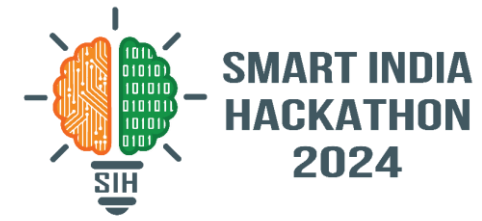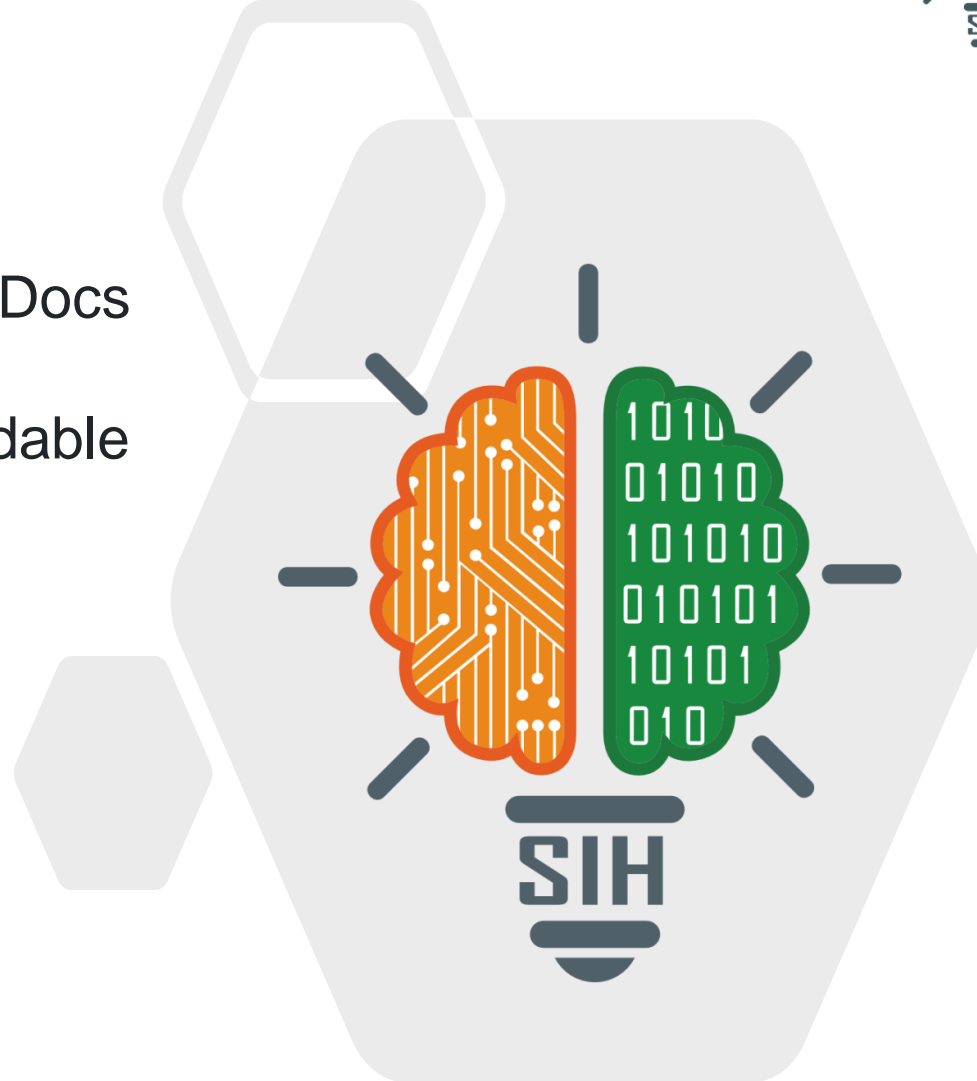# SMART INDIA HACKATHON 2024

- **Problem Statement ID –** SIH1669

- **Problem Statement Title-** Transformo Docs

  Application: Empowering Machine-Readable

  Document Management System.

