RV UNIVERSITY, BENGALURU-59

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A Project Report On

OCR for Electoral Roll Data

Submitted in partial Fulfillment for the award of degree of

BSc(Honors)

In

School of Computer Science and Engineering

Submitted By

Kishor Desai	USN: 1RVU22BSC044
Meghan D	USN: 1RVU22BSC049
Prabhas Bhat	USN: 1RVU22BSC069
Preetish Kumar Chanda	USN: 1RVU22BSC073

Under the Guidance of

Prof. Anoop A

Assistant Professor

School of CSE

RV University, Bengaluru-560059

2025-2026

RV UNIVERSITY, BENGALURU-59

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the mini project work titled OCR for Electoral Roll Data is carried out by Kishor Desai (USN: 1RVU22BSC044), Meghan D (USN: 1RVU22BSC049), Prabhas Bhat (1RVU22BSC069), Preetish Kumar Chanda (USN: 1RVU22BSC073) who are bonafide students of RV University, Bengaluru, in partial fulfilment of **Bachelor of Science(Hons) in School of Computer Science and Engineering** of the RV University, Bengaluru during the year 2025-2026. It is certified that all corrections/suggestions indicated for the Internal Assessment have been incorporated in the mini project report deposited in the departmental library. The Mini Project report has been approved as it satisfies the academic requirements in respect of mini project work prescribed by the institution for the said degree.

Signature of Guide	**Signature of Program Director**	**Signature of Dean**
Prof. Anoop A	**Dr. Lokanayaki K**	**Dr. Shobha G**

External Viva:

Name of Examiners	**Signature with Date**
1	
2	

DECLARATION

We, Kishor Desai, Meghan D, Prabhas Bhat and Preetish Kumar Chanda, students of seventh semester BSc(Hons), SoCSE, RV University, Bengaluru, hereby declare that the Major Project titled ‘OCR for Electoral Roll Data’ has been carried out by us and submitted in partial fulfillment of **Bachelor of Science (Hons)** in **School of Computer Science and Engineering** during the year 2025-26.

Further we declare that the content of the report has not been submitted previously by anybody or to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RV University will be the property of RV University, Bengaluru and we will be one of the authors of the same.

Place: Bengaluru

Date:

Name **Signature**

1. **Kishor Desai** (1RVU22BSC044)

2. **Meghan D** (1RVU22BSC049)

3. **Prabhas Bhat** (1RVU22BSC069)

4. **Preetish Kumar Chanda** (1RVU22BSC073)

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to School of Computer Science and Engineering, RV University for providing us with a great opportunity to pursue our bachelor's degree in this institution.

A special thanks to our Program Director **Dr. Lokanayaki K, Dean Dr. Shobha G** for their continuous support and providing the necessary facilities with guidance to carry out project work.

We would like to thank our guide **Prof. Anoop A, Assistant Professor**, School of Computer Science and Engineering, RV University, for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

ABSTRACT

Particularly in the last few election cycles, the reliability and accuracy of electoral rolls in India has increasingly come under the national and public scrutiny and concern. There is an increasing number of reports detailing issues regarding the rolls, such as duplicate and fraudulent entries, as well as invalid voter information. Even though the rolls can be accessed publicly, they're almost always posted as scanned PDF files fro uneditable documents. Moreover, these files do not conform to a standard uniform layout and are predominately in regional languages, unless the filed region has a certain threshold of literacy. This resulted in a scenario where large scale verification and investigation became infeasible. There are still digitization and analysis approaches that have used OCR and image processing that have been implemented to these documents in the past to attempt to make the documents more accessible and have faced these issues, to which the tools have shown efficienct but modest levels of success.

Now with improvements in DL models, new GPUs and LLMs, we believe we can achieve better results than the previous attempt. The structure of the project is to try 2 approaches – Image Processing + OCR and Image Processing + Multimodal LLM, compare the performance, execution times and accuracy for both approaches and decide on most optimal one. Finally, the extracted information will be structured and stored for further analysis.

TABLE OF CONTENTS

1. INTRODUCTION

The Election Commission of India (ECI) has made the Voter Records publicly available to ensure fairness and trust. Many times, systemic errors and inconsistencies like duplicate records, name-gender mismatch, photo-gender mismatch, etc have popped up and the most recent news is about deleted voters of Bihar. Even to identify this type of faults, it's very difficult because analysis of voter records is not easy.

