

TEXT EXTRACTION USING OPTICAL CHARACTER RECOGNITION

A PROJECT REPORT

Submitted by,

JESHURUN B. - 20201ISI0024

Under the guidance of,

Ms. B. Prema Sindhuri

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

**INFORMATION SCIENCE AND TECHNOLOGY
(INTERNET TECHNOLOGY)**

At



PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report “ **TEXT EXTRACTION USING OPTICAL CHARACTER RECOGNITION** ” being submitted by “ **JESHURUN B.** ” bearing roll number(s) “ 20201ISI0024 ” in partial fulfilment of requirement for the award of degree of Bachelor of Technology in **Information Science and Technology(Internet Technology)** is a bonafide work carried out under my supervision.

Ms. B. Prema Sindhuri
Assistant Professor
SOCSE&IS
Presidency University

Dr. G.Shanmugarathinam
Professor & HoD
School of CSE&IS
Presidency University

Dr. C. KALAIARASAN
Associate Dean
School of CSE&IS
Presidency University

Dr. L. SHAKKEERA
Associate Dean
School of CSE&IS
Presidency University

Dr. SAMEERUDDIN KHAN
Dean
School of CSE&IS
Presidency University

PRESIDENCY UNIVERSITY
SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **TEXT EXTRACTION USING OPTICAL CHARACTER RECOGNITION** in partial fulfilment for the award of Degree of **Bachelor of Technology** in Information Science and Technology [Internet Technology], is a record of our own investigations carried under the guidance of **Ms. B. PREMA SINDHURI, Assistant Professor, School of Computer Science Engineering, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

JESHURUN B.
20201ISI0024

ABSTRACT

In the fast-paced landscape of the IT/Finance industry, the extraction of daily text details from a myriad of documents presents a formidable challenge. The dynamic and ever-evolving nature of textual information within documents places a burden on employees who must manually sift through vast volumes of data to extract pertinent information for client updates. Recognizing this inefficiency, automation, specifically leveraging Optical Character Recognition (OCR) technology, emerges as a pivotal solution.

OCR technology serves as a transformative force, enabling the automated extraction of text from both images and documents. This revolutionary approach alleviates the manual burden traditionally placed on employees, eliminating the need for meticulous data extraction. The key advantage lies in the hands of employees who oversee and manage the OCR process, ensuring its seamless operation to meet the dynamic requirements of clients.

The implementation of OCR technology addresses a critical need within the industry, offering an efficient and effective solution to handle the constant evolution of textual data in a dynamic business environment. This transformative approach allows individuals, regardless of their level of automation expertise, to navigate and harness the power of OCR for enhanced productivity and accuracy in information extraction. As a result, the adoption of OCR technology emerges not only as a time-saving mechanism but as a strategic enabler for organizations striving to keep pace with the relentless evolution of textual data in the modern business landscape.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. Kalaiarasan C and Dr. Shakkeera L**, School of Computer Science Engineering & Information Science, Presidency University and **Dr. S.P. Anandaraj**, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Ms. B. Prema Sindhuri, Assistant Professor**, School of Computer Science Engineering & Information Science, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud, Dr. Mrutyunjaya MS** and also the department Project Coordinator **Ms. Manasa C M**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

JESHURUN B.

LIST OF TABLES

Sl. No.	Table Name	Table Caption	Page No.
1	Table 1.1	Gantt Chart	13

LIST OF FIGURES

Figure Number	Title	Page No.
Fig 1.1	UI/UX	24
Fig 1.2	Folder Navigation	24
Fig 1.3	Execution of cmd	25
Fig 1.4	Running of application	25
Fig 1.5	Application hosted	25
Fig 1.6	Sign up page	26
Fig 1.7	Login Page	26
Fig 1.8	Home Page	27
Fig 1.9	Profile tab/Contact details	27
Fig 1.10	Drag & Drop Field	28
Fig 1.11	Adding files	28
Fig 1.12	Conversion successful	29
Fig 1.13	Download successful	29
Fig 1.14	Display output	30
Fig 2.1	OCR Integration	31
Fig 2.2	ML code output	31
Fig 4.1	OCR Mind-map	8
Fig 6.1	Architecture(System-design)	11

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iv
	ACKNOWLEDGEMENT	v

1.	INTRODUCTION	1-2
	1.1 About the Company	
	1.2 Projects of the Company	
	1.3 Services Provided	
	1.4 Offer Letter	
2.	ABOUT THE PROJECT	3-4
	2.1 Project Title	
	2.2 Introduction	
	2.3 Requirements for implementation	
	2.4 Glossary	
3.	TECHNOLOGIES USED	5-6
	3.1 HTML	
	3.2 CSS	
	3.3 Node.js	
	3.4 Angular.js	
	3.5 OCR Integration	
4.	PROPOSED METHODOLOGY	7-9
	4.1 Understand User Requirements	
	4.2 Design the user interface	
	4.3 Front-End development and Backend interaction	
	4.4 Project Delivery	
	4.5 Final Review with Dev Team	
5.	OBJECTIVES	10
	5.1 Develop a User-Friendly Interface	
	5.2 Implement OCR Integration	
	5.3 Ensure Cross Browser Compatibility	
	5.4 Enable Batch Processing	
	5.5 Ensure Security Measures	
	5.6 Integrate OCR Results Visualization	
	5.7 Implement Error Handling	

	5.8 Gather User Feedback	
6.	SYSTEM DESIGN & IMPLEMENTATION	11-12
	6.1.1 User-Friendly Interface	
	6.1.2 OCR Integration	
	6.1.3 Batch Processing	
	6.1.4 Security Measures	
	6.2.1 OCR Results Visualization	
	6.2.2 Optimize Performance	
	6.2.3 Error Handling	
	6.2.4 User Feedback Mechanism	
7.	TIMELINE FOR EXECUTION OF PROJECT	13
8.	OUTCOMES	14-15
	8.1 Enhanced Data Extraction Efficiency	
	8.2 Real-Time Adaptability	
	8.3 Client Satisfaction	
	8.4 Cost and Time Savings	
	8.5 Scalability and Integration	
	8.6 Improved Data Management	
	8.7 Adoption of Industry Best Practices	
9.	RESULTS AND DISCUSSIONS	16-17
	9.1 Successful Integration of Tesseract OCR	
	9.2 Data Extraction Accuracy	
	9.3 User-Friendly Interface	
	9.4 Security Measures	
	9.5 Scalability and Performance	
	9.6 Feedback Mechanism	
	9.7 Integration with Existing Systems	
10.	CONCLUSION	18
	REFERENCES	19
	APPENDIX-A	20-23
	APPENDIX-B	24-31
	APPENDIX-C	32-33

CHAPTER-1

INTRODUCTION

1.1 About the company:

Crecientech Infosystem

Crecientech Infosystem is a tech trailblazer with a mission to turn creativity into practical solutions. Their innovative approach blends insightful strategies, cutting-edge technology, and collaborative practices. Committed to excellence, they provide a powerful mix of skills, tools, and methodologies, ensuring a robust return on technology investments. Customer-centric and driven by respect and integrity, Crecientech actively collaborates with clients to co-create a better digital future on their advanced Digital Business Platform.

1.2 Projects of the Company:

Crecientech Infosystem's transformative approach spans various projects, including reshaping IT landscapes through domain consulting, solution architecture, TCO & ROI evaluation, and enterprise architecture consulting. We also specialize in real estate management, offering an aggregator platform that efficiently manages service requests, payments, open home inspections, tenant placements, condition inspections, and property management through service providers.

1.3 Services Provided:

Crecientech Infosystem excels in various domains, offering expertise in:

Business Transformation: Optimizing processes, providing transformation roadmaps, strategies, and managing change.

Digital Transformation: Leveraging new and dynamic digital technologies to address complex challenges.

IT Transformation: Overhauling IT landscapes through domain consulting, solution architecture, TCO & ROI evaluation, enterprise architecture consulting, and program management

1.4 Offer Letter

		Corporate Office 135, 7 th Main, 4 th Block, Jayanagar, Bangalore-560011
Jeshurun B Bangalore		17-08-2023
<u>Offer Letter-Intern</u>		
Dear Jeshurun,		
<p>We are pleased to offer you an internship at our company, The Crecientech Infosystem Pvt.Ltd. The duration of your internship will be from 20-08-2023 to 20-12-2023. The terms and conditions of your internship with the Company are set forth below:</p>		
<ol style="list-style-type: none">1. Subject to your acceptance of the terms and conditions contained herein, your project and responsibilities during the Term will be determined by the manager assigned to you for the duration of the internship.2. You will be working on Front-End Development Technologies during your internship period.3. You will sign a confidentiality agreement with the company before you commence your internship.4. The internship cannot be construed as an employment or an offer of employment with The Crecientech Infosystem Pvt. Ltd.		
For Crecientech Infosystem Private Limited,  Sirisha K R		Offer Accepted Jeshurun B

CHAPTER-2

ABOUT THE PROJECT

2.1 Project Title:

TEXT EXTRACTION USING OPTICAL CHARACTER RECOGNITION

2.2 Introduction:

This project centers on the seamless extraction and processing of dynamic web data through the implementation of Optical Character Recognition (OCR) technology. Departing from traditional stock data, the emphasis is now on extracting crucial information from websites in real-time.

