosting engagement.

**Data Extraction and Validation:** Utilizing OCR technology, the system extracts textual data from marksheets with high accuracy, ensuring the correctness and completeness of extracted information. It employs validation algorithms to verify the accuracy of extracted data against predefined formats and ranges, minimizing errors.

**Data Storage:** Structured marksheet data is stored in a relational database using SQL, ensuring efficient data management and retrieval. Database optimization techniques such as indexing, and normalization are implemented to enhance query performance and minimize storage overhead.

**API Integration:** FastAPI is employed to create RESTful APIs, enabling seamless integration with existing academic systems. These APIs facilitate data exchange between the OCR-Based Marksheet Analysis system and external applications, ensuring interoperability and data consistency.

**Data Visualization:** Seaborn and Matplotlib are utilized to generate visual representations of marksheet data, including histograms, bar charts, and line graphs. Interactive visualization libraries enhance user experience, allowing users to explore and interact with data dynamically.

**Report Generation:** The system generates comprehensive reports in Excel format, providing detailed insights into student performance, attendance records, and other relevant metrics. Customizable report templates accommodate specific reporting requirements and preferences of educational institutions.

**Testing and Validation:** Thorough testing is conducted to validate system functionality, including OCR accuracy, data insertion processes, web application features, and APIs. User acceptance testing is performed to gather feedback and identify areas for improvement, ensuring the system meets user needs and expectations.

Deployment and Training: The system is deployed in educational institutions, accompanied by training sessions and comprehensive documentation for end-users. Ongoing support and maintenance activities ensure the system's smooth operation and address any issues that may arise.

Integration with Existing Systems: Seamless integration with existing academic record management systems is ensured, allowing for interoperability and data consistency across platforms. This integration enhances administrative efficiency and decision-making capabilities in educational institutions.

