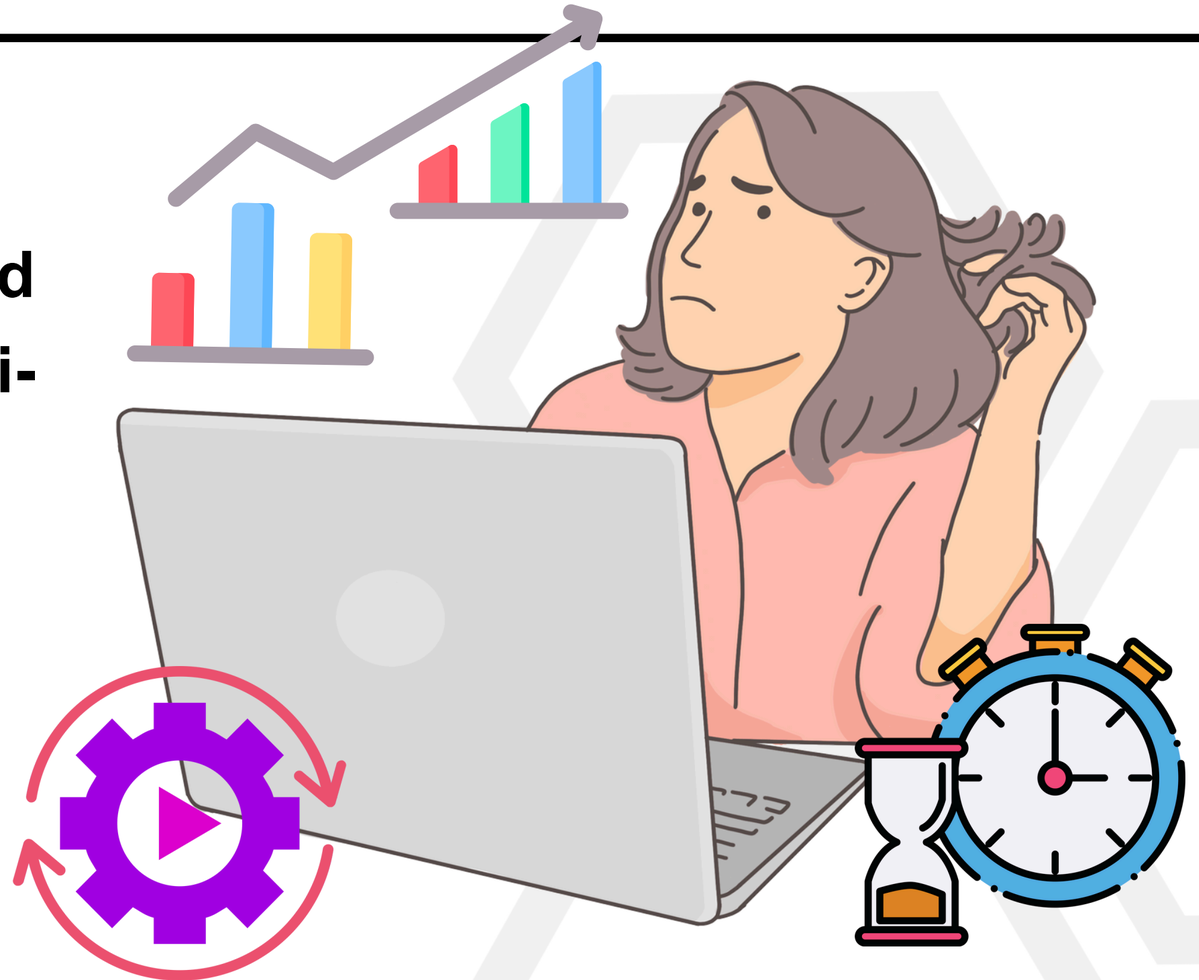
SMART INDIA HACKATHON 2024

Project DocLense

- **Problem Statement ID -** SIH1652
- **Problem Statement Title -** Extraction and Verification of Information from semi-categorised data.
- **Theme -** Smart Automation
- **PS Category -** Software
- **Team ID -** 18684
- **Team Name -** HIVE



HIVE

Project DocLens

SMART INDIA HACKATHON 2024

Approach of DocLens

The Process is mainly divided into two parts:

- **Document Template Configuration (One-time Training Phase) :**
  - The System contains all **sample documents in all possible formats** (eg. Aadhar, eAadhar, PVC Aadhar, etc). and more can be added if required.
  - **Fields to be added in the form are configured.**
  - **Documents to be verified are selected.**
  - **Document Template specific prompts are configured.**