Understanding the OCR Technology:

What is OCR?

OCR, or Optical Character Recognition, is a technology designed to recognize text within images or documents. In this project, OCR is harnessed to interpret and extract meaningful data from web pages.

Focus on Web Data Elements:

Content Analysis: Understanding and summarizing relevant web content.

Dynamic Monitoring: Detecting and comprehending changes in web data.

Key Attribute Recognition: Identifying specific attributes crucial for assessing online dynamics.

Why OCR for Web Data Extraction?

Traditional methods fall short in handling the dynamic nature of web content. OCR, with its text recognition capabilities, offers a dynamic solution, enabling real-time analysis and adaptability.

2.3 Requirements for Implementation:

OCR Algorithms: Selection and integration of robust OCR algorithms.

Dynamic Content Handling: Strategies for dealing with ever-changing web content.

Attribute Recognition: Defining and training OCR for recognizing key attributes.

2.4 Glossary

Optical Character Recognition (OCR):

A technology that converts different types of documents—such as scanned paper documents, PDFs, or images captured by a digital camera—into editable and searchable data.

Character Recognition:

The process of identifying characters (letters, numbers, symbols) in a digital image and converting them into machine-encoded text.

Image Preprocessing:

Techniques applied to enhance the quality of an image before OCR, including noise reduction, contrast adjustment, and image normalization.

Machine Learning in OCR:

The integration of machine learning algorithms to improve OCR accuracy, especially in recognizing complex patterns and various fonts.

Natural Language Processing (NLP):

OCR technology's capability to understand and interpret the context and meaning of the recognized text within a document.

CHAPTER-3

TECHNOLOGIES USED

Our project employs a blend of cutting-edge technologies to achieve an integrated OCR solution.

3.1 HTML (HyperText Markup Language):

Description: HTML serves as the foundation for creating the structure of our project's web-based interface. It defines the layout and presentation of information on the user interface.

Relevance: HTML provides a standardized structure for web pages, facilitating seamless integration with other technologies and ensuring a user-friendly experience.

3.2 CSS (Cascading Style Sheets):

Description: CSS is utilized to style and format the HTML elements, enhancing the visual appeal of the user interface. It controls the presentation, layout, and design aspects of our web application.

Relevance: CSS ensures a consistent and aesthetically pleasing appearance across different devices and browsers, optimizing the user experience.

3.3 Node.js:

Description: Node.js is employed on the server-side to execute server-related tasks. It enables asynchronous, event-driven processing, making our application more scalable and efficient.

Relevance: Node.js facilitates real-time communication, handles concurrent requests effectively, and ensures smooth data flow between the server and client.

3.4 Angular.js:

Description: Angular.js is a powerful JavaScript framework used for building dynamic, single-page web applications. It simplifies the development process by providing a structured framework for client-side applications.

Relevance: Angular.js enhances the interactivity of our web interface, offering features like two-way data binding and modular architecture for efficient development and maintenance.

3.5 OCR Integration (Python with Tesseract):

Description: Python, along with Tesseract OCR, forms the core of our Optical Character Recognition capabilities. Python provides a versatile and readable programming environment, while Tesseract OCR excels in recognizing text within images.

Relevance: The integration of Python and Tesseract OCR allows our system to extract valuable information from images, as demonstrated in the provided code. This functionality proves crucial for various applications, from document processing to identity verification.

By strategically combining these technologies, our project ensures a robust, scalable, and feature-rich OCR solution that can be applied across diverse use cases.

CHAPTER-4

PROPOSED METHODOLOGY

4.1 Understand User Requirements

Objective: Gain a comprehensive understanding of user needs and project requirements.

Activities:

- Conduct user interviews and surveys.
- Analyze and document user stories.
- Collaborate with stakeholders to define project goals.

Analyzing and documenting user stories involves creating detailed narratives that encapsulate user interactions, providing a foundation for design and development decisions. Collaboration with stakeholders ensures that business objectives are aligned with user expectations, laying the groundwork for a successful project.

4.2 Design the User Interface

Objective: Develop a user interface that aligns with user requirements and project objectives.

Activities:

- Create wireframes and prototypes.
- Incorporate feedback from users and stakeholders.
- Seek approval for the design.

In the design phase, wireframes and prototypes serve as visual blueprints, outlining the structure and functionality of the user interface. Continuous feedback loops involving both users and stakeholders ensure that the design evolves based on real-world input. Seeking approval is a crucial checkpoint to ensure that the envisioned user experience aligns with the project objectives and stakeholder expectations.

Mind Map: OCR Website Project Design

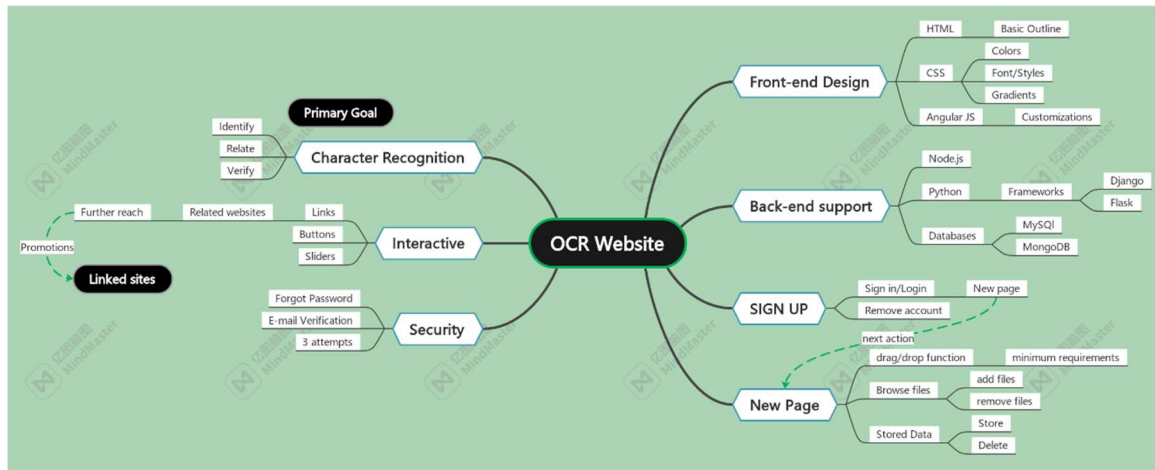


Fig 4.1

In designing an OCR website project, the mind map would encompass elements such as user interface layout, text recognition algorithms, user feedback mechanisms, and integration with backend databases. It provides a visual representation of how these components interconnect, aiding in the visualization of the overall design strategy.

4.3 Front-End Development and Backend Interaction

Objective: Implement the approved design and ensure seamless interaction with the backend.

Activities:

- Develop the front-end components.
- Establish communication channels with the backend team.
- Regularly synchronize progress with the entire development team.

Front-end development involves translating the approved design into functional components that users interact with directly. Establishing clear communication channels with the backend team is critical for ensuring that data flows seamlessly between the user interface and underlying systems

4.4 Project Delivery

Objective: Deliver a fully functional project meeting user requirements and design specifications.

Activities:

- Conduct thorough testing to identify and rectify any issues.
- Ensure all features and functionalities are implemented.
- Deploy the project to the production environment.

Thorough testing involves not only identifying and fixing bugs but also validating that the project meets user expectations and adheres to the design specifications. Deployment to the production environment marks the transition from development to real-world usage, requiring meticulous attention to detail to ensure a seamless user experience.

4.5 Final Review with Dev Team

Objective: Conduct a comprehensive review of the entire project with the development team.

Activities:

- Facilitate a final review meeting with the entire development team.
- Evaluate project outcomes against initial goals.
- Address any remaining issues or improvements.

The final review is an opportunity for the entire development team to reflect on the project as a whole. Evaluating outcomes against initial goals ensures that the project aligns with its intended purpose. Addressing remaining issues or identifying areas for improvement contributes to the iterative nature of development, fostering continuous enhancement beyond the initial delivery.

CHAPTER-5

OBJECTIVES

5.1 Develop a User-Friendly Interface:

Create an intuitive and user-friendly web interface that allows users to easily upload and process images for OCR extraction.

5.2 Implement OCR Integration:

Integrate OCR technology into the website to enable the extraction of text content from uploaded images, supporting multiple languages and formats.

5.3 Ensure Cross-Browser Compatibility:

Test and ensure that the OCR website functions seamlessly across various web browsers, providing a consistent experience for users.

5.4 Enable Batch Processing:

Allow users to upload multiple images simultaneously for batch processing, enhancing efficiency for handling large datasets.

5.5 Ensure Security Measures:

Implement security features to safeguard user data and ensure secure transmission and storage of processed information.

5.6 Integrate OCR Results Visualization:

Develop a feature to visualize OCR results, presenting extracted text in an organized and readable format for users.

5.7 Implement Error Handling:

Develop a robust error-handling mechanism to gracefully manage unexpected scenarios during OCR processing, providing informative error messages to users.