- **Theme-** Smart Automation

- **PS Category-** Software
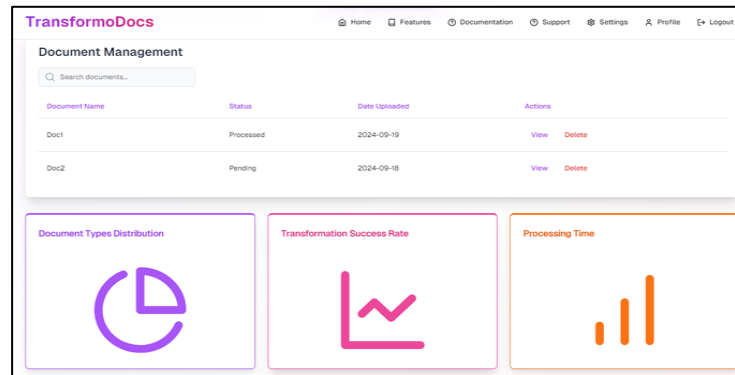
- **Team ID-**

- **Team Name –** Alert Aura Hive

IDEA TITLE

Alert Aura Hive

Proposed Solution

TransformoDocs is a document management application that uses advanced technologies to convert non-machine-readable documents into structured data. It leverages OCR, NLP, and machine learning to make documents searchable and easier to manage.
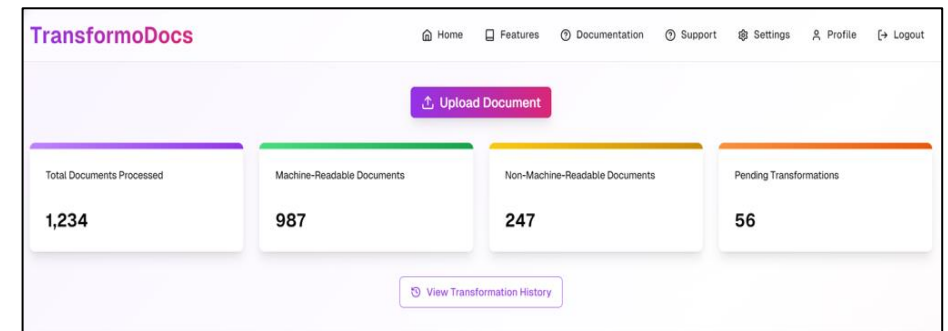
How it Addresses the problem

- **Format Restriction:** Prevents ingestion of unsupported documents.
- **Automated Conversion:** Uses OCR and NLP for seamless document conversion.
- **Data Accessibility:** Enables easy searching, analysis, and integration of extracted data.

Innovation & uniqueness of the solution

- **Comprehensive Support:** Handles various document formats and conversion methods.
- **Intelligent Extraction:** Accurately extracts data from complex documents.
- **Data Quality:** Ensures reliable and accurate data.
- **Seamless Integration:** Integrates with existing systems.
- **Scalability:** Handles large volumes of documents effectively.
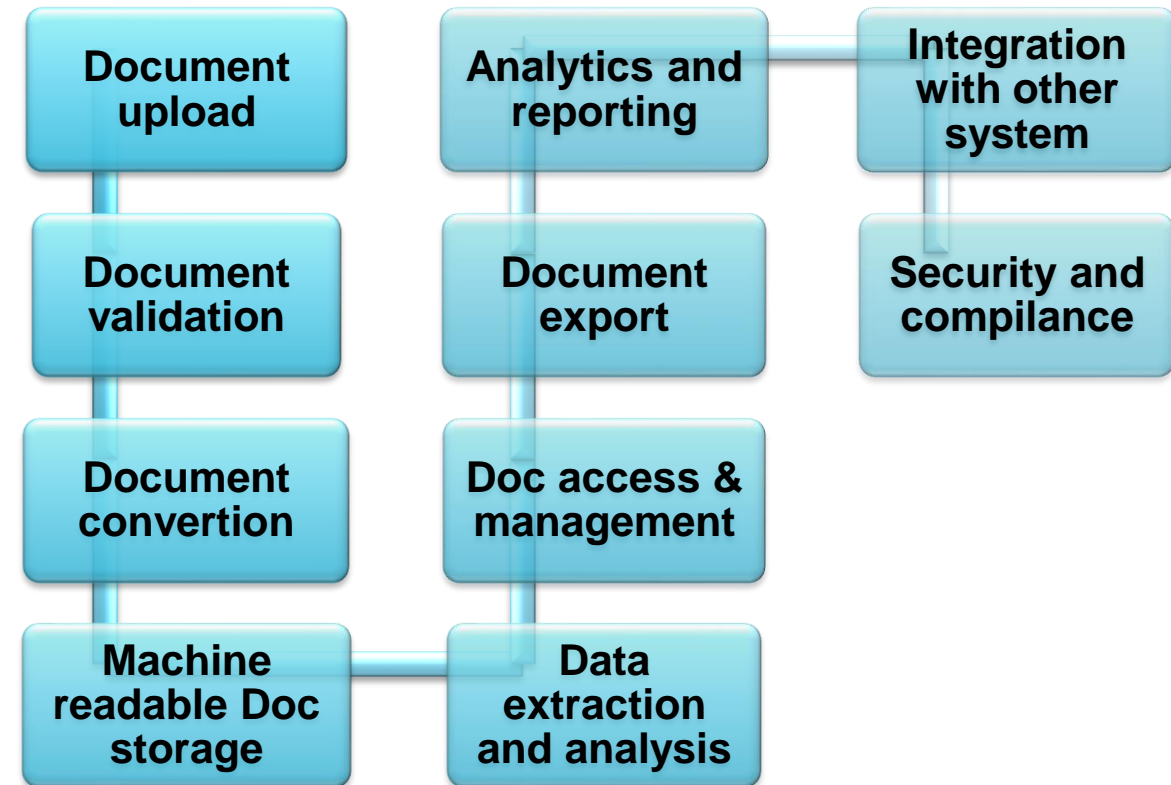
TECHNICAL APPROACH