The Voter data is available as scanned PDF, where each page of that pdf is not text, but instead it is a jpg image. So, for a human looking at the page, it looks normal but when trying to read it using computer, normal data analysis techniques fail because it is not text, it's an image. In such cases, OCR is the go-to solution but for this particular example, previous attempts and trials have shown that OCR does not perform well, does not have the accuracy needed and cannot support Indian regional languages. There are other issues of image quality, document layout inconsistencies, etc which further add to problem.

Now with improvements in DL models, new GPUs and LLMs, we believe we can achieve better results than the previous attempt. The structure of the project is to try 2 approaches – Image Processing + OCR and Image Processing + Multimodal LLM, compare the performance, execution times and accuracy for both approaches and decide on most optimal one. Finally, the extracted information will be structured and stored for further analysis.

2. OBJECTIVES

Our main expectation from the project is to develop a system to extract data from voter records and store it in a structured way to enable downstream tasks like analysis, visualization, drill-down, etc. As discussed earlier, we will try 2 approaches: Image Processing + OCR and Image Processing + Multimodal LLM

We are hoping that the challenges mentioned about Image quality, document layout, orientation, etc will be solved by proper Image Preprocessing techniques and our literature survey suggests the same. Handling regional languages, structure and formatting issues will be taken care by LLMs as they are trained on huge dataset of internet text and they are intelligent enough to fix these issues.

The ultimate objective is to develop an end-to-end solution whereby registered voter information, such as name, age, gender, and address, could be automatically identified and extracted, cleaned, and then presented in a structured digital format. This would make the process of data verification and analysis easier, as well as help in future research and administrative works to improve integrity and transparency in electoral data management.

3. LITERATURE SURVEY AND RESEARCH GAP

1. **Paper Title:** Ballot Character Recognition Based on Image Processing

   Date of Issue: 2021

   Methodology/Algorithms used: Image preprocessing, segmentation, and OCR pipeline

   Dataset: Scanned ballot images

   Evaluation Metrics: Character recognition accuracy

   Tools/Frameworks: OpenCV, MATLAB, conventional OCR

   Research gap: Needs integration with secure voting systems

   Result/Findings: Improved character recognition via image processing techniques

   Remarks: Applicable to ballot-specific extraction tasks

2. **Paper Title:** OCR-free Document Understanding Transformer (Donut / OCR-free approaches)

   Date of Issue: 2022

   Methodology/Algorithm used: End-to-end transformer-based document understanding without explicit OCR

   Dataset: DocVQA, synthetic document datasets

   Evaluation Metrics: F1, precision, recall on extraction tasks

   Tools/Frameworks: PyTorch, Hugging Face Transformers

   Research gap: Needs domain adaptation for electoral roll formats

   Result/Findings: Achieved SOTA on structured document understanding without OCR

   Remarks: Promising alternative to pipeline OCR for structured extraction

3. **Paper Title:** Optical Character Recognition with Neural Networks and Post-correction with Finite State Methods

   Date of Issue: 2020

   Methodology/Algorithm used: Deep neural network OCR plus finite-state post-correction

   Dataset: Historical Finnish/Swedish corpora

   Evaluation Metrics: CER/WER improvements after post-correction

   Tools/Frameworks: Neural models + FST toolkits

   Research gap: Multilingual scalability and computational cost concerns

   Result/Findings: Achieved high accuracy across fonts and historical texts

   Remarks: Shows benefit of post-correction strategies

4. **Project Title:** Document Verification Using OCR

   Date of Issue: 2024

   Methodology/Algorithm used: YOLOv8 for ROI detection + OCR + signature matching for verification

   Dataset: Passport and ID datasets

   Evaluation Metrics: Overall verification accuracy (~75% reported)

   Tools/Frameworks: YOLOv8, OCR engines, signature matching libraries

   Research gap: Need robustness across diverse formats and qualities

   Result/Findings: Shows practical verification pipeline with moderate accuracy

   Remarks: Applicable concept for voter ID verification

5. **Project Title:** Integrating OCR and LLMs for Enhanced Document Digitization in ERP Systems