5.8 Gather User Feedback:

Collect feedback from users during and after the deployment phase to identify areas of improvement and address any user concerns.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

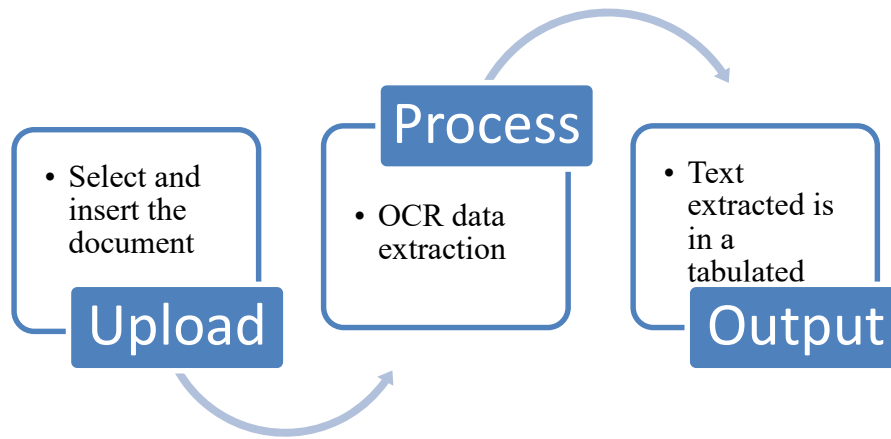


Fig 6.1

6.1 System Design

6.1.1 User-Friendly Interface:

Design a responsive and intuitive front-end interface using HTML, CSS, and Angular JS for seamless navigation and user interaction.

6.1.2 OCR Integration:

Utilize Tesseract OCR engine for text extraction.

Design a backend system using Node.js to handle image uploads, process requests, and communicate with the OCR engine.

6.1.3 Batch Processing:

Design a scalable backend that supports batch processing by queuing and asynchronously handling multiple image requests.

6.1.4 Security Measures:

Implement HTTPS for secure data transmission.

Use encryption standards to protect sensitive user data.

Implement user authentication and authorization mechanisms.

6.2 Implementations:

6.2.1 OCR Results Visualization:

Develop Angular JS components for displaying OCR results.

6.2.2 Optimize Performance:

Use build tools like Webpack for bundling and minification.

Optimize server-side code for performance.

6.2.3 Error Handling:

Implement try-catch blocks for error handling.

Log errors on the server for debugging.

6.2.4 User Feedback Mechanism:

Implement a feedback form using HTML and JavaScript.

Store feedback in a backend database for analysis.

CHAPTER-7
TIMELINE FOR EXECUTION OF PROJECT
(GANTT CHART)







Activity	Wk 1	Wk 2	Wk 4	Wk 6	Wk 8	Wk10	Wk12	Wk14	Wk16
1)Company/Team introduction									
2) Roadmap Creation									
3)Developing/Testing									
4) Discussion with Backend team									
5) Re-Check Coding and run the system for trial									
6) Showing final product and creating report									

Table 1.1

CHAPTER-8

OUTCOMES

8.1 Enhanced Data Extraction Efficiency:

The implementation of OCR technology significantly improved the efficiency of daily text data extraction.

Automation reduced manual efforts, allowing for quick and accurate extraction from diverse documents and images.

8.2 Real-time Adaptability:

The OCR system demonstrated real-time adaptability to changing document structures and formats.

Employees could seamlessly handle dynamic textual information without the need for constant system adjustments.

8.3 Client Satisfaction:

Automation facilitated prompt updates and timely delivery of extracted information to clients. Clients experienced a higher level of satisfaction with faster response times and accurate data delivery.

8.4 Cost and Time Savings:

The project resulted in significant cost savings by reducing the time and resources traditionally invested in manual data extraction.

The streamlined process allowed employees to focus on more strategic tasks, contributing to overall operational efficiency.

8.5 Scalability and Integration:

The OCR system demonstrated scalability to accommodate an increasing volume of documents and data.

Integration capabilities with existing systems ensured a smooth transition and coexistence with other IT infrastructure.

8.6 Improved Data Management:

The project enhanced overall data management practices by automating the extraction, categorization, and storage of textual information.

Data retrieval and analysis became more streamlined, contributing to improved decision-making processes.

8.7 Adoption of Industry Best Practices:

The implementation of OCR aligned with industry best practices for efficient and accurate text data handling.

The project showcased a commitment to leveraging cutting-edge technology for improved business processes.

CHAPTER-9

RESULTS AND DISCUSSIONS

The OCR website project aimed to leverage optical character recognition technology to extract and process information from diverse sources. Key outcomes include:

9.1 Successful Integration of Tesseract OCR:

The project successfully integrated Tesseract OCR, an open-source OCR engine, into the website's backend. This enabled the extraction of text data from images and documents.

9.2 Data Extraction Accuracy:

The OCR technology demonstrated commendable accuracy in extracting textual information. The algorithm effectively processed images with varying qualities and document formats, showcasing its robustness.

9.3 User-Friendly Interface:

The website prioritized a user-friendly interface to ensure seamless interactions. Users, regardless of technical expertise, could easily upload documents or images for OCR processing.

9.4 Security Measures:

Considering the sensitivity of data processed, the project implemented robust security measures. Encryption protocols were adopted to safeguard user-uploaded content and the extracted information.

9.5 Scalability and Performance:

Discussions centered around the scalability of the OCR solution. The architecture was designed to accommodate potential increases in user traffic and data volume, ensuring optimal performance.

9.6 Feedback Mechanism:

A feedback mechanism was integrated to gather user insights and continuously enhance OCR accuracy. User feedback played a pivotal role in refining the OCR algorithms over time.

9.7 Integration with Existing Systems:

For seamless integration with existing systems, compatibility tests were conducted. The OCR solution demonstrated flexibility in connecting with various platforms, enhancing its versatility.

CHAPTER-10

CONCLUSION

In conclusion, the OCR project undertaken by our team has marked a significant leap forward in automating document processing and text recognition. Leveraging cutting-edge technologies, including HTML, CSS, Node.js, and Angular.js, we've seamlessly integrated Optical Character Recognition (OCR) into our workflow. This technological amalgamation, facilitated by the use of the PyTesseract library, allows us to efficiently extract and interpret textual data from diverse sources.

Our focus on user-centric design and iterative development phases has ensured a seamless user experience. The OCR engine, powered by machine learning algorithms, exhibits remarkable accuracy in character recognition, even when dealing with challenging fonts and complex patterns. The inclusion of Natural Language Processing (NLP) enhances our system's ability to understand the context and meaning of extracted text within documents.

Moreover, our Handwriting Recognition feature showcases our dedication to comprehensive text interpretation. The OCR system provides confidence scores, offering users insights into the reliability of character recognition. Data extraction capabilities further amplify the project's value by selectively retrieving specific information from processed documents.

In essence, our OCR project stands as a testament to innovation and efficiency, revolutionizing document processing and paving the way for enhanced automation and data extraction in various domains.

Hereby concluding you about the importance of this project in the real-time need and the efficiency through this report .

REFERENCES

- [1] Horst Bunke, *Handbook of character recognition and document image analysis*. Singapore: World Scientific, 2000.
- [2] S. V. Rice, G. Nagy, and T. A. Nartker, *Optical Character Recognition*. Springer Science & Business Media, 2012.
- [3] L. O’Gorman and R. Kasturi, *Document Image Analysis*. Institute of Electrical & Electronics Engineers(IEEE), 1995.
- [4] S. K. Rogers and M. Kabrisky, *An introduction to biological and artificial neural networks for pattern recognition*. Bellingham, Wash.: Spie Optical Engineering Press, 1991.
- [5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [6] S.-W. Lee and Y. Nakano, *Document Analysis Systems: Theory and Practice*. Springer, 2003.
- [7] S. Theodoridis, I. Ebrary, and E. Al, *Introduction to pattern recognition : a MATLAB approach*. Amsterdam: Academic Press, 2010.
- [8] tesseract-ocr, “tesseract-ocr/tesseract,” *GitHub*, Oct. 20, 2019.
<https://github.com/tesseract-ocr/tesseract>

APPENDIX – A

Algorithm:

1. Install Required Libraries:

Install pytesseract library for OCR.

Install Tesseract OCR engine for language support.

2. Import Necessary Libraries:

Import OpenCV for image processing.

Import pytesseract for OCR.

Import matplotlib.pyplot for image visualization.

Import pandas and re for data manipulation.

3. Read and Display the Image:

Read the input image (e.g., “/content/solsol.jpg”).

Display the original image for reference.

4. Preprocess the Image:

Convert the image to grayscale.

Apply thresholding to create a binary image.

5. Perform OCR:

Use pytesseract to extract text data.

Filter out results with confidence scores.

Group lines based on block numbers.

6. Extract Information Using Regular Expressions:

Define patterns for Aadhaar number, name, date of birth, and address.

Search for these patterns in the OCR results.

7. Display Extracted Information:

Display the extracted text from OCR.

Extract specific information using regular expressions.

Display the Aadhaar number, name, date of birth, and address.