- **Application Form Process:**
  - The applicant **fills** out the form and uploads supporting **documents**.
  - **Pre-processing of the input fields for eg.: validation of Aadhaar Number or phone number and other fields is done.**
  - **Text Extraction of only specific fields is performed by the vision model.**
  - **The Extracted text is post processes for fuzzy values**
    - **for eg.: Different Naming Formats: <Name><Surname> & <Surname><Name>**
    - **Different date formats : DD/MM/YYYY & MM/DD/YYYY**
  - **The processed extracted text is verified with the inputs entered by the user.**
  - **If verified then the data and documents with verification report is stored in the database.**
  - **If not verified, then the applicant has the access to request for a manual document verification.**

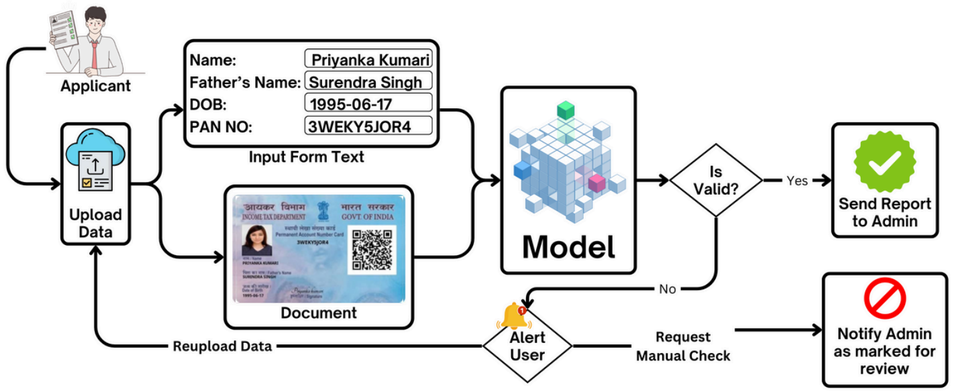
Handling of tables in the document:

a. If a table exists, its **position is identified.**
b. **OCR** is applied specifically to the table area.
c. The extracted table content is **reconstructed into CSV or XLS format** with the help of their **positions**.
d. **Calculations** like **totals, percentages, and averages** are computed for values which are **not explicitly mentioned in the document.**

Handling Documents with No Definite/Specific Template

- NLP techniques like **NER and pattern matching extract key details** such as employment duration, skills, and sentiment from experience certificates.