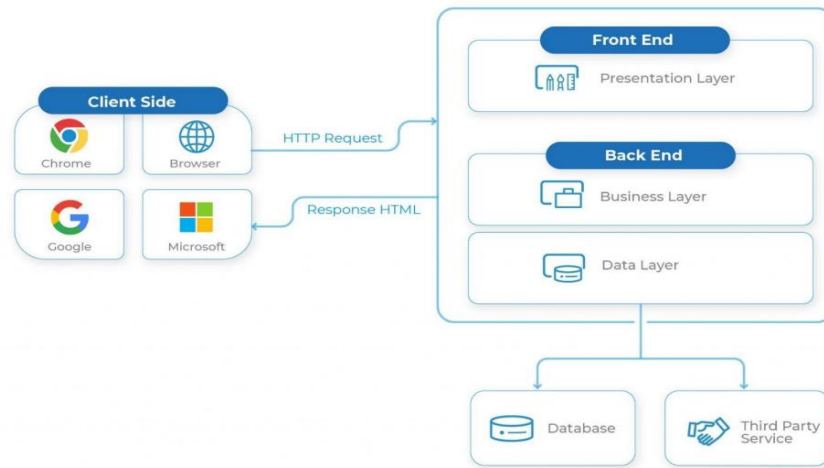


Figure 1. Architecture

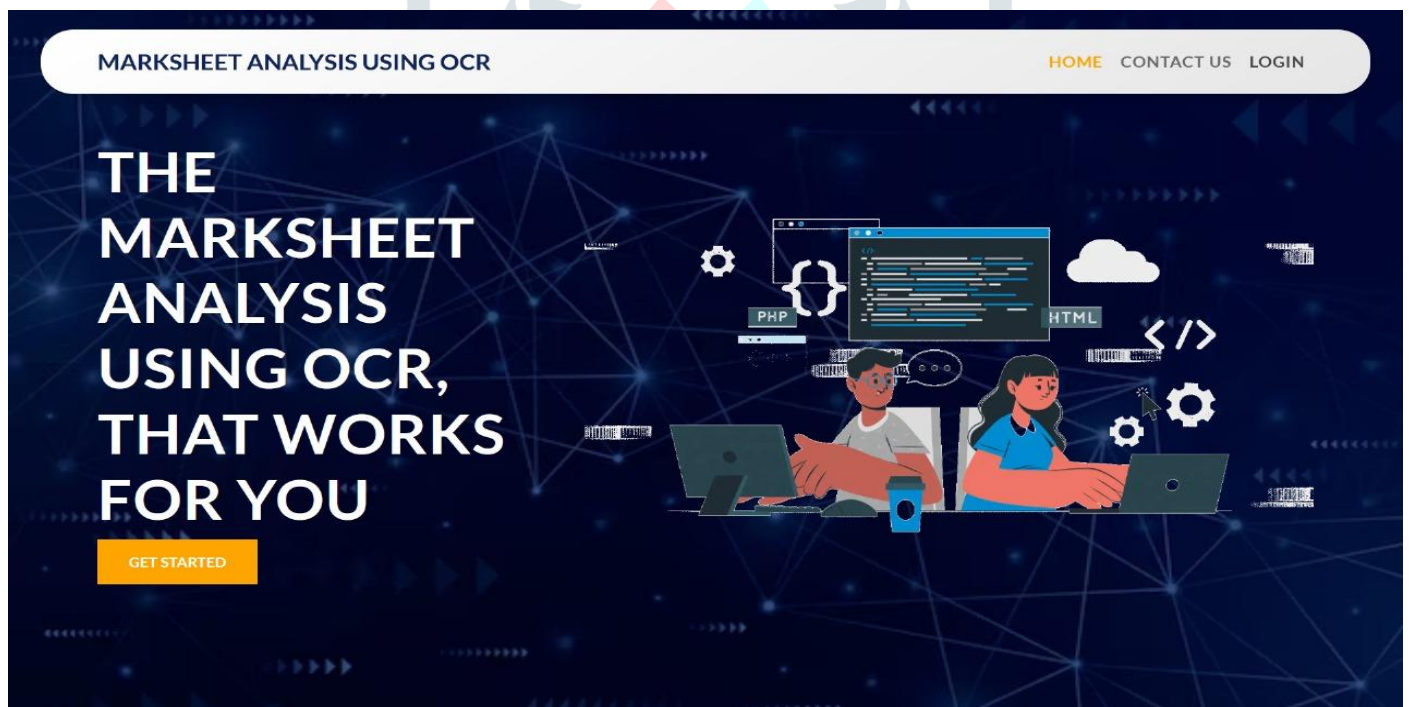


Figure 2. Website Homepage

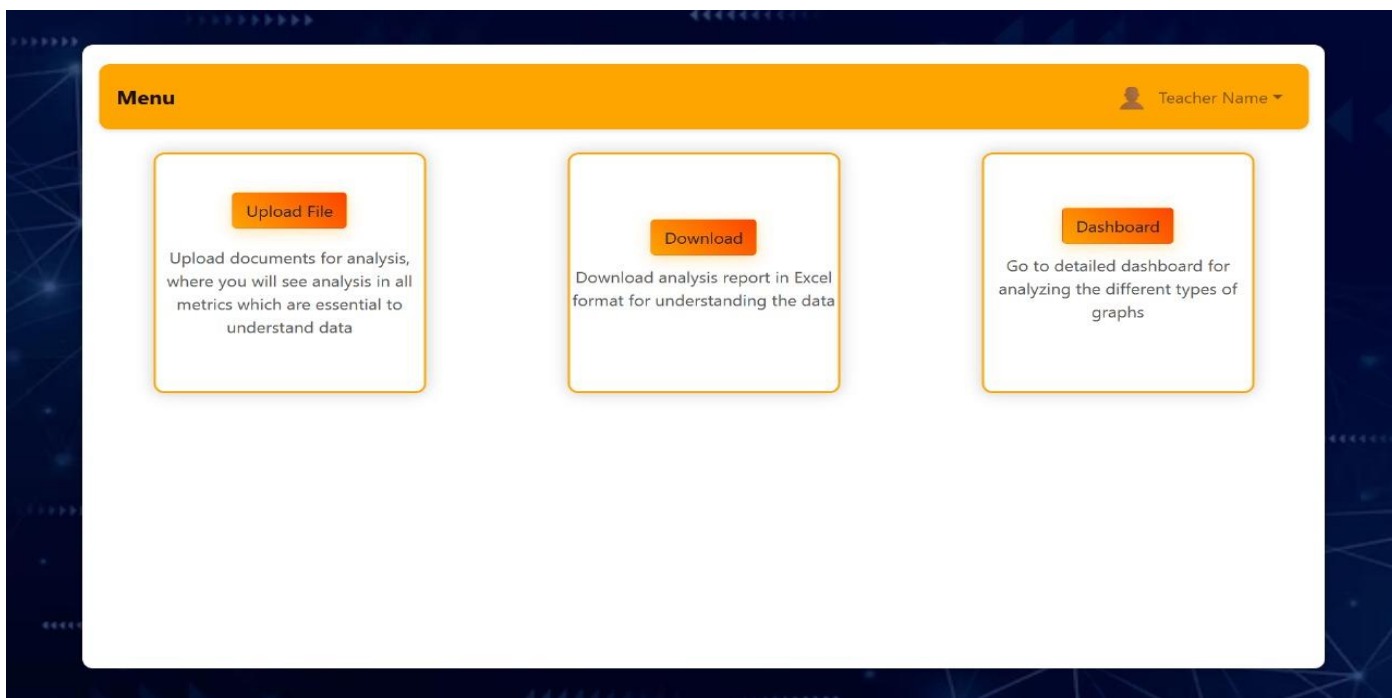


Figure 3. Teacher Functionality

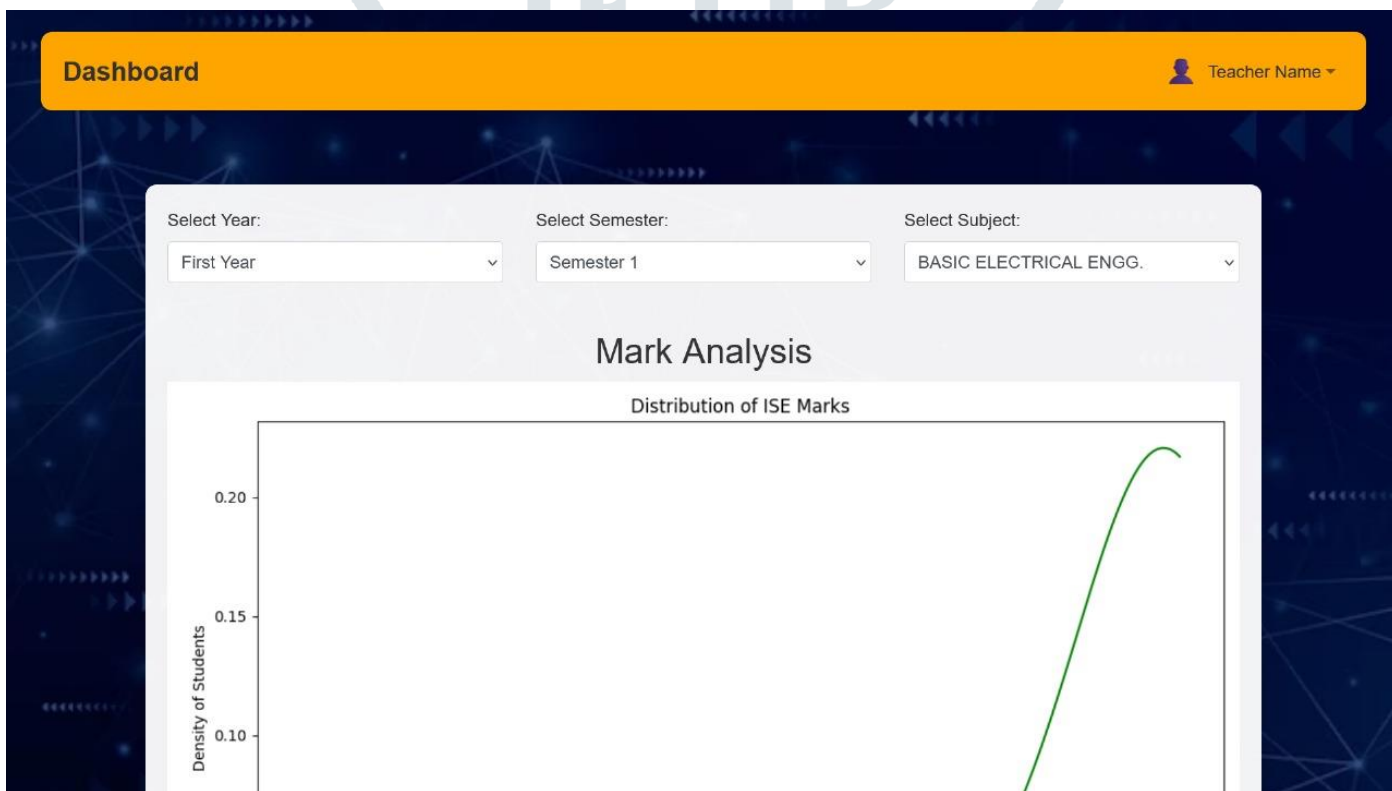


Figure 4. Teacher Classroom Analysis

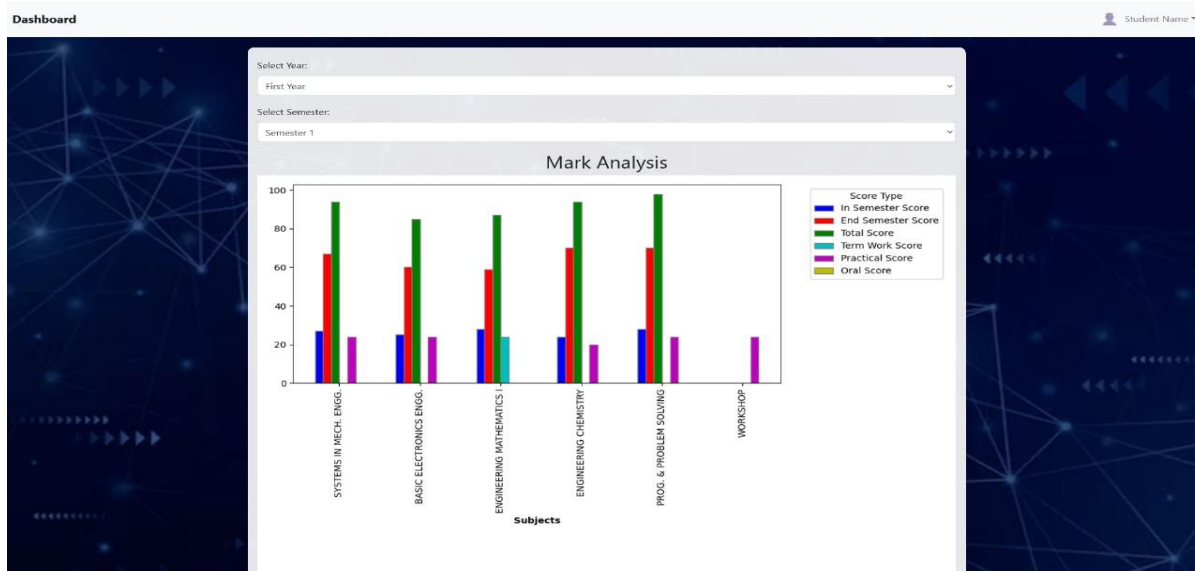