   Date of Issue: 2024

Methodology/Algorithm used: Pipeline combining coordinate-based grouping + OCR + LLM few-shot learning

Dataset: ERP document corpora

Evaluation Metrics: Extraction accuracy and throughput

Tools/Frameworks: OCR engines, LLMs, grouping heuristics

Research gap: Struggles with diverse/unstructured formats

Result/Findings: Improved digitization performance in ERP contexts

Remarks: Shows practical integration patterns

6. **Paper Title:** Basic research on a handwritten note image recognition system that combines two OCRs

   Date of Issue: 2021

   Methodology/Algorithm used: Combining outputs from multiple OCRs (e.g., Tesseract + Mathpix) with selection heuristics

   Dataset: Handwritten Japanese/math datasets

   Evaluation Metrics: Recognition accuracy improvements

   Tools/Frameworks: Multiple OCRs, selection heuristics

   Research gap: Selection based only on OCR scores unreliable

   Result/Findings: Combined OCRs improved recognition in experiments

   Remarks: Ensemble selection requires smarter decision logic

7. **Paper Title:** Leveraging LLMs for Post-OCR Correction of Historical Newspapers

   Date of Issue: 2024

   Methodology/Algorithm used: Fine-tuning Llama 2 for OCR post-correction

   Dataset: Historical newspaper OCR outputs

   Evaluation Metrics: Error rate reduction metrics

   Tools/Frameworks: LLMs (Llama 2), fine-tuning frameworks

Research gap: Need multilingual fine-tuning and generalization

Result/Findings: Llama 2 outperformed BART; significant error reduction reported

Remarks: Shows LLMs effectiveness in post-correction tasks

8. **Paper Title:** Scrambled text: Fine tuning Language models for OCR Error correction using synthetic data.

   Date of Issue: 2025

   Methodology/Algorithm used: Fine-tuning LMs on synthetic OCR errors for post-correction

   Dataset: Synthetic OCR-error corpora (English focus)

   Evaluation Metrics: Error reduction metrics (CER/WER), downstream task F1

   Tools/Frameworks: Transformers, IJDAR implementations

   Research gap: Focused mainly on English; multilingual challenges remain

   Result/Findin**gs:** Synthetic fine-tuning improves post-correction effectiveness

   Remarks: Shows potential for synthetic-data-driven postprocessing

4. PROJECT DESCRIPTION

The voter records are publicly available as scanned pdfs where each page of pdf is an image, which makes data extraction very difficult. Moreover, the tightly packed nature of data and variations in layouts, languages, etc adds to difficulty. The goal of our project is to develop a system to extract voter data from electoral roll records. The difficulties will be handled through image processing techniques, integration of newer OCR tools and large language models

SYSTEM OVERVIEW

The system will have 4 main modules. First module is the input and pre-processing module. Second module is the data extraction module which is heart of the system. Post processing of data will be handled in module-3. Module-4 for enabling storage and analysis of data.

SYSTEM ARCHITECTURE

There will be 4 Modules.

1) Input and Preprocessing: Individual pages of pdf will be saved as jpg images, then these images will be preprocessed for noise removal, contrast enhancement, etc

2) Data Extraction Module: We will be trying 2 approaches - one to break down the larger image into per-record image and then run OCR on this smaller image to extract data; second approach will be to pass the large image to an multi-modal LLM to extract data.

3) Post Processing: Post-processing of data like text cleaning, field matching, formatting, structuring, etc.

4) Storage: Storing data in structured/relational format.