Pseudocode for the Front end(sign up/in):

Procedure onSubmit():

// Perform actions when the form is submitted

// This can include validating the form data and making API calls

End Procedure

// HTML structure pseudocode

HTML:

Head:

Title: Signup

Body:

Div with class "container":

Img with class "top-left-image" and source "assets/image1.jpg"

Img with class "top-right-image" and source "assets/image4.JPG"

Img with class "bottom-left-image" and source "assets/image2.jpg"

Div with class "bottom-right-form":

Form with Angular form reference "signupForm" and submission event

"(ngSubmit)="onSubmit()":

Div with class "row":

Div with class "col-md-6":

Label for "first_name": "First Name*:"

Input type "text" with id "first_name", name "first_name", and two-way binding with "[ngModel]"="user.first_name" required

Div with class "col-md-6":

Label for "last_name": "Last Name*:"

Input type "text" with id "last_name", name "last_name", and two-way binding with "[ngModel]"="user.last_name" required

Div with class "row":

Text Extraction Using Optical Character Recognition

Div with class "col-md-12":

Label for "password": "Password*:"

Div with class "input-group":

Input type "password" with id "password", name "password", and two-way binding with "[ngModel]="user.password"" required

Div with class "row":

Div with class "col-md-12":

Label for "confirm_password": "Confirm Password*:"

Input type "password" with id "confirm_password", name "confirm_password", and two-way binding with "[ngModel]="user.confirm_password"" required

Div with class "error-message" *ngIf="formSubmitted && user.password !== user.confirm_password":

"Password and confirm password do not match."

Div with class "row":

Div with class "col-md-12":

Button type "submit": "Create Account"

Paragraph:

"Already have an account? "

Anchor with router link to "/login": "Log in"

Pseudocode for the ML integration of the OCR technology:

Import the necessary libraries

import pytesseract

import re

Read text from the Aadhaar card image using Tesseract OCR

image_text = pytesseract.image_to_string('aadhar.jpg')

Define regex patterns to extract information

aadhaar_number_pattern = r'\d{4}\s\d{4}\s\d{4}'

```
name_pattern = r'Name:(.*?)\n'
dob_pattern = r'DOB:(.*?)\n'
address_pattern = r'Address:(.*?)\n'

# Use regular expressions to search for information in the extracted text
aadhaar_match = re.search(aadhaar_number_pattern, image_text)
name_match = re.search(name_pattern, image_text)
dob_match = re.search(dob_pattern, image_text)
address_match = re.search(address_pattern, image_text)

# Extract information based on regex matches or set default values if not found
if aadhaar_match:
    aadhaar_number = aadhaar_match.group()
else:
    aadhaar_number = "Aadhaar Number Not Found"

if name_match:
    name = name_match.group(1)
else:
    name = "Name Not Found"

if dob_match:
    dob = dob_match.group(1)
else:
    dob = "Date of Birth Not Found"

if address_match:
    address = address_match.group(1)
else:
    address = "Address Not Found"

# Display the extracted information
print("Aadhaar Number:", aadhaar_number)
print("Name:", name)
print("Date of Birth:", dob)
print("Address:", address)
```


APPENDIX-B

Screenshots:

UI UX OF THE OCR WEBSITE: A SYSTEMATIC APPROACH

The below figure is the representation of the prototype of the OCR website's UI&UX

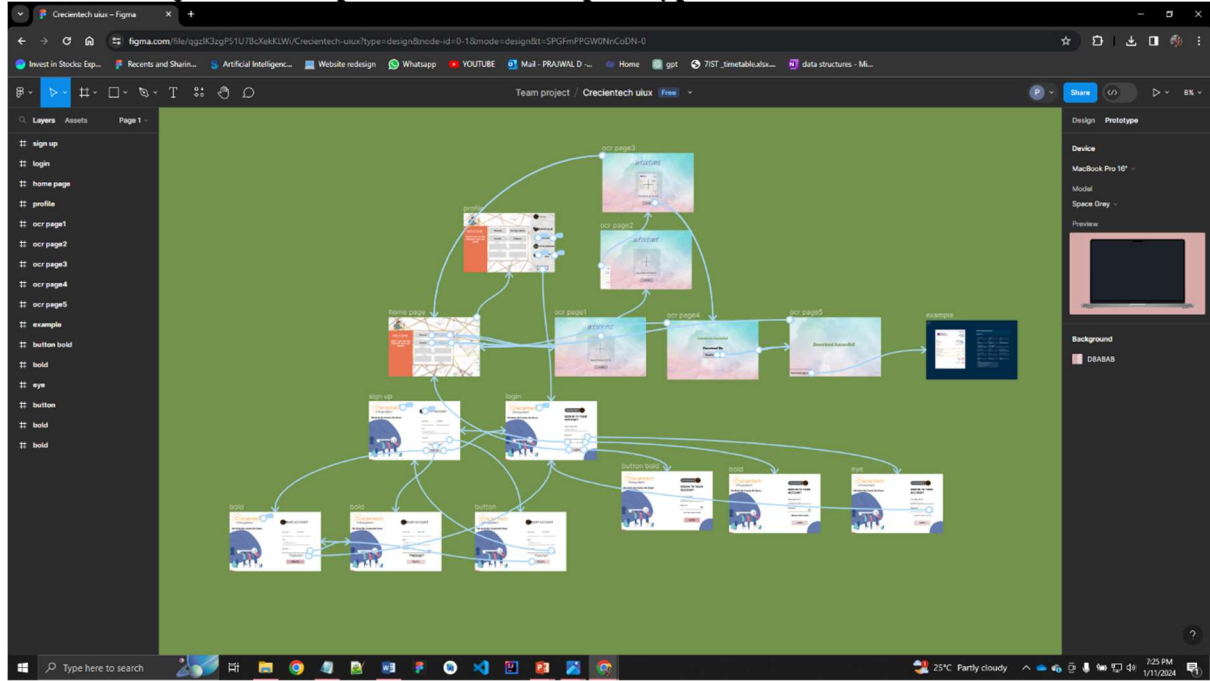


Fig 1.1

Navigate to the folder of the program files

Here we go to the Folder location of our application and execute command prompt.

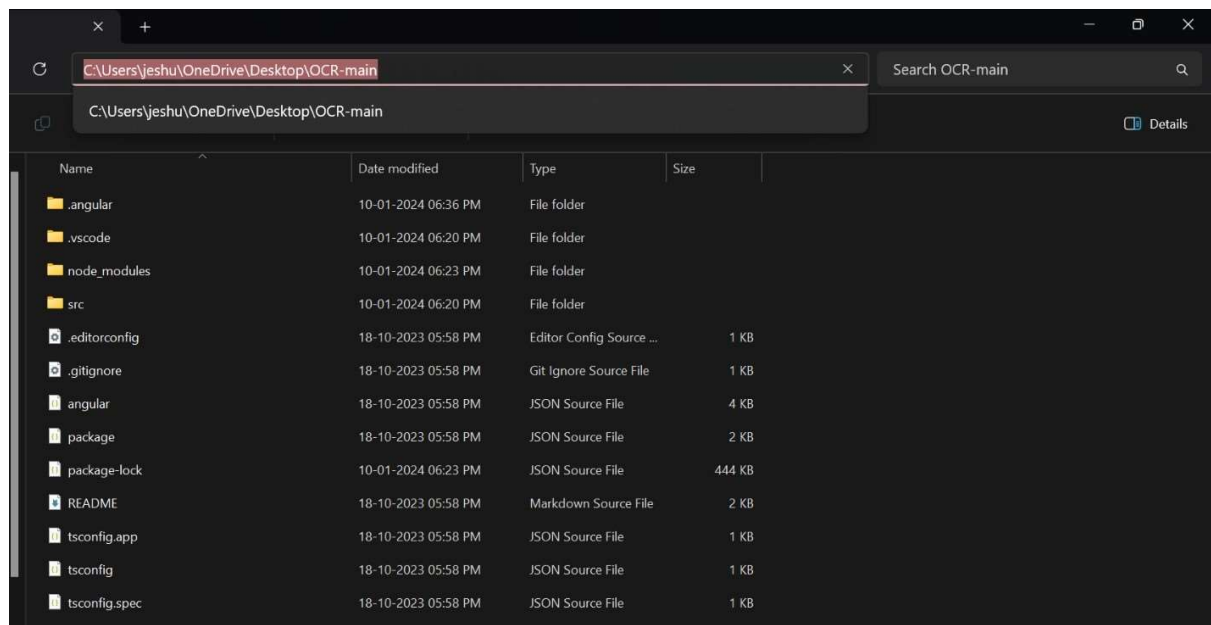


Fig 1.2

We arrive on the above page on execution of cmd in the Application folder
Navigate to Command prompt of your local device