Figure 5. Student personal analysis webpage

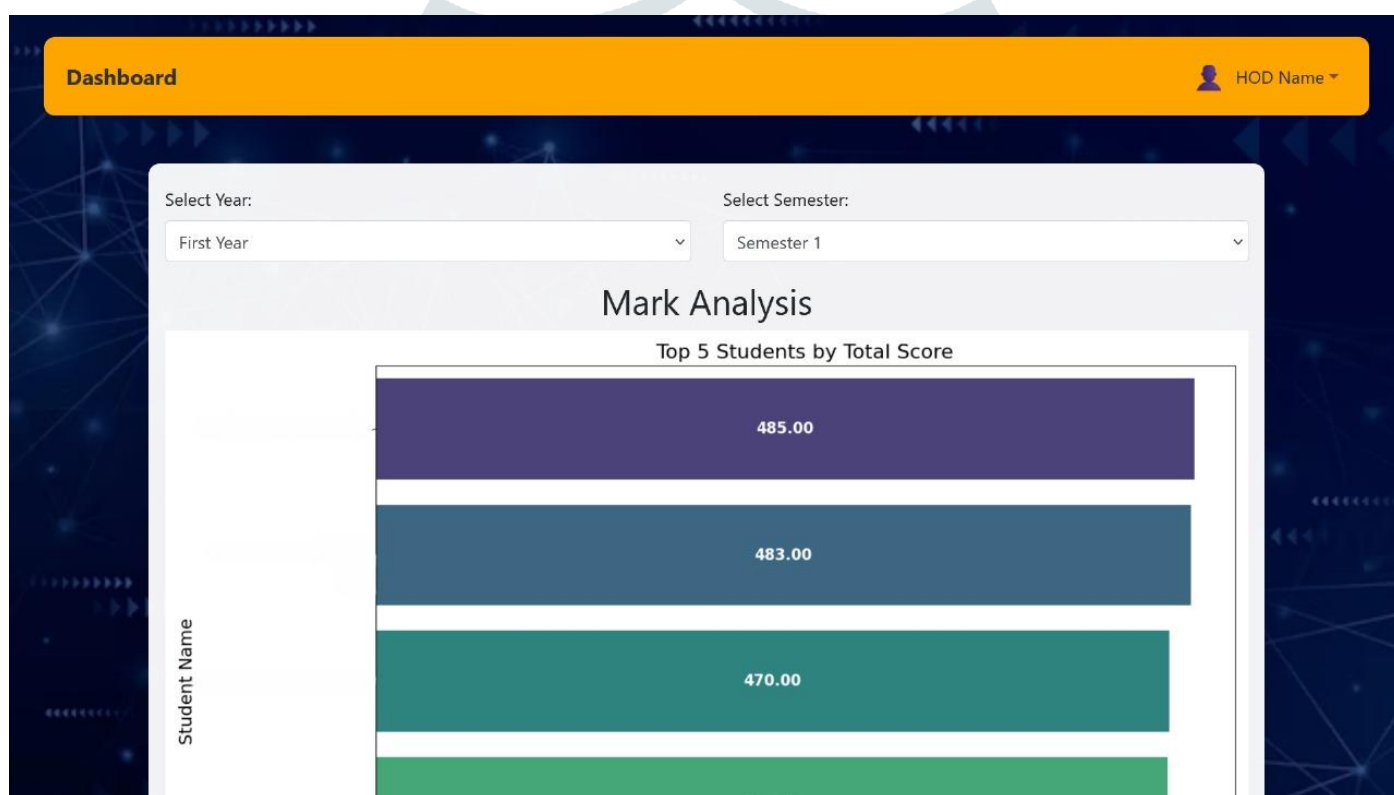


Figure 6. HOD Class analysis diagram

**VII. RELATED WORK**

*A. Impact on Educational Decision-Making:*

The goal of this research topic is to comprehend how decision-making procedures in educational institutions are impacted by the OCR-Based Marksheet Analysis method. Important findings from this research include:

*System Utilization:* To improve data management procedures, analysis shows that educational institutions frequently use the OCR-Based Marksheet Analysis system. Institutions extensively depend on this technology, especially those looking for precision and efficiency in marksheet analysis.

*Data Accuracy:* Results indicate that decision-making processes in educational contexts are positively impacted by the system's capacity to guarantee correct marksheet data extraction and validation. When educational administrators have access to accurate and current data, they are more likely to make well-informed judgments.