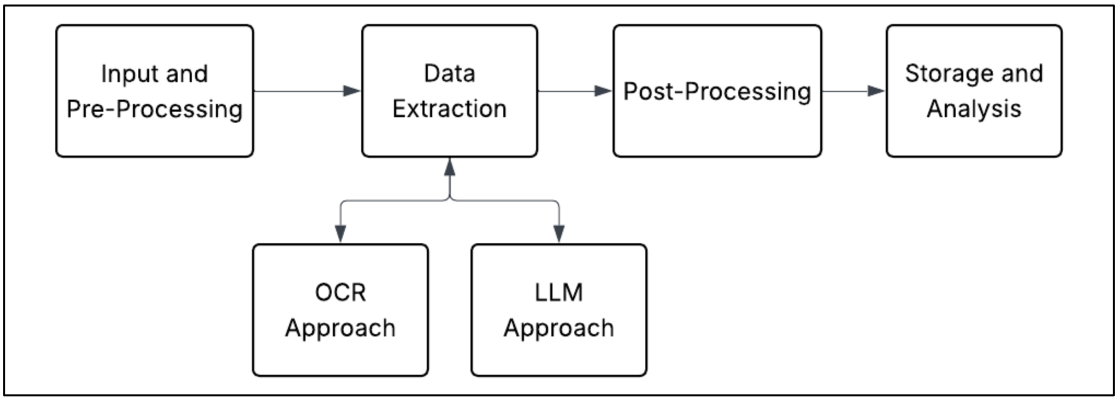


Figure 1: System Architecture with Modules Involved

TOOLS AND TECHNOLOGIES

1) Python and its libraries like pypdf, opencv, Pandas, Numpy, PyTorch, TensorFlow, others.

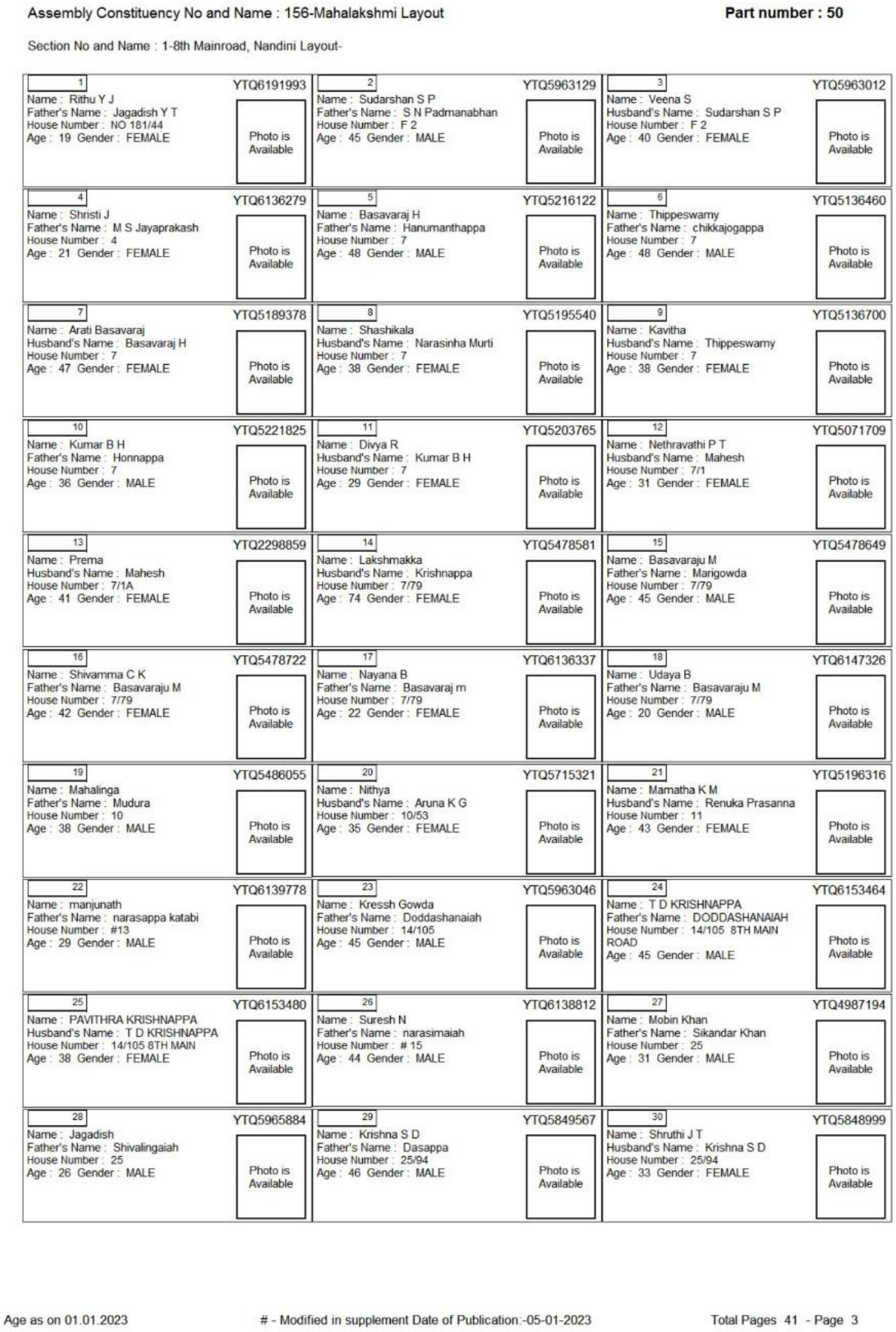
2) Tesseract OCR for text recognition.