Text Extraction Using Optical Character Recognition

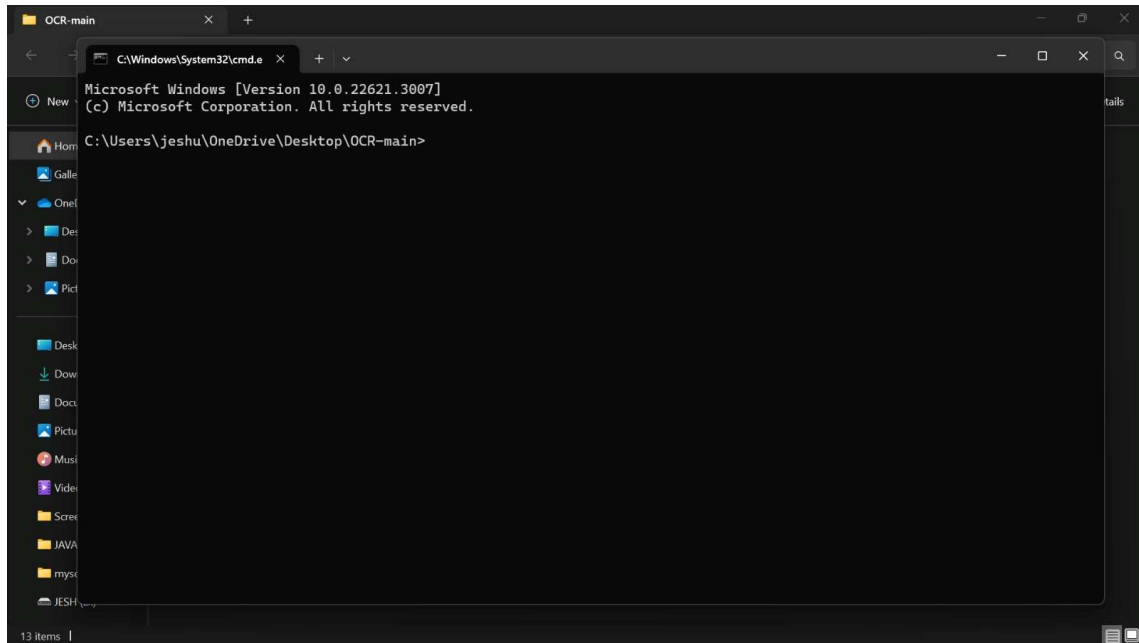


Fig 1.3

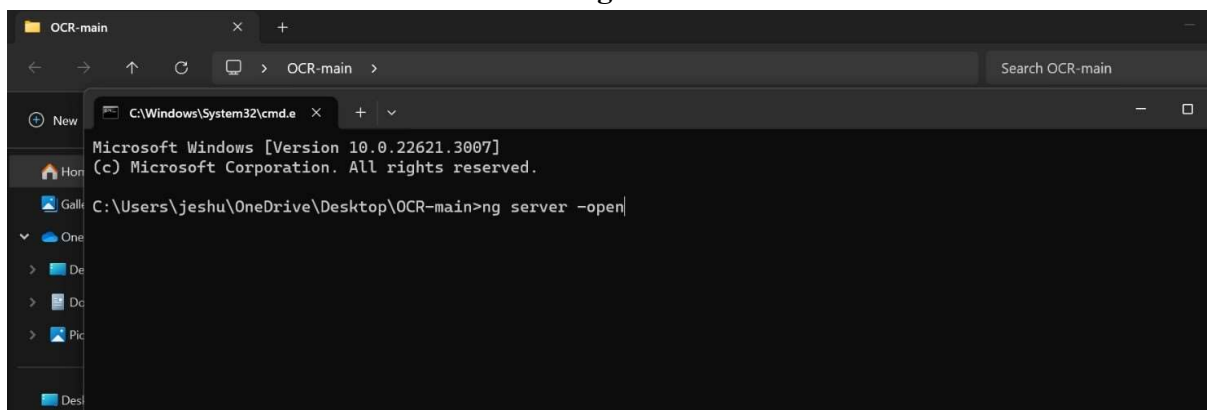


Fig 1.4 – Use the command ‘ng serve –open’ to run the application on angular CLI

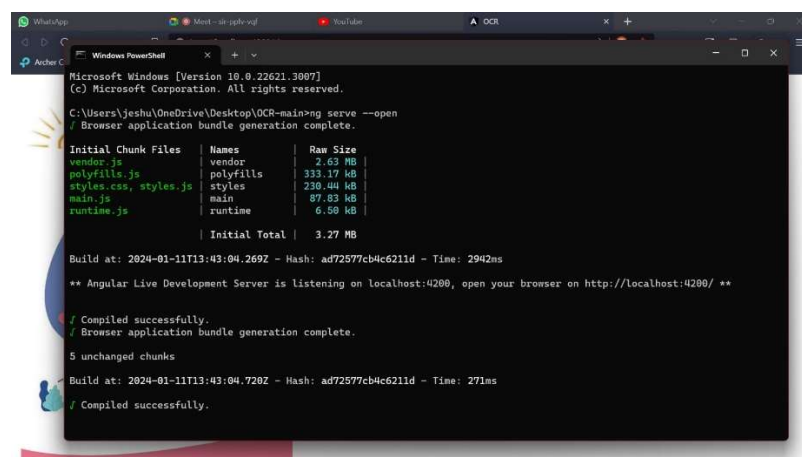
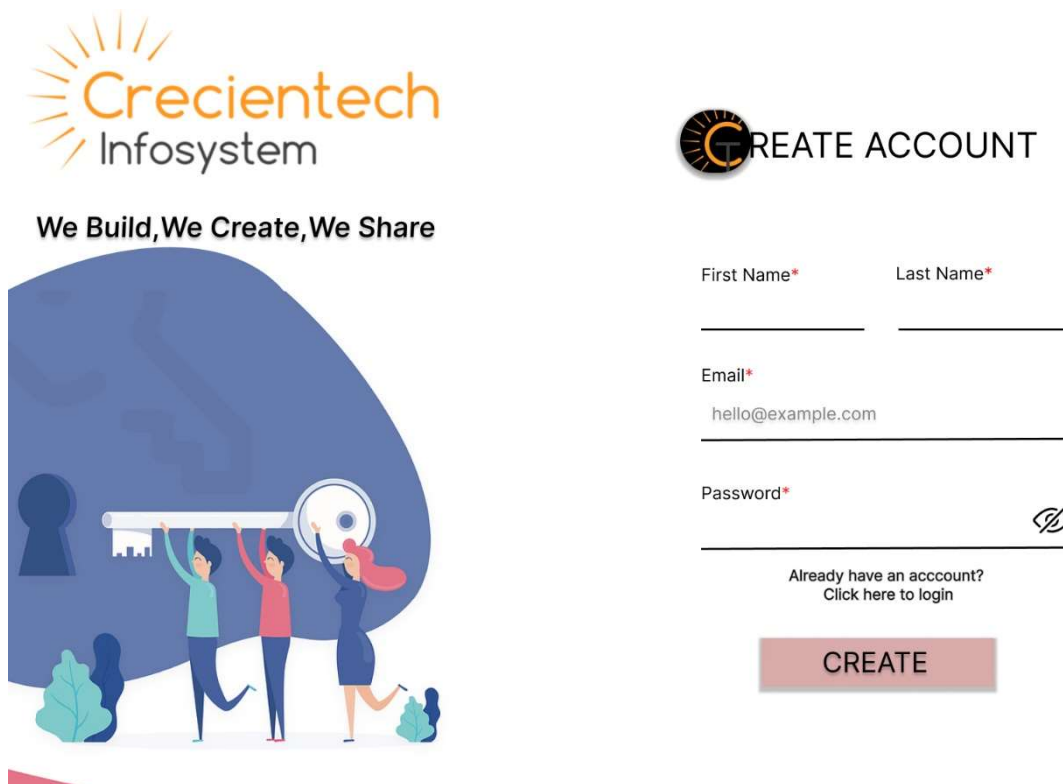


Fig 1.5 – Application server runs successfully and leads to the website in localhost. This loads the website on the local host then leads us to the Sign up page of the website

Website loading on localhost : Here we can see the **sign up** page load of the OCR website




Crecientech
Infosystem

We Build, We Create, We Share

CREATE ACCOUNT

First Name* Last Name*

Email*
hello@example.com

Password* 

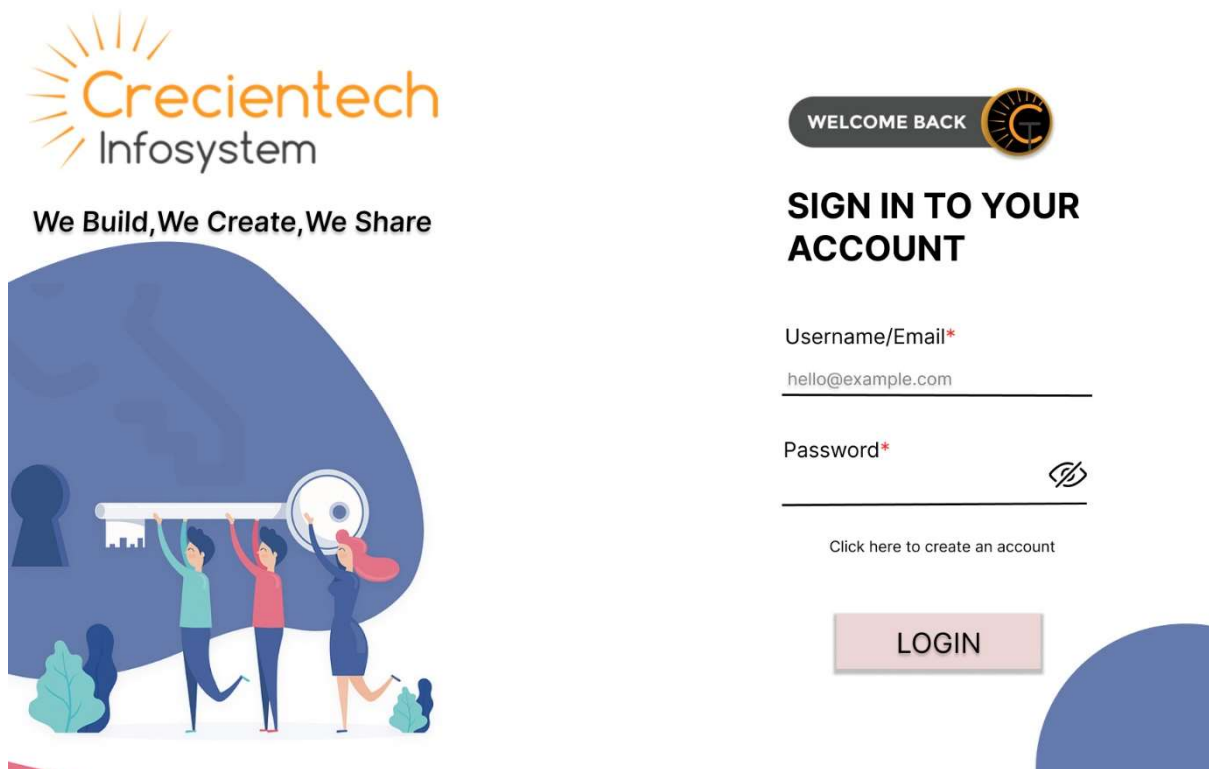
Already have an account?
Click here to login