### B. User Experience and System Design:

The goal of this study field is to raise user engagement levels and optimize the OCR-Based Marksheet Analysis system's user experience. Important conclusions consist of:

*User-Friendly Interfaces:* Research emphasizes how crucial it is for the OCR-Based Marksheet Analysis system to have an intuitive and user-friendly interface. Platforms that are simple to use and comprehend are valued by users—including educators and administrators—and increase user happiness in general.

*Options for Search and Filtering:* According to research, adding search and filtering capabilities to the system enhances efficiency and user experience. Increased productivity and satisfaction arise from users' ability to swiftly search for specific student data or filter results based on parameters like grades or attendance.

To sum up, the first research topic emphasizes the critical function that the OCR-Based Marksheet Analysis system plays in highlighting the significance of data accuracy and reliability in educational decision-making. In order to promote user engagement and happiness, the second study area focuses on maximizing the user experience through interface design and functionality enhancements. The OCR-Based Marksheet Analysis system can be improved with the use of the insights gained from this research, increasing its usefulness for managing educational data and decision-making procedures.

## VIII. FUTURE SCOPE

*Integration with Learning Management Systems (LMS):* By integrating the system with well-known LMS platforms like Moodle or Canvas, educational institutions might simplify data management and enable the smooth transfer of grades and student performance information between systems.

*Enhanced Data Analytics Features:* Educational administrators would be better equipped to make data-driven choices if they implemented advanced data analytics capabilities, such as trend analysis and predictive analytics for academic achievement forecasts.

*Real-time Collaboration Tools:* By enabling real-time collaboration capabilities like collaborative marksheet editing and live data sharing, educators and administrators can work together more effectively and efficiently.

*Integration with Student Information Systems (SIS):* Systems (SIS): Integrating the system with SIS platforms commonly used in educational institutions would ensure interoperability and data consistency, allowing for seamless data exchange between administrative and academic systems.

## IX. CONCLUSION

An effective way to automate the extraction, storing, and analysis of marksheet data at educational institutions is provided by the "Marksheet Analysis using OCR" project. Using OCR technology and data analysis tools, the system streamlines administrative chores that take a lot of time and offers insightful information for making decisions based on data. Ensuring data correctness, security, and user-friendliness, this project answers the changing needs of educational institutions with strong security measures, scalability, and future enhancement possibilities.

## X. ACKNOWLEDGMENT

We would like to thank Prof Pranjali Bahalkar for their support and guidance in completing our project, the topic Marksheet Analysis using OCR. Your expertise and guidance have been instrumental in its success. A heartfelt gratitude for your invaluable assistance and support throughout the project.

## REFERENCES

- [1] Dr. Vishwanath Burkpalli, Abhishek Joshi, Abhishek B Warad, Akash Patil " Automatic Number Plate Recognition Using TensorFlow and EasyOCR " Volume:04/Issue:09/September-2022
- [2] Jason Ravagli, Zahra Ziran, Simone Marinai. " Text Recognition and Classification in Floor plan Images " Jason Ravagli, Zahra Ziran, Simone Marinai.
- [3] Lloyd Alan Fletcher and Rangachar Kasturi " A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images " IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. IO. NO. 6, NOVEMBER 1988
- [4] Adithya T. G, Pavithra G, Praveen N. T. C. Manjunath " Automatic Number Plate Recognition Idea Development using AI-based ANNs " Vol. 8 No. 1 (2022) Journal of Communication Engineering and its Innovations
- [5] Matsakis, S. N., & Cham, W. K. "OCRopus: A community-driven open-source OCR system" Document Recognition and Retrieval XV, DRR 2008, 15th Document Recognition and Retrieval Conference, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 29-31, 2008. Proceedings
- [6] Nicholas Arnold, Plamen P Angelov," Automatic Extraction and Labelling of Memorial Objects From 3D Point Clouds" April 2021 Journal of Computer Applications in Archaeology
- [7] "Data-driven Decision Making in Higher Education Institutions: State-of-play" Silvia Gaftandzhieva, Sadiq Hussain, Slavoljub Hilcenko, Rositsa Doneva January 2023

- [8] An overview on web scraping techniques and tools AV Saurkar, KG Pathare, SA Gode - International Journal on Future ..., 2018 - ijfrcsce.org