3) YOLO model for bounding box detection

4) Multi-modal LLMs and Lightweight LLMs from repositories like HuggingFace for Module-2 and Module-3 respectively

DATASET

The dataset is publicly available as scanned pdfs where each page of pdf is an image. The information is present in multiple Indian regional languages. Each page will have 3 columns fixed and up to 10 rows, making it a 10x3 table. Each cell of this table holds information for an individual and each cell will have 8 data fields of importance. Additionally, the header and footer of a page also contain some information.



Assembly Constituency No and Name : 156-Mahalakshmi Layout

Part number : 50

Section No and Name : 1-8th Mainroad, Nandini Layout-

1 YTQ6191993 Name : Rithu Y J Father's Name : Jagadish Y T House Number : NO 181/44 Age : 19 Gender : FEMALE Photo is Available	2 YTQ5963129 Name : Sudarshan S P Father's Name : S N Padmanabhan House Number : F 2 Age : 45 Gender : MALE Photo is Available	3 YTQ5963012 Name : Veena S Husband's Name : Sudarshan S P House Number : F 2 Age : 40 Gender : FEMALE Photo is Available
4 YTQ6136279 Name : Shristi J Father's Name : M S Jayaprakash House Number : 4 Age : 21 Gender : FEMALE Photo is Available	5 YTQ5216122 Name : Basavaraj H Father's Name : Hanumanthappa House Number : 7 Age : 48 Gender : MALE Photo is Available	6 YTQ5136460 Name : Thippeswamy Father's Name : chikkajogappa House Number : 7 Age : 48 Gender : MALE Photo is Available
7 YTQ5189378 Name : Arati Basavaraj Husband's Name : Basavaraj H House Number : 7 Age : 47 Gender : FEMALE Photo is Available	8 YTQ5195540 Name : Shashikala Husband's Name : Narasinha Murti House Number : 7 Age : 38 Gender : FEMALE Photo is Available	9 YTQ5136700 Name : Kavitha Husband's Name : Thippeswamy House Number : 7 Age : 38 Gender : FEMALE Photo is Available
10 YTQ5221825 Name : Kumar B H Father's Name : Honnappa House Number : 7 Age : 36 Gender : MALE Photo is Available	11 YTQ5203765 Name : Divya R Husband's Name : Kumar B H House Number : 7 Age : 29 Gender : FEMALE Photo is Available	12 YTQ5071709 Name : Nethravathi P T Husband's Name : Mahesh House Number : 7/1 Age : 31 Gender : FEMALE Photo is Available
13 YTQ2298859 Name : Prema Husband's Name : Mahesh House Number : 7/1A Age : 41 Gender : FEMALE Photo is Available	14 YTQ5478581 Name : Lakshmakka Husband's Name : Krishnappa House Number : 7/79 Age : 74 Gender : FEMALE Photo is Available	15 YTQ5478649 Name : Basavaraju M Father's Name : Marigowda House Number : 7/79 Age : 45 Gender : MALE Photo is Available
16 YTQ5478722 Name : Shivamma C K Father's Name : Basavaraju M House Number : 7/79 Age : 42 Gender : FEMALE Photo is Available	17 YTQ6136337 Name : Nayana B Father's Name : Basavaraj m House Number : 7/79 Age : 22 Gender : FEMALE Photo is Available	18 YTQ6147326 Name : Udaya B Father's Name : Basavaraju M House Number : 7/79 Age : 20 Gender : MALE Photo is Available
19 YTQ5486055 Name : Mahalinga Father's Name : Mudura House Number : 10 Age : 38 Gender : MALE Photo is Available	20 YTQ5715321 Name : Nithya Husband's Name : Aruna K G House Number : 10/53 Age : 35 Gender : FEMALE Photo is Available	21 YTQ5196316 Name : Mamatha K M Husband's Name : Renuka Prasanna House Number : 11 Age : 43 Gender : FEMALE Photo is Available
22 YTQ6139778 Name : manjunath Father's Name : narasappa katabi House Number : #13 Age : 29 Gender : MALE Photo is Available	23 YTQ5963046 Name : Kressh Gowda Father's Name : Doddashanaiah House Number : 14/105 Age : 45 Gender : MALE Photo is Available	24 YTQ6153464 Name : T D KRISHNAPPA Father's Name : DODDASHANAIAH House Number : 14/105 8TH MAIN ROAD Age : 45 Gender : MALE Photo is Available
25 YTQ6153480 Name : PAVITHRA KRISHNAPPA Husband's Name : T D KRISHNAPPA House Number : 14/105 8TH MAIN Age : 38 Gender : FEMALE Photo is Available	26 YTQ6138812 Name : Suresh N Father's Name : narasimaiah House Number : # 15 Age : 44 Gender : MALE Photo is Available	27 YTQ4987194 Name : Mobin Khan Father's Name : Sikandar Khan House Number : 25 Age : 31 Gender : MALE Photo is Available
28 YTQ5965884 Name : Jagadish Father's Name : Shivalingaiah House Number : 25 Age : 26 Gender : MALE Photo is Available	29 YTQ5849567 Name : Krishna S D Father's Name : Dasappa House Number : 25/94 Age : 46 Gender : MALE Photo is Available	30 YTQ5848999 Name : Shruthi J T Husband's Name : Krishna S D House Number : 25/94 Age : 33 Gender : FEMALE Photo is Available

