

Fornova: An Intelligent Multi-Domain Document Processing and Analysis System Using AI

Shrihari Suresh Kasar¹, Rohit Pramod Khadangle², Utkarsha Bajirao Kakulte³, Ishwari Sandip Gamne⁴, Rohini D. Bhaskar⁵

¹²³⁴Student, Computer Engineering, Matoshri College of Engineering and Research Centre, Nashik, Maharashtra, India.

⁵Assistant Professor, Computer Engineering, Matoshri College of Engineering and Research Centre, Nashik, Maharashtra, India.

Abstract - This paper represents FORNOVA, an AI-powered document automation platform developed to efficiently process, organize, and analyze large volumes of unstructured data across multiple domains such as healthcare, finance, education, legal, industrial, and human resources. The system supports document uploads in formats like PDFs and images and extracts structured information using a hybrid approach combining rule-based parsing and AI techniques. Key data fields are presented with confidence scores to ensure reliability and transparency. It also provides validation checks and risk indicators for better analysis. A domain-specific dashboard displays performance metrics, while role-based access control enables document review and approval. The platform enhances efficiency, reduces manual effort, and supports data-driven decision-making for small and medium enterprises effectively.

Key Words: AI Document Processing, Document Automation, Information Extraction, Flask, MySQL, Data Analytics, OCR, Domain-Based Systems, Intelligent Systems, Data Visualization

1. INTRODUCTION

In the modern digital era, organizations across various industries generate and manage an enormous volume of documents daily, including invoices, medical records, legal contracts, academic certificates, industrial logs, and employee records. Most of this data exists in unstructured or semi-structured formats, making manual processing time-consuming, error-prone, and inefficient. Traditional methods rely heavily on human effort, increasing operational costs and reducing productivity.

To address these challenges, this project introduces FORNOVA, an AI-based document automation system designed to process and manage documents across domains such as healthcare, finance, education, legal, industrial, and human resources. The system uses a hybrid approach combining rule-based extraction and AI techniques to convert unstructured data into structured information. It extracts key fields with confidence scores,

provides validation and risk insights, and features a dashboard for performance monitoring, ensuring improved efficiency and decision-making.

2. LITERATURE SURVEY

- 1) Early document processing relied on Optical Character Recognition systems like Tesseract OCR, which convert scanned images into text but lack contextual understanding and structural interpretation capabilities.
- 2) Deep learning approaches using CNNs and transformer models like BERT improve document understanding by enabling feature extraction, semantic analysis, and accurate entity recognition from unstructured textual data.
- 3) Commercial platforms such as Google Document AI, Amazon Textract, and Microsoft Azure Form Recognizer offer automated extraction but often require templates and involve high costs.
- 4) Hybrid document processing systems combining rule-based techniques with AI models enhance accuracy and flexibility, effectively handling both structured formats like invoices and unstructured documents across multiple domains.

3. METHODOLOGY

The methodology of the FORNOVA system follows a structured, multi-stage pipeline to ensure efficient document processing and accurate data extraction. Initially, users upload documents in formats such as PDFs or images, which are preprocessed using OCR techniques to convert them into machine-readable text. The extracted text is then passed through a hybrid processing layer that combines rule-based parsing and AI-driven models to identify key fields like names, dates, and amounts. Each extracted field is assigned a confidence score to indicate reliability. The processed data is stored in a MySQL database, while analysis modules perform validation checks and detect potential risks. Finally, results are visualized through a domain-specific dashboard, and users can export structured data for further use, ensuring scalability and efficiency.

4. SYSTEM OVERVIEW

The proposed FORNOVA system follows a structured methodology to automate document processing and intelligent data extraction across multiple domains. Users upload documents in formats such as PDFs or images, which are first preprocessed using OCR techniques to extract machine-readable text. This text is then processed through a hybrid pipeline combining rule-based parsing and AI models to identify key fields such as names, dates, amounts, and domain-specific attributes.

4.1 Text Processing:

The extracted document content is cleaned, tokenized, and normalized using NLP techniques. Relevant entities are identified and converted into structured representations, enabling accurate extraction of important information.

4.2 Hybrid Data Extraction:

The system applies rule-based methods for structured patterns and AI models for unstructured data, ensuring flexibility and high accuracy. Confidence scores are assigned to each extracted field.

4.3 Analysis and Storage:

The processed data is stored in a MySQL database, where validation checks and risk analysis are performed. Results are visualized through dashboards, and structured data can be exported for further use.