Technologies

- **Programming:** Python (Django/Flask)
- **Database:** PostgreSQL/MongoDB
- **Conversion:** PyPDF2, PDFMiner.six, Python-docx, openpyxl, Tesseract OCR
- **Extraction:** NLTK, spaCy, Scikit-learn, TensorFlow, PyTorch
- **Metadata:** EXIF Data, Custom Parsing
- **Search:** Elasticsearch/Solr
- **UI:** HTML, CSS, JavaScript
- **Cloud:** AWS

Methodology

- Problem Definition and Requirements Gathering
- Technology Selection
- System Design and Architecture
- Data Preparation and Collection

Document upload → Document validation → Document convertion → Machine readable Doc storage → Data extraction and analysis → Doc access & management → Document export → Analytics and reporting → Integration with other system → Security and compilance

- Document Conversion and Data Extraction
- User Interface Development
- Integration with Other Systems
- Testing and Quality Assurance
- Deployment and Maintenance

FEASIBILITY AND VIABILITY

Feasibility of the idea

- **Technological advancements:** Recent progress in AI and NLP make it feasible.
- **Market demand:** Organizations seek to digitize and analyze documents.
- **Scalability and integration:** Can handle large volumes and integrate with existing systems.
- **Economic viability:** Potential benefits justify investment.

Risk or Challenge	Strategies to Overcome
Data Quality	Data cleaning and validation
Document Diversity	Preprocessing techniques
Privacy and Security	Strong security measures, compliance
Scalability	Cloud-based infrastructure, distributed systems
Integration Complexity	Integration frameworks, APIs
Model Limitations	Model evaluation, refinement
User Experience	Intuitive interface

IMPACT AND BENEFITS

Potential Impact on Target Audience

Businesses:
- **Efficiency:** Streamlined processes, reduced costs.
- **Decisions:** Data-driven decisions.
- **Compliance:** Adherence to regulations.
- **Collaboration:** Seamless sharing.
- **Environment:** Reduced paper and energy.

Individuals:
- **Simplified:** Easy document management.
- **Productivity:** Efficient searching, analyzing, sharing.
- **Accessibility:** Accommodation for disabilities.

Benefits of the Solution

Social:
- **Accessibility:** Accommodates disabilities.
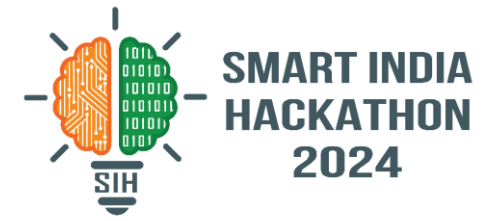- **Collaboration:** Seamless sharing.

Economic:
- **Cost Savings:** Reduced labor, efficiency, and improved decisions.
- **Productivity:** Streamlined document management.
- **Revenue:** Potential revenue from services.

Environmental:
- **Paper Reduction:** Digitization reduces paper usage and waste.
- **Energy Reduction:** Lower energy consumption.

- Details / Links of the reference and research work