Age as on 01.01.2023 # - Modified in supplement Date of Publication:-05-01-2023 Total Pages 41 - Page 3

Figure 2: Sample Dataset

5. METHODOLOGY / TECHNICAL IMPLEMENTATION

As discussed earlier, the system will have 4 modules and 2 methods/approaches for data extraction. Key stages are as below:

1. DATA COLLECTION AND PREPROCESSING

First step is downloading the dataset from the source. Dataset will be a pdf file and each page of this pdf will be split into individual jpg images. The images will be then processed to improve downstream results.

1. Grayscale conversion: to reduce noise and simplify the image.
2. Thresholding and binarization in order to separate the text from the background.
3. Noise removal and morphological operations: small mark cleaning and enhancement of text regions.
4. Skew correction: to properly align the text for correct OCR recognition.
5. This will ensure that the images are clean and prepared for accurate text extraction.

2. LAYOUT AND STRUCTURE DETECTION

The general appearance of an electoral roll page is tabular. The contour detection and projection-based segmentation techniques identify boundaries of each rectangular cell accurately to capture the actual structure. Each of the identified cells is cropped out individually and fed to the OCR module for extraction at the field level. This provides cell-level segmentation so that each and every voter record gets processed in isolation, reducing overlap errors and enhancing data consistency.

3. OPTICAL CHARACTER RECOGNITION (OCR)

The OCR step is responsible for converting processed image regions to machine-readable text. In this regard, two different OCR engines are tested and compared:

Tesseract OCR is widely popular due to its flexibility and support for most languages.

Google Cloud Vision OCR API: cloud-based service with better performance for complex layouts.

Each cropped cell then undergoes processing with the OCR engine, while the recognized text is stored together with its position coordinates. Post-processing is applied to correct common OCR errors, such as misread characters and inconsistent formatting.

4. TEXT POST-PROCESSING AND FIELD EXTRACTION

After OCR, the raw text output often contains noise and inconsistencies. String processing techniques, therefore, extract the relevant voter fields such as Name, Age, Gender, House Number, and Address. Various techniques such as regular expressions, keyword matching, and rule-based parsing are used to identify key-value pairs in each cell. Extracted fields are cleaned, standardized, and validated before adding them to the final dataset.