4.4 SYSTEM ARCHITECTURE

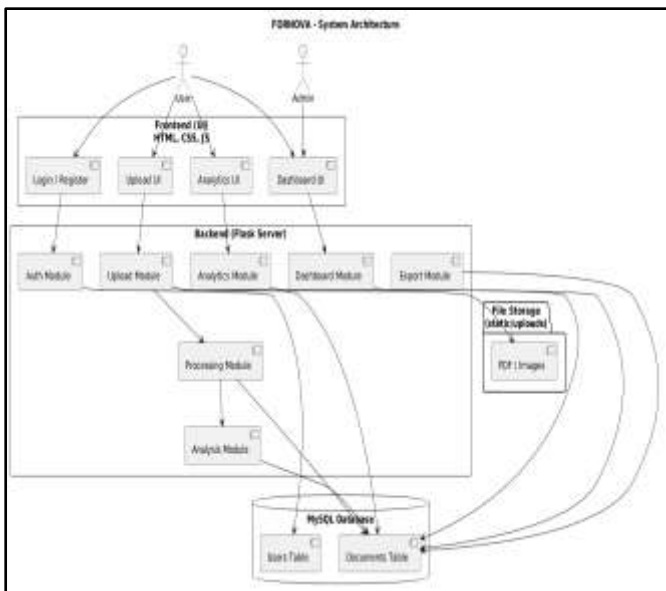


Fig No. 4.4: System Architecture

5. MODELING AND ANALYSIS:

The modeling and analysis of the FORNOVA system focus on designing an efficient framework for extracting, processing, and analyzing unstructured document data across multiple domains. The system is modeled using a modular architecture that integrates document input,

preprocessing, data extraction, analysis, and visualization components. Each module operates independently while maintaining seamless data flow through well-defined interfaces.

The analysis begins with document ingestion, where different formats such as PDFs and images are standardized through preprocessing and OCR techniques. The extracted text is then passed to a hybrid extraction model that combines rule-based algorithms and AI techniques for identifying structured fields. This hybrid model is analyzed for accuracy, flexibility, and performance across varied document types.

Performance analysis is conducted using metrics such as extraction accuracy, processing time, and confidence scores. The system also evaluates domain-wise efficiency and error rates. Analytical dashboards provide insights into trends, document volumes, and system performance. Overall, the modeling ensures scalability and robustness, while analysis validates the system's effectiveness in improving automation and decision-making.

The following describes the suggested FORNOVA workflow:

1. Document Input: Users upload documents such as invoices, medical records, or certificates in formats like PDF or images for processing.
2. Preprocessing (OCR & Cleaning): The system extracts text using OCR, followed by cleaning, normalization, and tokenization to prepare data for analysis.
3. Text Representation: The processed text is converted into structured formats and embeddings to capture semantic meaning and domain-specific attributes.
4. Hybrid Extraction Engine: A combination of rule-based parsing and AI models identifies key fields such as names, dates, amounts, and other relevant entities.
5. Validation Module: Extracted data is verified using logical rules and consistency checks to ensure correctness and reliability.
6. Confidence Scoring: Each extracted field is assigned a confidence score indicating the accuracy of extraction.
7. Analysis Module: The system performs risk detection, anomaly identification, and domain-specific insights for better decision-making.
8. Output & Storage: Final structured data is displayed on dashboards and securely stored in a database, with options to export in CSV format for further use.

6. RESULTS AND DISCUSSION

The FORNOVA system demonstrates effective performance in automating document processing and extracting structured data from unstructured inputs across multiple domains. The system was tested on various document types such as invoices, certificates, and records, and showed significant improvement in processing speed, accuracy, and consistency compared to manual methods. The hybrid extraction approach combining rule-based logic and AI techniques proved to be efficient in handling both structured and unstructured data. The dashboard and analytics module further enhanced usability by providing real-time insights into system performance and document trends.

Key Results and Discussion Points:

- **High Extraction Accuracy:** The system achieved reliable extraction of key fields such as names, dates, and amounts with strong confidence scores.
- **Reduced Processing Time:** Automated processing significantly decreased the time required compared to manual data entry.
- **Improved Efficiency:** The system minimized human effort and reduced errors in document handling workflows.
- **Multi-Domain Capability:** Successfully handled documents from healthcare, finance, education, legal, and HR domains.
- **Effective Visualization:** Dashboards provided clear insights into document volume, accuracy, and processing metrics.
- **Scalability:** The modular architecture supports scaling for larger datasets and multiple users.
- **Limitations:** Performance may vary with poor-quality or highly complex documents, indicating scope for further AI model enhancement.

