Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

LOK RAKSHA : INDIAN Personally Identifiable Information System

Mihir Gosavi
Artificial Intelligence and Data Science
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

Vipulkumar Gupta
Artificial Intelligence and Data Science
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

Nisarga Lande
Artificial Intelligence and Data Science
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

Purva Khandagale
Artificial Intelligence and Data Science
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

Shipali Bansu
Artificial Intelligence and Data Science
A.C.Patil College of Engineering
Kharghar, Navi Mumbai

***Abstract*—The increasing adoption of digital platforms across India has led to a significant rise in the collection and processing of sensitive personal information. Data such as Aadhaar numbers, PAN details, and contact information is frequently handled across sectors, raising concerns about privacy, security, and regulatory compliance. Existing solutions for Personally Identifiable Information (PII) detection are largely designed for Western datasets and often fail to accommodate India's multilingual and structurally diverse documents. To address these challenges, this paper introduces LOK RAKSHA, a unified system for identifying and protecting Indian PII document. The proposed framework combines Optical Character Recognition (OCR), rule-based methods, and machine learning models to enable accurate and context-aware detection. It is specifically tailored to support regional languages and align with the Digital Personal Data Protection (DPDP) Act, 2023. The system also incorporates explainability features to improve transparency and usability in real-world applications.**

***Index Terms*—Personally Identifiable Information, Data Privacy, Optical Character Recognition, Natural Language Processing, Indian Data Protection, Compliance.**

I. Introduction

The rapid digitization of services in India, particularly across government, banking, and healthcare sectors, has significantly increased the volume of personal data being generated and stored. Systems such as Aadhaar and PAN have become central to identity verification, making sensitive information more frequently processed in digital environments. While this shift has improved accessibility and efficiency, it has also introduced new risks related to data exposure and misuse. Similar concerns regarding large-scale handling of Personally Identifiable Information (PII) have been highlighted in recent studies on privacy and data protection systems [2], [4]. During our initial observations, we found that many existing tools struggle when applied to Indian documents, especially those containing multiple languages or non-standard formats. For example, documents often include a mix of English and regional scripts, which reduces the effectiveness of traditional detection systems. Research on multilingual datasets further emphasizes the importance of region-specific models for accurate PII detection [5]. Another limitation is the lack of alignment with India's Digital Personal Data Protection (DPDP) Act, 2023. Most available solutions are designed around Western regulations and fail to address local compliance requirements. Existing frameworks for automated PII detection and redaction often focus on global standards, such as GDPR, without considering region-specific legal constraints [4], [6]. To address these challenges, we propose LOK RAKSHA, a system designed specifically for Indian use cases. Instead of relying on a single technique, the system combines OCR, rule-based detection, and hybrid AI model to improve accuracy across different document types. Prior work has shown that hybrid approaches integrating OCR and AI models can enhance detection performance in complex documents [1], [3]. The goal is not only to detect sensitive information but also to ensure that it is handled in a compliant and interpretable manner.

II. Related Work

Earlier approaches to PII detection largely depended on regular expressions and predefined rules. These methods work well for clearly structured data such as phone numbers or identification codes, but their performance drops when the input becomes unstructured or multilingual. In several studies, rule-based systems were shown to be effective only in controlled environments, with limited adaptability to real-world data variations [2]. To improve this, researchers have incorporated Optical Character Recognition (OCR) techniques

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

to process scanned documents and images. For instance, OCR combined with hybrid AI models has been used to detect sensitive information in visual data, including identity cards and signatures [1]. Similarly, hybrid systems integrating OCR with regex-based detection have demonstrated improved performance in extracting and masking PII from document images [3]. However, these approaches are often trained on limited datasets and may not generalize well to region-specific formats. Hybrid models that combine rule-based techniques with machine learning have been proposed to balance precision and flexibility. These systems can handle both structured and semi-structured data, but their effectiveness depends heavily on the availability of diverse and high-quality training data [2]. In multilingual environments like India, this remains a significant challenge. More recent work has explored the use of Named Entity Recognition (NER), to detect context-dependent personal information. These approaches improve the identification of implicit PII in unstructured text and are increasingly used in modern privacy-preserving systems [4]. Additionally, large language models have been applied for context-aware redaction tasks, enhancing semantic understanding and masking accuracy [6]. Despite these advancements, most existing systems are designed with a focus on Western datasets and regulatory frameworks. Datasets developed for specific languages, such as Korean or other non-English corpora, highlight the importance of localized training for improving detection accuracy [5]. However, solutions tailored specifically for Indian datasets characterized by multiple languages, mixed scripts, and unique identification formats remain limited. This gap directly motivates the development of LOK RAKSHA.