5. DEEP LEARNING / YOLO-BASED APPROACH

In hope of improving data extraction, the cells in the 10x3 table will be annotated and boundary boxes will be drawn to train a deep learning model like YOLO to identify cells containing information and then individual cells will be passed to the OCR engine to extract data. Instead of extracting 240 fields from a large image, the OCR has to now extract just 8 fields from a smaller image. This is the divide-and-conquer technique to improve quality of the approach.

6. EXTRACTION BY LARGE LANGUAGE MODEL (LLM)

The other approach for data extraction is to use a multi-modal LLM. individual image from pdf containing the 10x3 table will be directly given as input to a multi-modal LLM to extract information from the image. In the data postprocessing module, an light-weight text only LLM will be enough for tasks like text cleaning, formatting, structuring, etc.

7. DATA STRUCTURING AND OUTPUT GENERATION

Since each voter record has 8 data fields, the postprocessed data from module-3 will be stored in a structured, relational format like csv or relational database which will also enable downstream tasks like analysis, visualization, etc. Quantitative metrics like Accuracy (number of correctly extracted fields / total number of fields), execution times, etc will be recorded

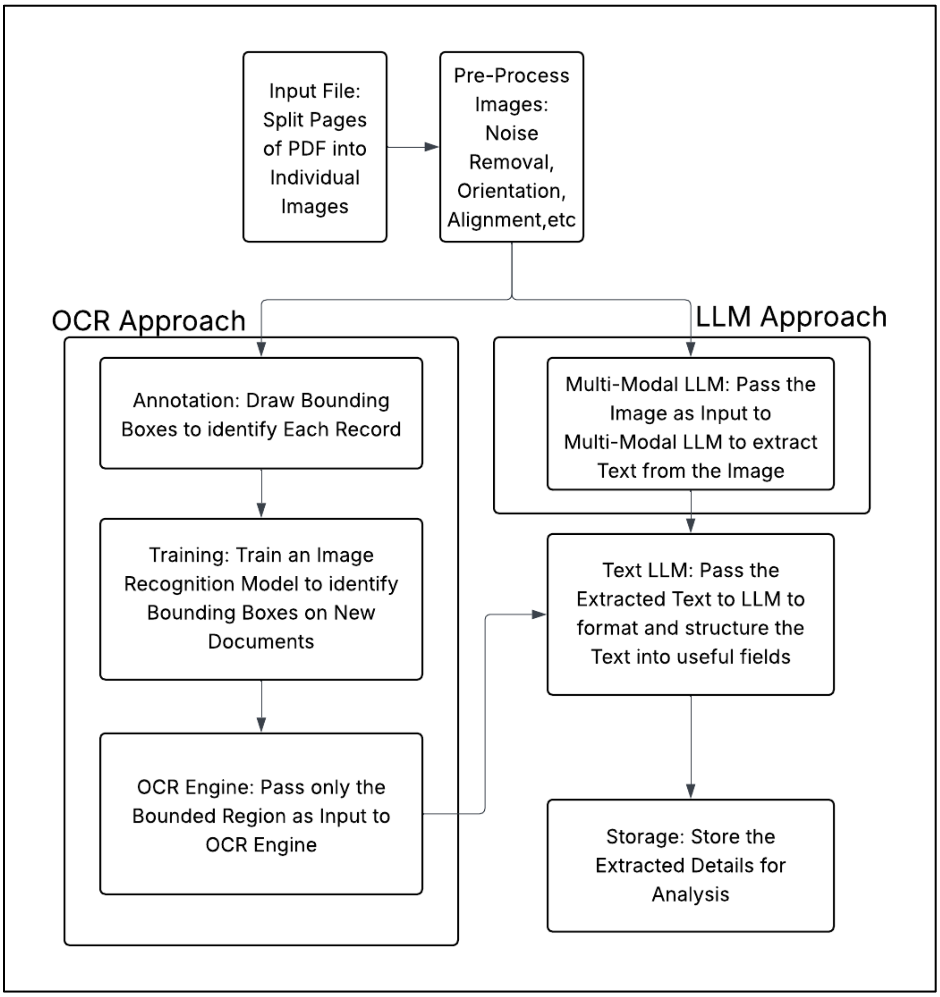

Input File: Split Pages of PDF into Individual Images
Pre-Process Images: Noise Removal, Orientation, Alignment,etc
OCR Approach
LLM Approach
Annotation: Draw Bounding Boxes to identify Each Record
Training: Train an Image Recognition Model to identify Bounding Boxes on New Documents
OCR Engine: Pass only the Bounded Region as Input to OCR Engine
Multi-Modal LLM: Pass the Image as Input to Multi-Modal LLM to extract Text from the Image
Text LLM: Pass the Extracted Text to LLM to format and structure the Text into useful fields
Storage: Store the Extracted Details for Analysis