High Level Overview



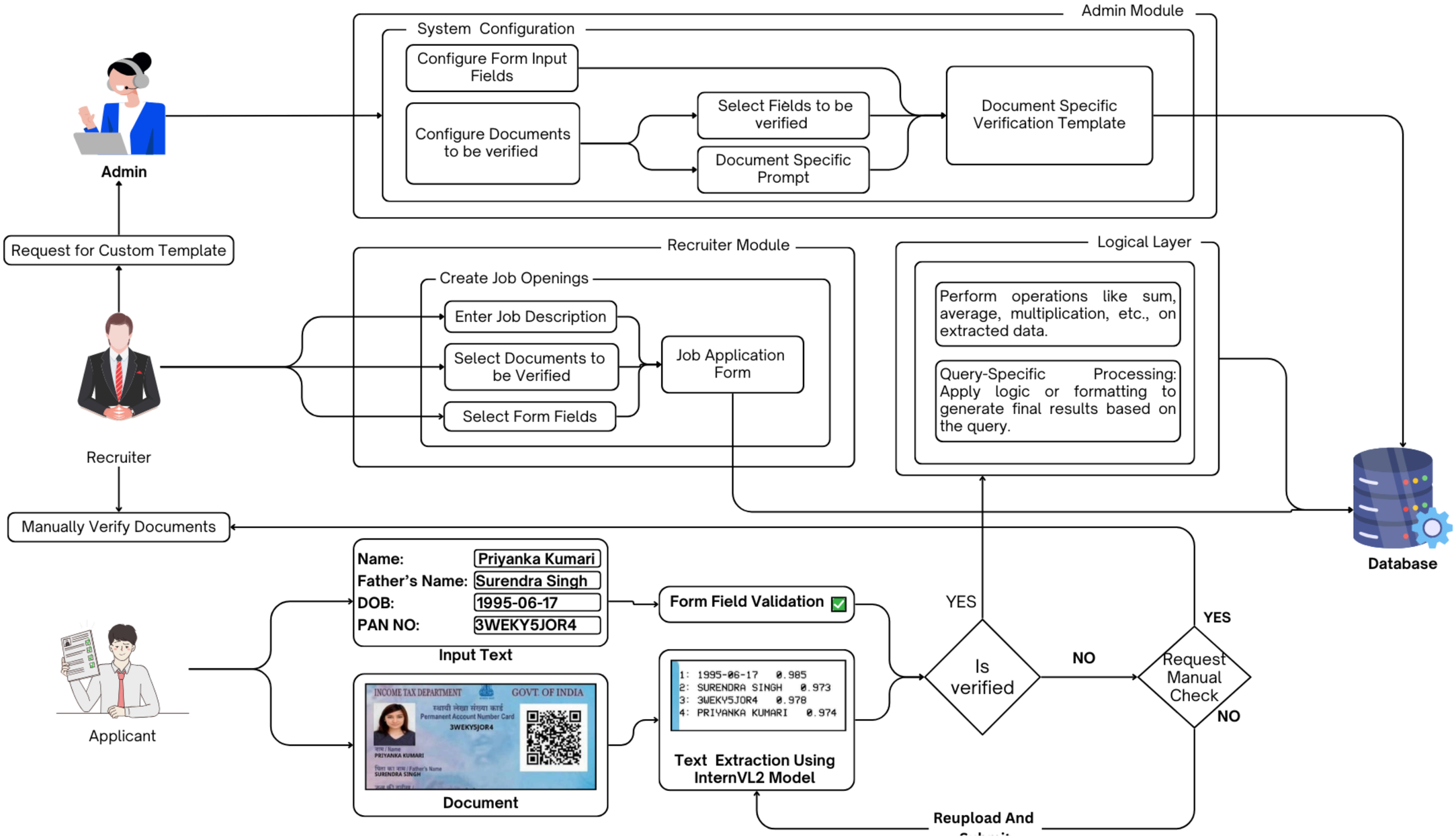
Need for Solution

- The Process is currently done **manually.**
- Manual verification is **time-consuming.** Hence the applicants **receives delayed feedback/results.**
- **The manual process is prone to human error.**
- To work on documents in different languages more manpower is required, along with proficiency in other languages.
- Indian languages are not fully supported by the existing software
- Currently **no major homegrown solution** is available

Uniqueness of Solution

- Easy Initial Onetime Customizable Setup
- No huge training datasets required
- Supports Indian Languages
- Simple System Architecture

Technical Approach



Tech Stack

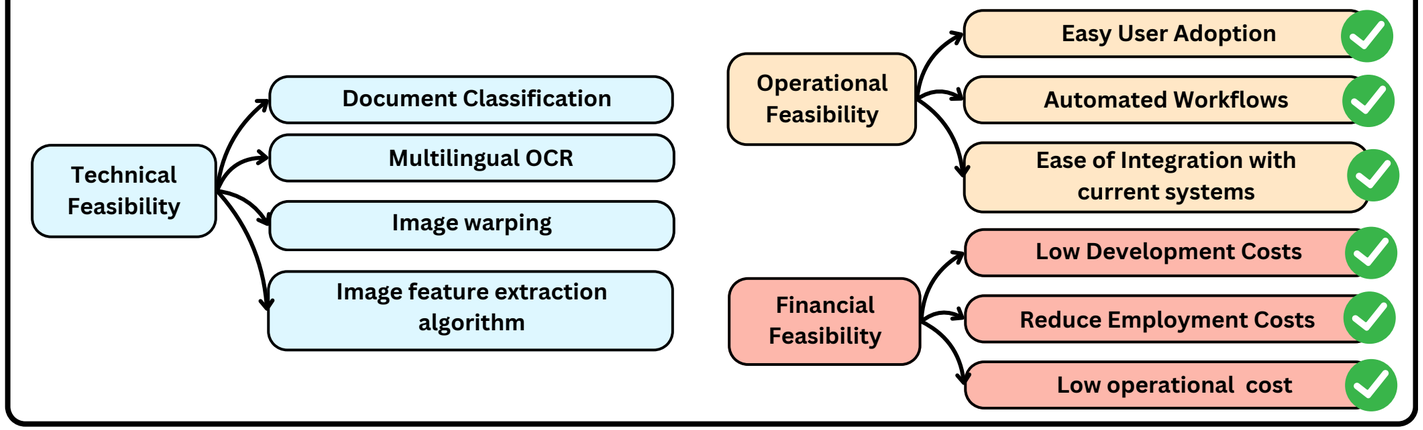
- OpenCV
- PaddleOCR — OCR (Optical Character Recognition)
- Python
- Django
- PostgreSQL
- Javascript

HIVE

Feasibility and Viability

SMART INDIA HACKATHON 2024

Feasibility and Viability in Different Aspects



Future Scope

- Integrating **super-resolution of low-resolution images/docs** algorithms.
- **Integration with national databases** like DigiLocker, Aadhaar, and PAN for real-time validation of authenticity of uploaded documents.
- Extending the **application of this technology in the form of an library/API** to be used in other applications of government sectors, educational institutions, and private enterprises, making it a universal solution for document verification.

Security Measures

- **Data Access Control only to RAC Officials.** No Personal Data Stored in our system
- No need for any **personal documents** for **training**, instead document templates is used.
- **Local Data Processing** without use of Third-Party APIs
- Role-Based Access Control (RBAC) to the verification reports
- Data Encryption in Transit.