III. Proposed Framework

A. System Overview

LOK RAKSHA is designed as a modular and scalable framework aimed at protecting PII in Indian digital documents. The system consists of multiple interconnected components that process input files and generate redacted outputs along with risk assessments and explanatory insights.

B. Text Extraction with OCR

In real-world scenarios, documents are often available in formats such as scanned images, PDFs, and handwritten forms. To process such data, the framework incorporates an OCR module optimized for Indian languages. This module converts visual content into structured text, enabling further analysis in subsequent stages.

C. Hybrid PII Detection

The detection process combines multiple techniques to achieve high accuracy:

- Rule-based regular expressions are used to identify structured identifiers such as Aadhaar numbers, PAN details, passports, and phone numbers.
- NER models trained on Indian datasets are employed to detect entities such as names, email addresses, and locations.

This layered approach minimizes false positives while increasing recall across different document types and formats.

D. Compilance Engine

To ensure adherence to the DPDP Act, 2023 and UIDAI masking standards, the compliance module enforces legal rules on how PII should be masked or stored. It validates detection results and guides the redaction process to ensure lawful handling of sensitive data.

E. Risk Scoring and Explainability

Detected PII is evaluated by a risk scoring module that categorizes the sensitivity and potential exposure level. A Generative AI layer generates human-readable explanations and summarizations of detection results, improving transparency and user interpretability.

IV. Results and Discussion

The LOK RAKSHA framework was tested on a diverse collection of Indian digital documents, including scanned identity proofs, structured records, and multilingual datasets containing both structured and unstructured content. The evaluation focused on the system's ability to accurately detect and redact

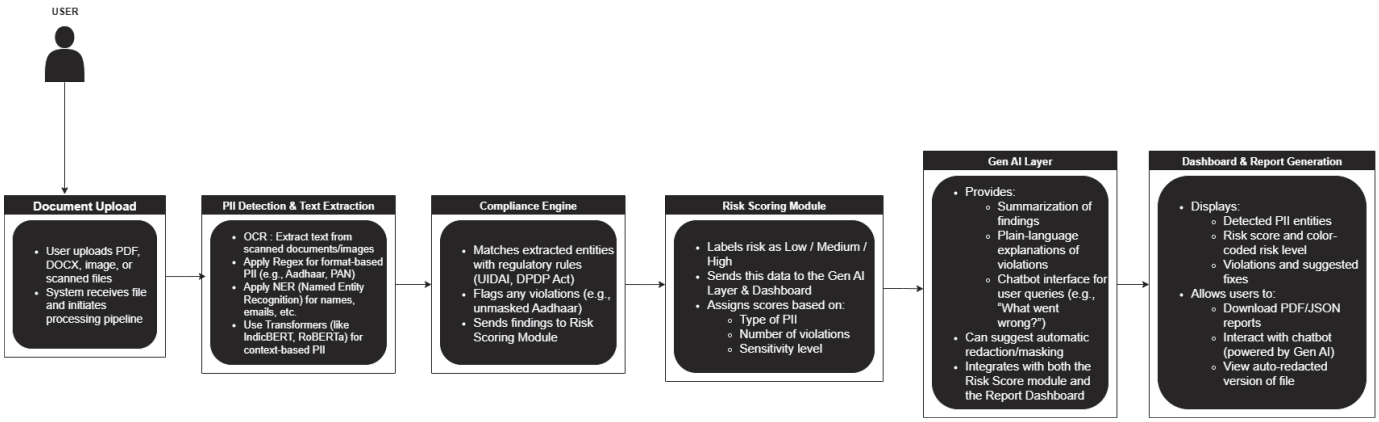

Fig. 1. Overall Architecture of the Proposed LOK RAKSHA System

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Indian-specific PII such as Aadhaar numbers, PAN details, phone numbers, names, and address-related information.

The hybrid detection approach demonstrated strong performance in identifying structured data using regex-based rules, while NER and tesseract models significantly improved detection in unstructured and context-dependent scenarios. Compared to purely rule-based systems, the proposed approach showed a noticeable reduction in false positives, especially in documents containing unrelated numerical data.