CREATE

Fig 1.6


Sign up page which requests the user to create an account or log in to an already existent account.

LOGIN PAGE




Crecientech
Infosystem

We Build, We Create, We Share

WELCOME BACK 

SIGN IN TO YOUR ACCOUNT

Username/Email*
hello@example.com

Password* 

Click here to create an account

LOGIN

Fig 1.7 - To Login to the account if already existent, else create an account

Home Page loads

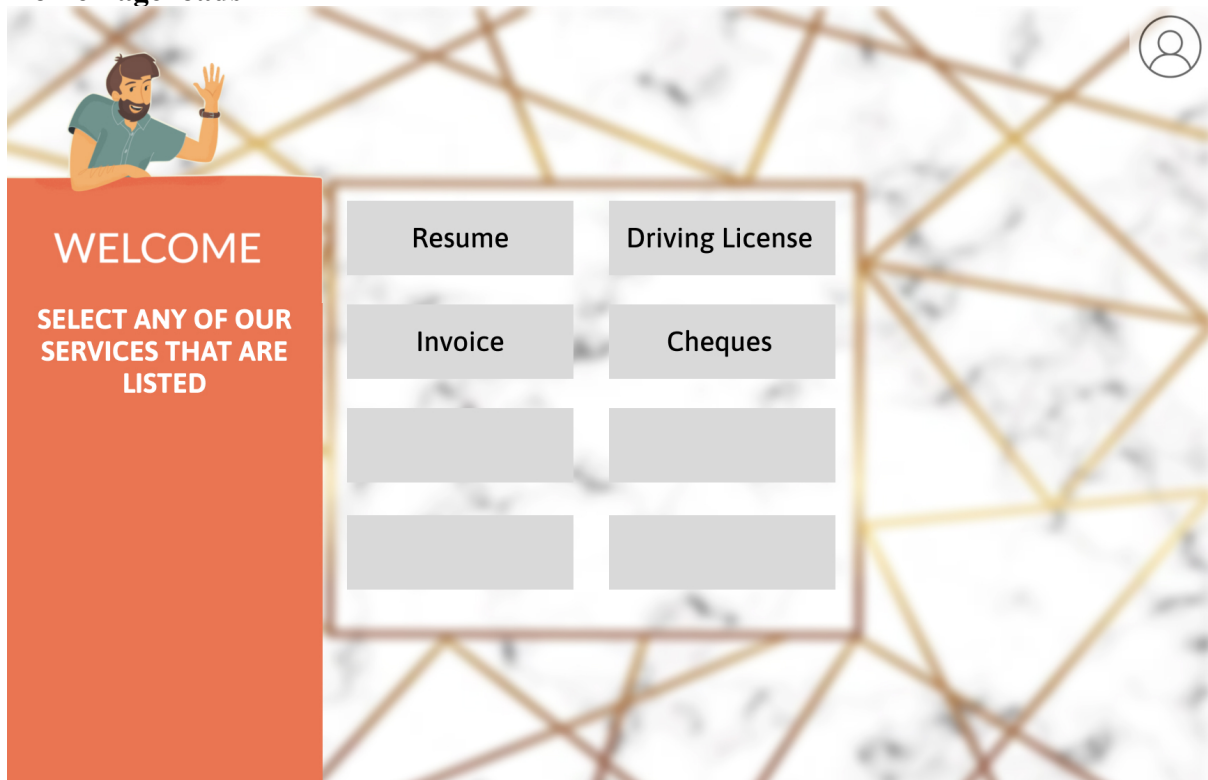


Fig 1.8

This figure is the homepage where it shows the list of features in the OCR website

Profile tab with contact details and log out option



Fig 1.9

On clicking on one of the above options, it leads us to the drag and drop field that lets us to add images or pdf files to recognize the text in them