Potential Challenges, Risks and How To Mitigate Them

	Challenges	Risks	Mitigation
	Documents in diverse Indian languages	Inconsistent data extraction and unreliable verification	Integrate OCR and NLP models for multiple Indian languages.
	Low-quality images hinder OCR accuracy	Faulty or incomplete verification of documents	Prompt users to resubmit or process images for super-resolution
	High document volumes during peak periods	Delays or failures in verification	Leverage cloud based scalability & distributed processing tools.
	Document formats across India are diverse	Increase in processing time and prone to errors	Use flexible OCR models and AI-powered layout recognition

Dependencies

- Vision Model - Llama
- Infrastructure - GUPs for processing
- Low Image Sizes for faster processing

HIVE

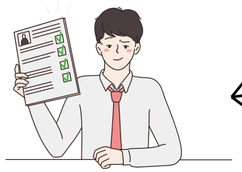
Impact and Benefits

SMART INDIA HACKATHON 2024

Impact on involved entities

Administrative Staff

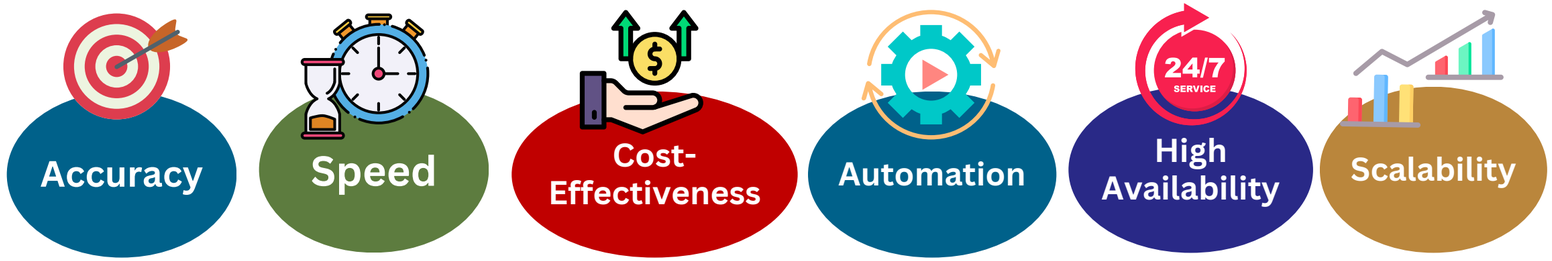
- Cost savings in manpower
- Reduction in Workload
- Improved Efficiency and Accuracy
- Faster Recruitment Cycles
- Data-Driven Decision Making
- Multilingual Document Support



Applicant

- Real -Time Feedback
- Reduced Waiting Period
- Enhanced User Experience

Benefits of DocLens to the Recruitment Process:



How Does DocLens Achieve 3 Sigma Accuracy

- Llama Vision already gives **95+ accuracy** this can be improved by training on custom document datasets using techniques like **transfer learning.**
- By using the **exact template** of the document we can get the proper position of desired fields thus increasing the accuracy.

How Does DocLens Achieve *Speed*

- **Feature Extraction using efficient algorithms** significantly reduces the time needed for categorizing and identifying different formats of documents
- To maximize speed, our solution **applies OCR only to essential regions of the document**, instead of the entire document.
- **Utilizing GPUs** for parallel processing and **cloud-based hardware** can significantly reduce document processing time
- Since we are using template we can predefine what language OCR we have to apply **reducing the need to add a layer of language detection.**

Scalability Plan

- Cloud Infrastructure
- Distributed Processing
- Microservices Architecture
- Caching Mechanisms
- Queueing Systems

HIVE

Research and References

SMART INDIA HACKATHON 2024

- **OCR(Optical Character Recognition :**
  - https://github.com/PaddlePaddle/PaddleOCR
  - https://openaccess.thecvf.com/content_cvpr_2016/papers/Shi_Robust_Scene_Text_CVPR_2016_paper.pdft
  - https://www.researchgate.net/publication/311609204_Robust_Scene_Text_Recognition_with_Automatic_Rectification

- **Feature Detection :**
  - https://medium.com/@deepanshut041/introduction-to-feature-detection-and-matching-65e27179885d
  - https://ai.stanford.edu/~syyeung/cvweb/tutorial2.html
  - https://docs.opencv.org/4.x/da/df5/tutorial_py_sift_intro.html
  - https://link.springer.com/chapter/10.1007/11744023_32

- **Image stitching:**
  - https://doi.org/10.1142/S0218001417540064
  - https://ouci.dntb.gov.ua/en/works/9QdjBOgl

Document

Click Here

The document offers a thorough explanation of the project's approach, detailing the process flow and methodology used in the implementation of the solution.

Video

Click Here

Explore the platform's potential by engaging with our unique approach and seamless user experience