The OCR component successfully extracted text from image-based and multilingual documents, enabling the system to handle real-world data commonly found in government, banking, and healthcare sectors. The compliance module ensured that all detected PII was masked according to DPDP Act guidelines, strengthening data protection and regulatory alignment.

From a usability perspective, the inclusion of a risk scoring system allowed prioritization based on sensitivity levels, while the Generative AI module provided clear and understandable explanations. These features enhance the practical usability of the system for organizations managing large volumes of sensitive data.

Overall, the results indicate that LOK RAKSHA is an effective and scalable solution for PII protection in India, addressing the limitations of existing systems by incorporating localization, contextual understanding, and explainability.

TABLE I
COMPARISON OF PII DETECTION TECHNIQUES USED IN LOK RAKSHA

Technique	**PII Type Detected**	**Key Advantage**
Regex-Based Detection	Aadhaar, PAN, Passport, Phone Numbers	High precision for structured and format-based identifiers
Named Entity Recognition (NER)	Names, Addresses, Email IDs, Locations	Context-aware detection in unstructured text
OCR Layer	Scanned documents and images	Enables text extraction from multilingual and image-based records
Generative AI Module	Explanation and compliance reasoning	Enhances transparency and user interpretability

Table I highlights the hybrid detection strategy adopted in LOK RAKSHA, where rule-based, statistical, and hybrid AI approaches are combined to achieve accurate and context-aware PII detection across diverse Indian document formats.

TABLE II
FEATURE COMPARISON WITH EXISTING PII PROTECTION SYSTEMS

Feature	**Existing Systems**	**LOK RAKSHA**
Indian Government ID Support	Partial	Yes
Multilingual Document Handling	Limited	Yes
Context-Aware Detection	No	Yes
DPDP Act Compliance	No	Yes
Explainability via GenAI	No	Yes
Risk Scoring Mechanism	No	Yes

V. CONCLUSION

LOK RAKSHA provides a strong foundation for addressing India's growing data privacy challenges. Unlike conventional systems that rely solely on pattern matching, the proposed framework integrates regex-based methods, NER models, and advanced language models to effectively process multilingual and complex document structures. By embedding compliance with the DPDP Act, 2023 directly into the system, it demonstrates how technological solutions can align with regulatory requirements. This approach offers a reliable and scalable solution for organizations aiming to secure sensitive personal data in an evolving digital landscape.

VI. FUTURE WORK

Future enhancements will focus on incorporating multimodal AI techniques capable of processing text, images, and audio simultaneously to improve detection accuracy. Additional efforts will be directed toward expanding support for more Indian languages and dialects. Furthermore, the integration of privacy-preserving techniques such as federated learning will be explored to enhance model performance while ensuring data confidentiality.

REFERENCES

[1] O. Shaikh *et al.*, "Detection and Classification of Personally Identifiable Information in Images Using Artificial Intelligence," *TechRxiv*, May 2024.

[2] J. Jaikumar, Mohana, and P. Suresh, "Privacy-Preserving Personal Identifiable Information (PII) Label Detection Using Machine Learning," in *Proc. Int. Conf. Computing, Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–5, doi: 10.1109/ICCCNT56998.2023.10307924.

[3] D. K. Tunwal *et al.*, "A Hybrid OCR and Regex-based PII Detection and Masking Tool with Deepfake and Forgery Detection Capabilities," *Int. J. Advance Research and Innovative Ideas in Education*, vol. 11, no. 3, pp. 1460–1466, 2025.

[4] S. Asthana *et al.*, "Adaptive PII Mitigation Framework for Large Language Models," *arXiv preprint arXiv:2501.12465*, 2025.

[5] L. Fei *et al.*, "KDPII: A New Korean Dialogic Dataset for the Deidentification of Personally Identifiable Information," *IEEE Access*, vol. 12, pp. 135626–135641, 2024.

[6] P. Thetbanthad, B. Sathanarugsawait, and P. Praneetpolgrang, "Automated Redaction of Personally Identifiable Information on Drug Labels Using Optical Character Recognition and Large Language Models for Compliance with Thailand's Personal Data Protection Act," *Applied Sciences*, vol. 15, no. 9, p. 4923, 2025.

[7] G. Gambarelli, A. Gangemi, and R. Tripodi, "SPeDaC: A New Resource and Benchmark for Training Sensitive Personal Data Classifiers," *IEEE Access*, vol. 11, pp. 10864–10880, 2023.