6.1 OUTPUT

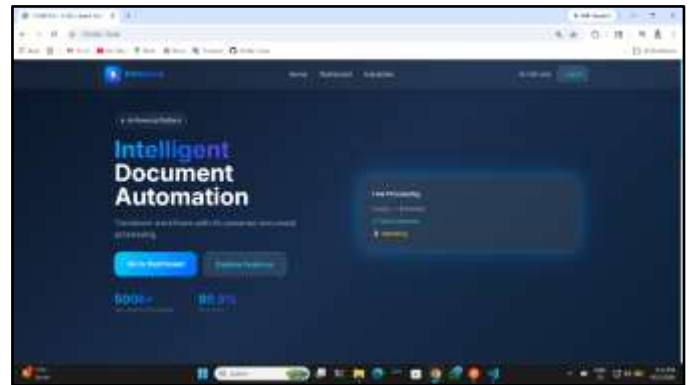


Fig 6.1 Home Page

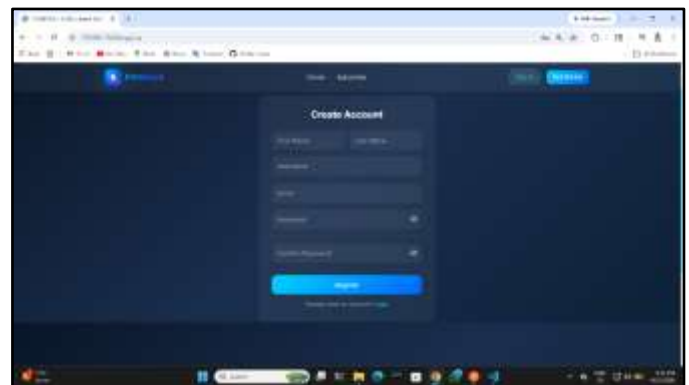


Fig 6.2 Registration Page

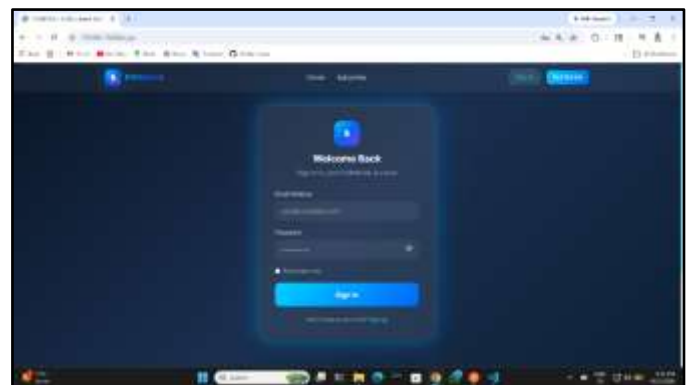


Fig 6.3 Login Page

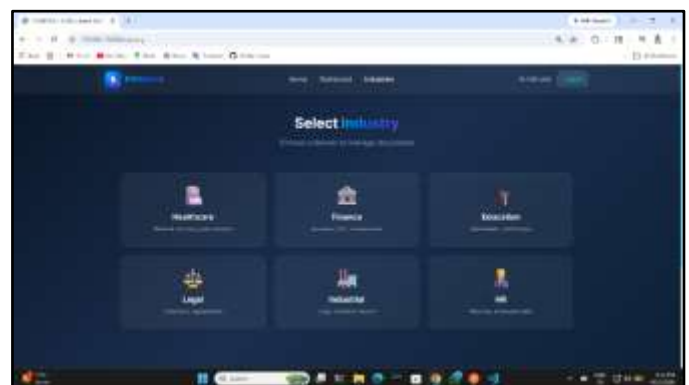


Fig 6.4 Domain Selection Page

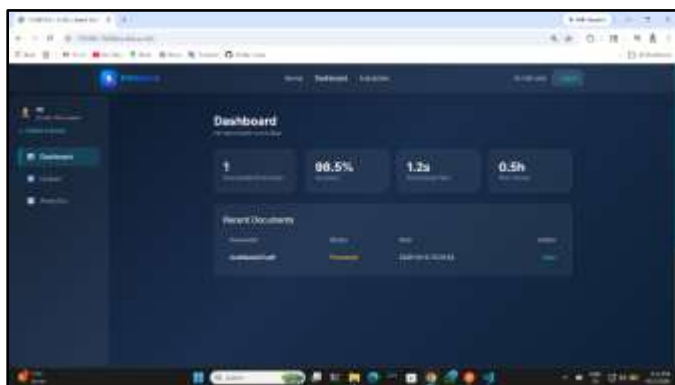


Fig 6.5 User Dashboard

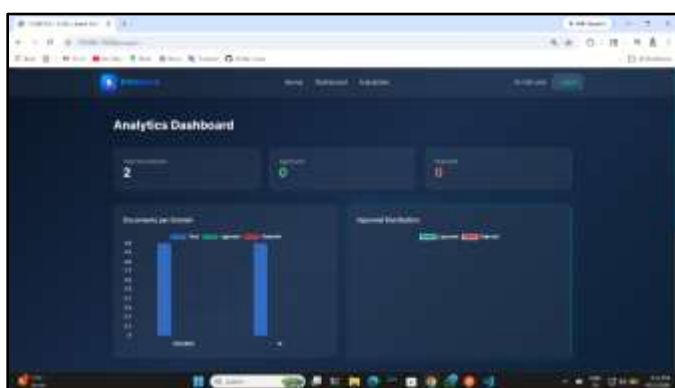


Fig 6.6 Analytics Dashboard

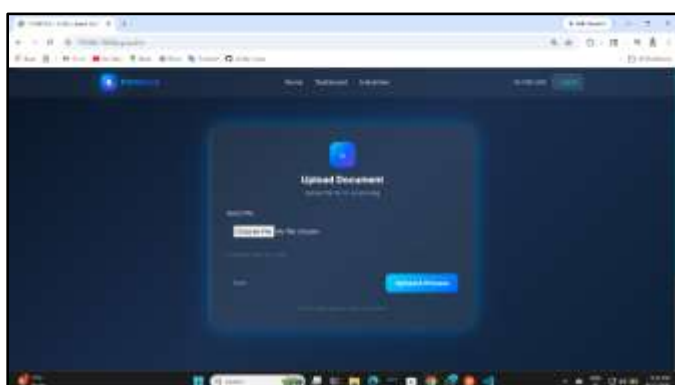


Fig 6.7 Document Upload Page

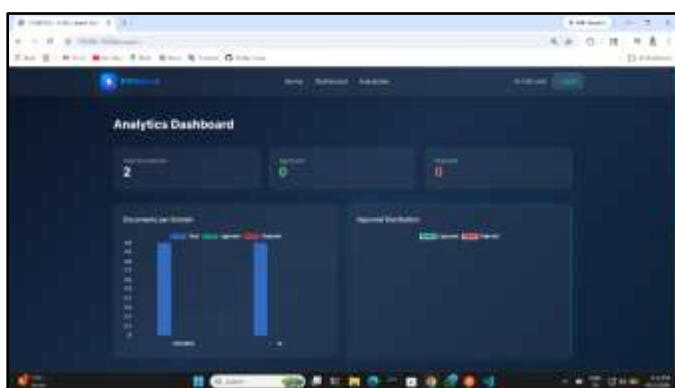


Fig 6.8 Document Overview Page

- Integration of deep learning models (CNN, NLP) for higher accuracy in document understanding
- Real-time document processing using APIs and cloud services
- Support for multilingual document extraction
- Advanced analytics with predictive insights and trend analysis
- Role-based access control with enhanced security mechanisms
- Mobile application for on-the-go document management
- Integration with enterprise systems such as ERP and CRM

8. CHALLENGES

1. Handling Unstructured and Diverse Formats: Documents vary widely in layout, quality, and structure, making it difficult to consistently extract accurate information across different domains.
2. OCR Limitations: Poor image quality, handwritten text, or complex formatting can reduce OCR accuracy, affecting overall data extraction performance.
3. Data Accuracy and Validation: Ensuring high precision in extracted fields while avoiding incorrect or mixed data remains a critical challenge.
4. Scalability and Performance: Processing large volumes of documents in real time while maintaining speed, accuracy, and system stability requires efficient resource management.

9. CONCLUSION