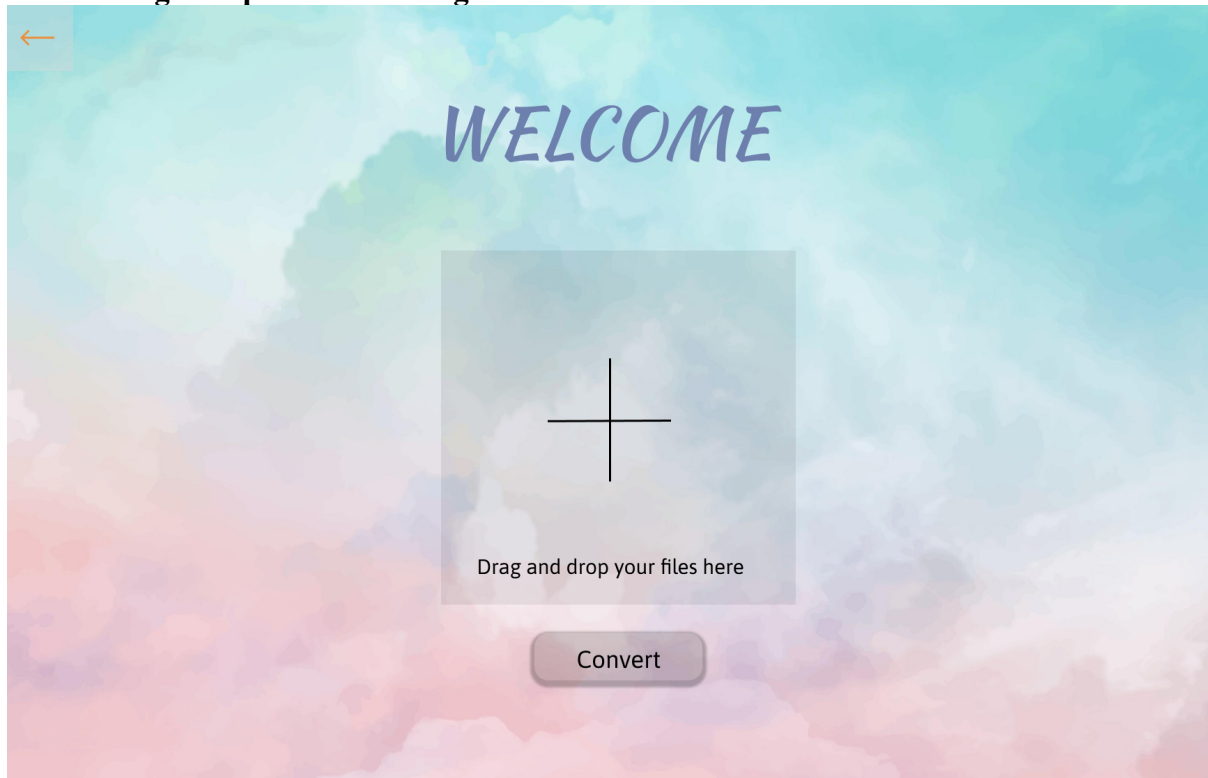


Fig 1.10

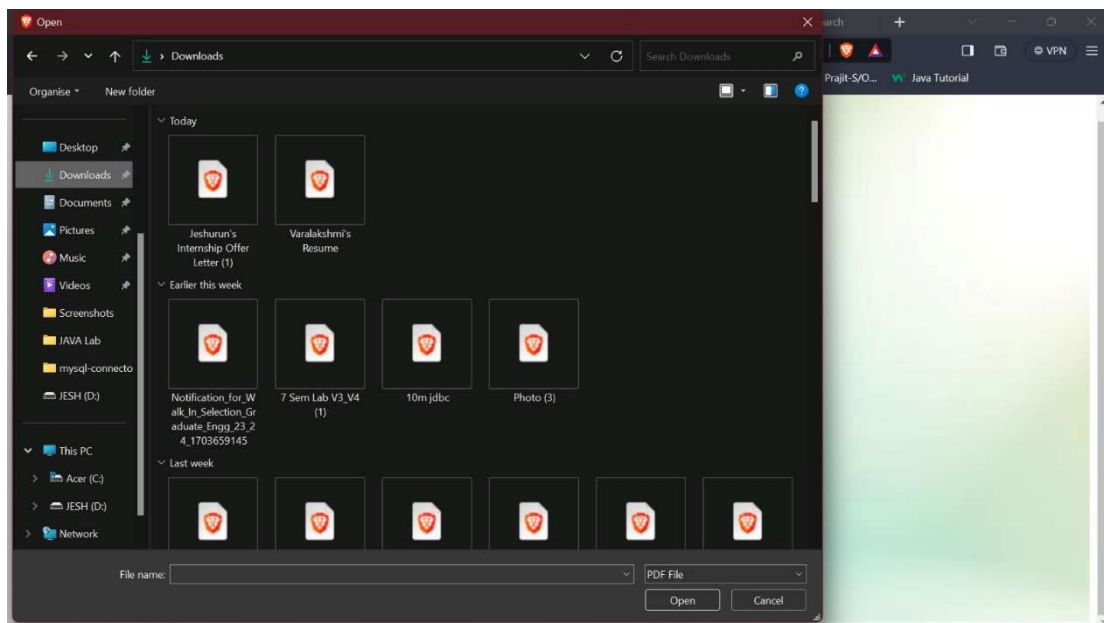


Fig 1.11 - On clicking the area, we can browse and add files

Now that we have dropped the file into the field, we can now hit on the convert button and then wait for the model to convert the file's text into a tabular format.

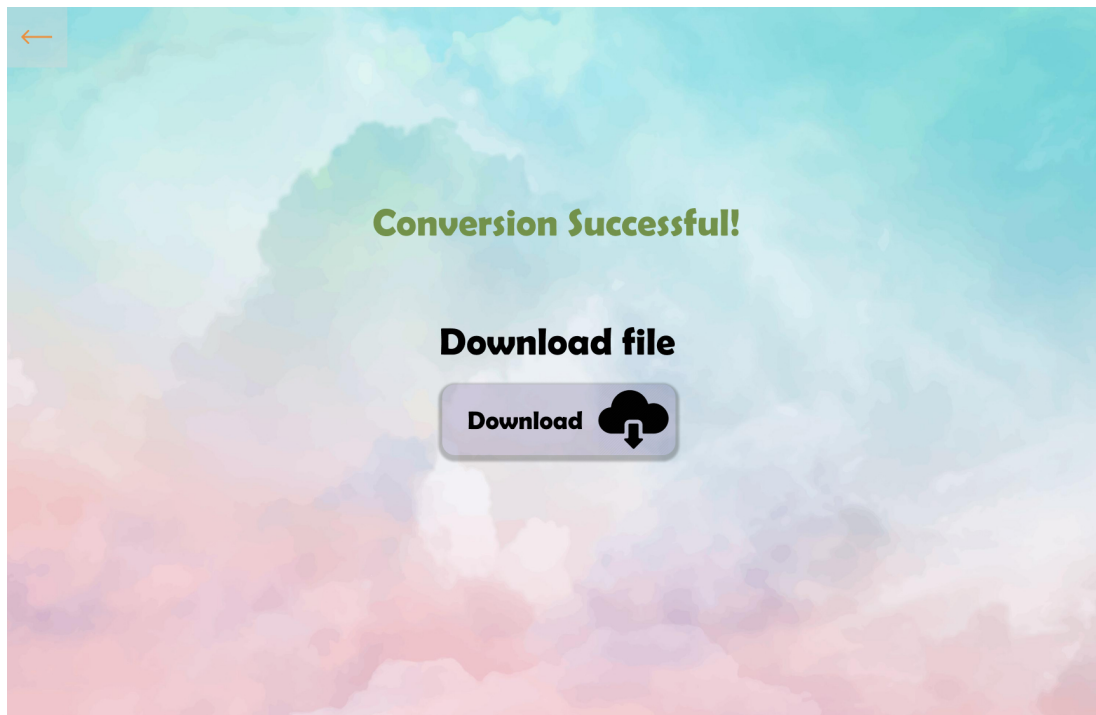


Fig 1.12

Shows an image of the file after conversion and now all we need to do is download the converted file in our desired format

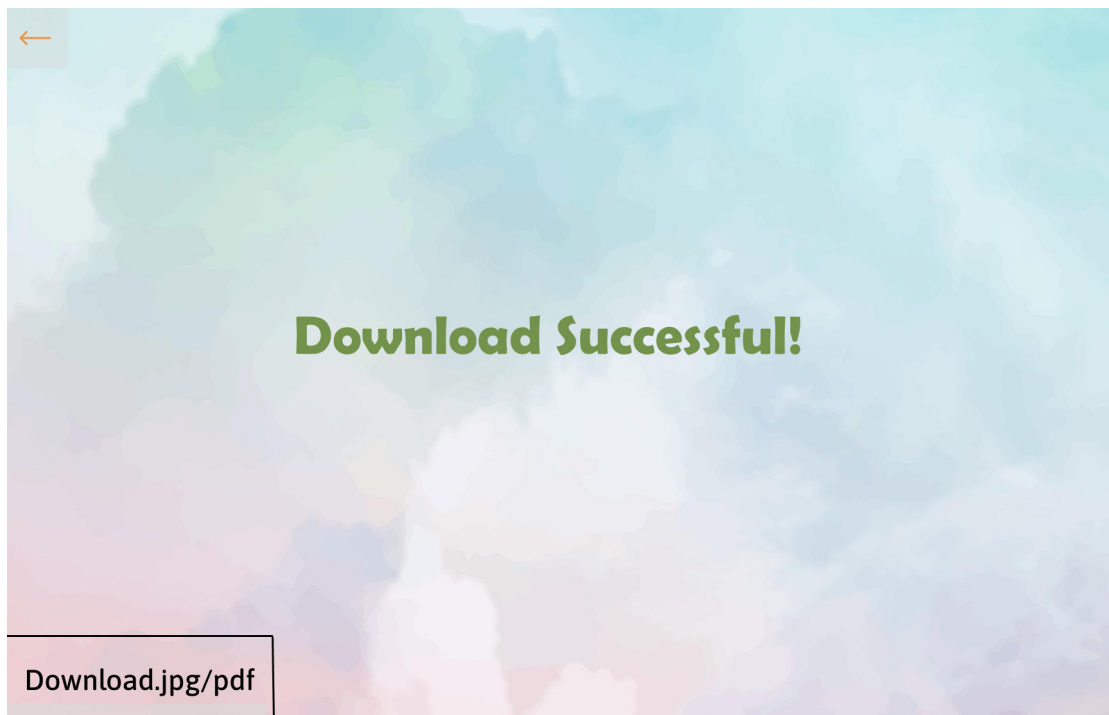


Fig 1.13

Now the download is successful and it can be viewed by clicking the bottom left corner as shown in the above snapshot

Text Extraction Using Optical Character Recognition

The image shows a screenshot of an invoice from Turnpike Designs Co. on the left and its corresponding tabular data extraction on the right. The invoice is titled "INVOICE" and includes the company logo and name. The "BILL TO" section lists the customer's name, address, and contact information. The "Invoice Number" and "Invoice Date" are also provided. The "Amount Due (USD)" is \$2,608.20. The "Services" section lists three items: Platinum web hosting package, 2 page website design, and Mobile designs. The "Total" is \$2,608.20. The "Amount due (CAD)" is \$2,608.20. The "EXTRACTED DATA" section on the right lists various fields and their values in a tabular format.

Field	Value
Invoice Number	14
Total amount	2608.2
Total net	2415
Date	2018-09-25
Due Date	2018-09-25
Supplier	TURNPIKE DESIGNS CO
Supplier Address	156 University Ave, Toront...
Supplier Registration	TIN: 770493581
Payment details	N/A
Customer	Jiro Doi
Customer Address	1954 Bloor Street West Tor...
Customer Registration	N/A
Language	en
Orientation (degrees)	0
Currency	CAD
Taxes	8% - 193.2
PO #	AD23094
Line items	1 - Platinum web hosting package - 65.00 3 - 2 page website design - 2100 1 - Mobile designs - 250

Fig 1.14

Here after the processing, we can see that the left side is the image inserted and the right side shows the text extracted from the image in a tabular format