Figure 3: Methodology and Implementation

6. CONCLUSION AND FUTURE SCOPE

CONCLUSION

This project accomplishments only include the efficient and dependable system that it put in place for the processing of structured voter information obtained. Through the use of image preprocessing, the system was able to accomplish the task of turning intricate, image-based files into machine readable organized data. Experimentation with cell-based segmentation, modern LLM-assisted extraction, and conventional OCR pipelines has provided information into the numerous advantages and disadvantages of each of the varied approaches.

As highlighted in the results, there was very high accuracy for structured documents based on cell level segmentation and preprocessing support. While the LLMs on the other hand had certain advantageous features when it came to complicated layout processing and contextual OCR error correction. Overall, the project created a foundation that will assist in the automation of SSD functionalities that are related to electoral rolls.

FUTURE SCOPE

Improvement and expanding this project even further are possible and this project areas are:

Increasing Model Fidelity: Identifying unrefined poorly defined scanned documents using either hybrid OCR-LLM systems or OCR with fine-tuned deep learning models for deep learning systems.

Support for More than One Language: Different than English scripts are published for India’s electoral rolls increasing recognition systems for including more regional languages is a worthy expansion.

Automated Checks: Automated duplicate entry and invalid entry detection software integration for non-closed documents is vital for cross verification with government and open record databases.

Developing Scalable Systems for Available and Usable Applications: Building systems that are scalable, in this case, systems that are web-based or cloud-based, and that provide services for structured and validated electronic roll documents while uploading pdfs for the electronic documents would be very useful.

Training and Future Model Development: Future deep learning systems require, specifically for new models training, the availability of a high-quality labeled data set or an annotation tool and generating a labeled dataset is vital and a simple annotation tool for generating labeled datasets would be very useful.

Improvement in processing and development of this project has the potential to make a great impact for both the researchers analyzing voters' data for optimizing and increasing transparency in the electoral systems as well as the election authorities. This project can facilitate further development and provide the foundation for a full functional solution.

7. REFERENCES

[1] A. Neupane, A. Lamichhane, A. Paudel, and A. Shakya, “Structured Information Extraction from Nepali Scanned Documents using Layout Transformer and LLMs,” *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pp. 134–143, Jan. 2025.

[2] U. Gupta, A. Kumar, A. Gupta, G. Raj, and A. P. Agrawal, “Advances in Handwritten Character Recognition: A Comparison of OCR and Large Language Model-Based Approaches,” *2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)*, IEEE, pp. 192–198, 2024.

[3] E. Sabir, S. Rawls, and P. Natarajan, “Implicit Language Model in LSTM for OCR,” *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 27–31, 2017.

[4] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, and F. Shafait, “High-Performance OCR for Printed English and Fraktur Using LSTM Networks,” *12th International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 683–687, 2013.

[5] J. Zhang, W. Haverals, M. Naydan, and B. W. Kernighan, “Post-OCR Correction with OpenAI’s GPT Models on Challenging English Prosody Texts,” *Proceedings of the ACM Symposium on Document Engineering 2024*, pp. 1–4, 2024.

[6] H. Vaswani et al., “Attention is All You Need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *arXiv preprint arXiv:1301.3781*, 2013.

[8] Y. Xu et al., “LayoutLMv2: Multi-modal Pre-training for Visually rich Document Understanding,” *arXiv preprint arXiv:2012.14740*, 2022.

[9] LLaMA Team, “The LLaMA 3 Herd of Models,” *arXiv preprint arXiv:2407.21783*, 2024.

APPENDICES

APPENDIX A: WORK DIARY

Include weekly progress logs, guide meetings, and reflections.

APPENDIX B: PLAGIARISM REPORT

Attach Turnitin plagiarism report (<20%).

APPENDIX C: REVIEW PAPER DRAFT

Attach draft or published version of the review paper with proof of submission.