Fornova is a comprehensive document automation system that integrates user management, document processing, analytics, and export functionalities into a single platform. The system successfully automates the extraction and analysis of data from various document types, reducing manual workload and improving efficiency.

The modular architecture ensures flexibility and scalability, allowing easy integration of advanced AI models in the future. The system provides a user-friendly interface along with powerful backend processing, making it suitable for both technical and non-technical users.

10. REFERENCES

7. FUTURE ENHANCEMENTS

- [1] N. Gupta, A. Sharma, "Intelligent Document Processing using AI and NLP," *International Journal of Computer Applications*, vol. 182, no. 25, 2020.
- [2] R. Kumar, S. Verma, "OCR and Machine Learning Techniques for Document Automation," *IEEE Access*, vol. 8, pp. 14567–14576, 2020.
- [3] PlantUML Documentation:
<https://plantuml.com/>.
- [4] M. Singh, "Natural Language Processing for Automated Data Extraction," *Journal of AI Research*, vol. 7, pp. 55–68, 2019.
- [5] A. Joshi, "AI-driven Document Management Systems: A Survey," *International Conference on Computing and Communication Technologies*, 2021.
- [6] R. Smith, "An Overview of the Tesseract OCR Engine," *Proc. ICDAR*, pp. 629–633, 2007.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL*, 2019.
- [8] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [9] Google Document AI, Google Cloud, 2023.
- [10] Amazon Textract, Amazon Web Services, 2023.
- [11] Microsoft Azure Form Recognizer, Microsoft, 2023.
- [12] T. Kieninger and A. Dengel, "Applying the Tesseract OCR Engine for Document Analysis," *Document Analysis Systems*, 1998.
- [13] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Pearson, 2020.
- [14] I. Goodfellow et al., "Generative Adversarial Networks," *NeurIPS*, 2014.
- [15] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly, 2009.