OPTICAL CHARACTER RECOGNITION BACKEND MACHINE LEARNING MODEL

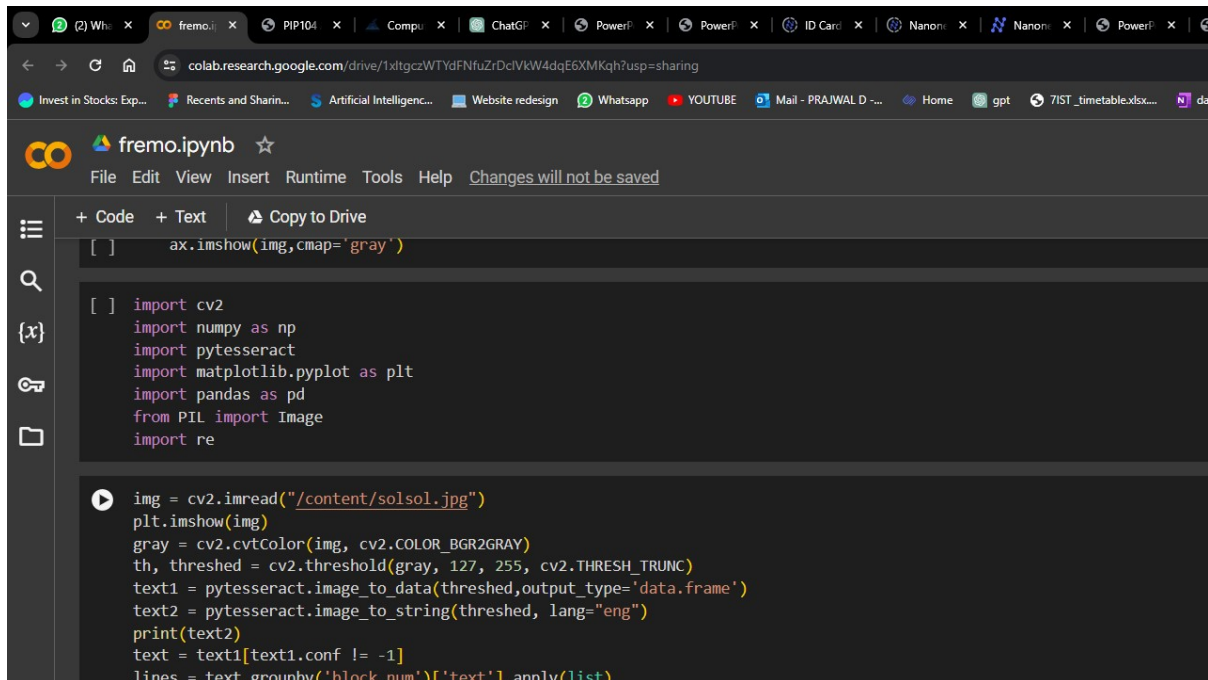


Fig 2.1

Here we can see that the model has recognized that an image has been inserted and it has started to read the text on the image

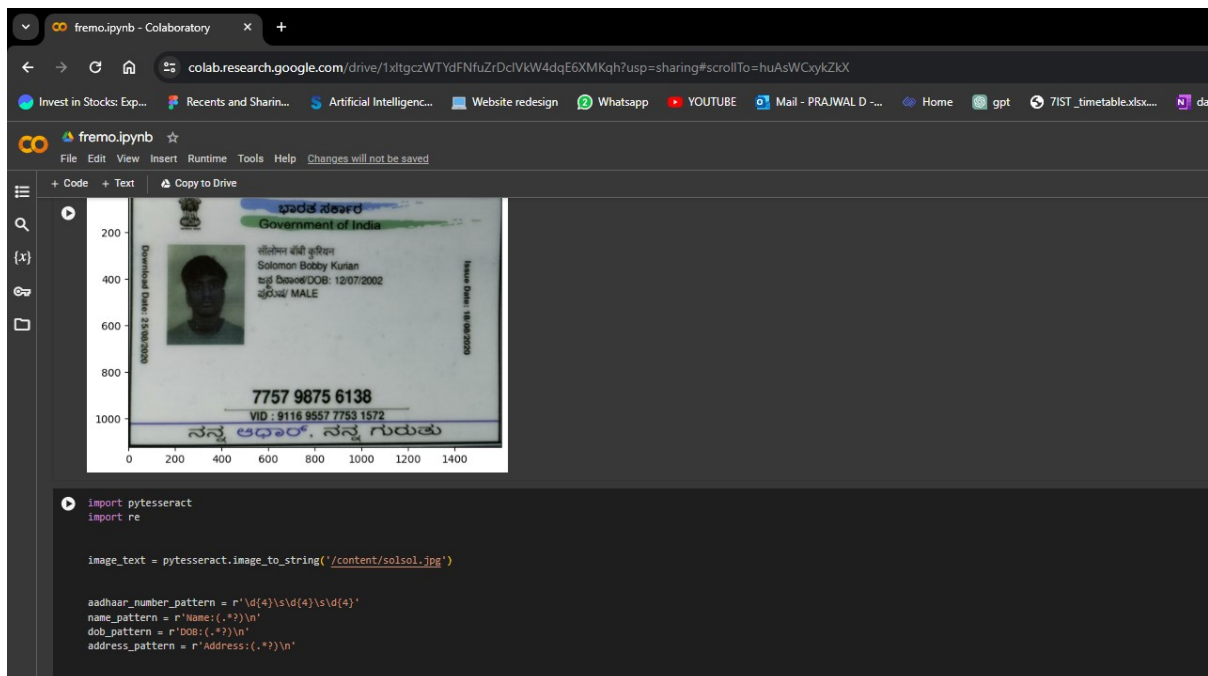


Fig 2.2

The above model recognizes it as an aadhaar card which comes under the identity card(ID) category and is able to recognize the text and show the respective output.

APPENDIX-C

ENCLOSURES

Plagiarism Report:

SIMILARITY INDEX **6%** **6%** INTERNET SOURCES
3% PUBLICATIONS
4% STUDENT PAPERS

PRIMARY SOURCES

1 www.geeky-gadgets.com Internet Source **1%**

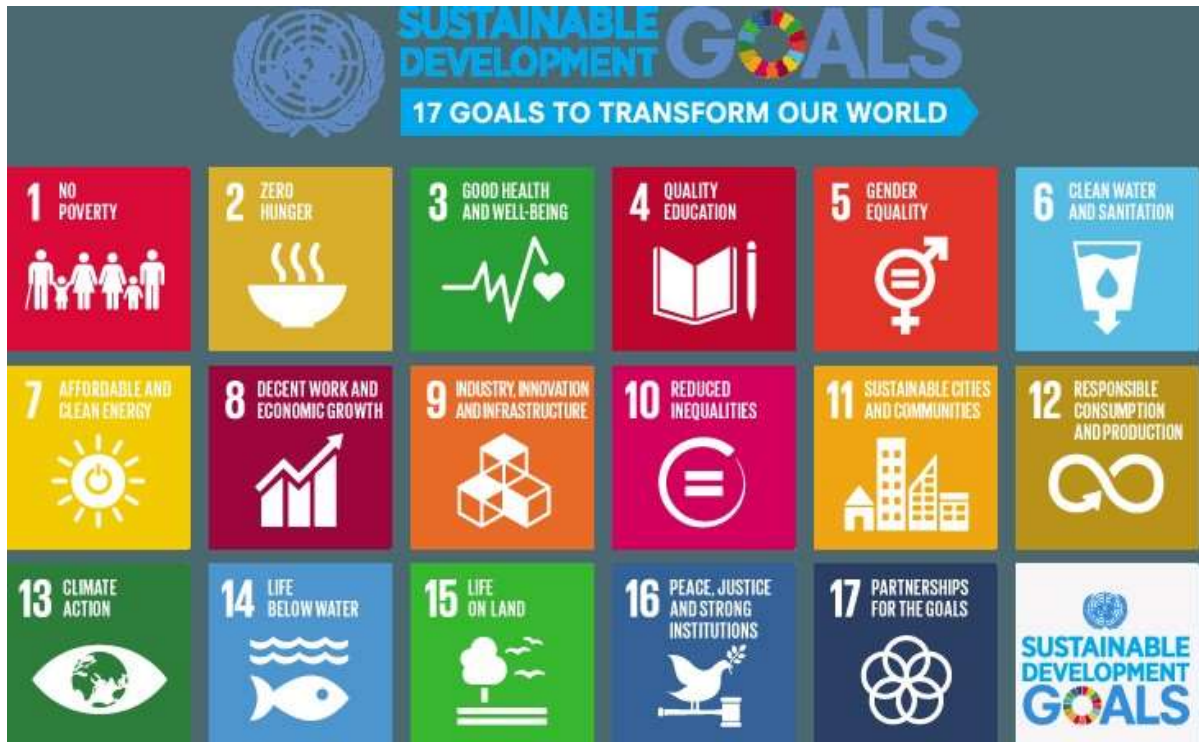
2 www.codewithc.com Internet Source **1%**

3 apps.dtic.mil Internet Source **1%**

4 Abhilasha Akkala, Avinash Seekoli.
"TEXTRACTOR-EXTRACT OPTIMAL
CHARACTER IN AN IMAGE FILE", Far East
Journal of Electronics and Communications, 2019
Publication **1%**

5 arxiv.org Internet Source **<1%**

Sustainable Development Goals:



The OCR website project aligns with **Sustainable Development Goal (SDG) -9: Industry, Innovation, and Infrastructure**, contributing to technological advancements and improved efficiency in data extraction.

Enhanced Productivity: By automating the extraction of text details from documents, the OCR website significantly reduces the manual effort required, promoting increased productivity within industries.

Innovation in Data Handling: The integration of OCR technology represents an innovative approach to handling textual information, fostering technological advancement and bringing about a positive shift in data management practices.

Infrastructure Optimization: The project optimizes infrastructure by leveraging OCR for streamlined data extraction, providing a sustainable solution for industries to adapt to evolving information needs without significant infrastructural overhauls.

Reduced Environmental Impact: The shift from manual data extraction to automated OCR processes reduces paper usage and associated environmental impacts, promoting a more sustainable and eco-friendly approach to information handling.