2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) | 979-8-3315-0724-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICAISS61471.2025.11042182

Secure Data Masking Through Synthetic Data Generation using ML

Dr.M.SUGANTHI,
Assisstant Professor/IT
Kongunadu College of Engineering and Technology,Trichy,India
sugi.mp@gmail.com

K.TAMIZH SELVAN
Student,Department of IT
Kongunadu College of Engineering and Technology,Trichy,India
tamiltazil@gmail.com

A.RISEN BRIGHT
Student,Department of IT
Kongunadu College of Engineering and Technology,Trichy,India
risenbright5@gmail.com

S.LOGANATHAN
Student,Department of IT
Kongunadu College of Engineering and Technology,Trichy,India
loganathanloganathan775@gmail.com

R.SATHYA
Assisstant Professor/IT
Kongunadu College of Engineering and Technology,Trichy,India
sathyaphdkncet@gmail.com

Abstract— As the need for privacy and data security, organizations face the challenge of using sensitive data for development, testing, and analytics while ensuring compliance with data protection laws. This paper introduces a novel approach to secure data masking by leveraging Synthetic Data Generation through machine learning, particularly Generative Adversarial Networks (GANs). Instead of traditional masking or anonymization techniques for sensitive information such as Personally Identifiable Information (PII), this approach generates synthetic yet realistic data that retains the statistical properties and structure of the original dataset. The generated synthetic data ensures privacy compliance by eliminating real sensitive information while preserving essential attributes for effective analysis, testing, and development. This project addresses the need for robust synthetic data generation frameworks that can be seamlessly integrated into existing data pipelines, providing both scalability and security.

Keywords — Data Masking, Privacy Presentation, GANs, Document Parsing, Machine Learning.

I. INTRODUCTION

In today's digital landscape, safeguarding data privacy is a critical challenge, particularly when handling Personally Identifiable Information (PII) in identity and financial documents. Traditional data masking techniques often compromise data integrity, leading to reduced usability and increased security risks. To address this, we propose a machine learning (ML)-driven synthetic data generation system designed to enhance security by effectively masking sensitive information while preserving its analytical and operational value. Our approach provides organizations dealing with high volumes of sensitive data with a robust solution that strengthens privacy protections without sacrificing data utility. This study details the architecture, implementation, and evaluation of a secure document processing framework capable of extracting, masking, and generating synthetic alternatives for sensitive fields such as PAN and Aadhaar numbers. By integrating Generative Adversarial Networks (GANs) for secure synthetic data generation with Optical Character Recognition (OCR) for text extraction, the system ensures that generated data retains the statistical properties of the original dataset while preventing exposure of actual identities. This approach enhances data security, mitigates privacy risks, and supports regulatory compliance in sensitive data environments.

II. RELATED WORKS

In financial transactions, research highlights how GANs help address data imbalance and enhance fraud detection by generating synthetic transaction records that preserve data integrity and improve accuracy without exposing real transaction details [1]. By offering privacy- preserving synthetic data for AI-driven diagnostics, patient record analysis, and medical research, GANs have completely changed the healthcare industry. GANs provide synthetic EEG data that maintains analytical correctness while guaranteeing patient privacy in one important study that focuses on EEG brain signals [2].

By spotting unusual patterns in datasets, GANs enhance threat identification in cybersecurity [3]. In order to improve security models and preserve data privacy, they are also employed for data masking and adversarial training. Differentially private GANs have been shown to improve secure database query processing, guaranteeing data analysis without disclosing actual records. This is especially helpful in the healthcare and financial industries [4]. GANs are used in AI and machine learning to augment data, which solves data shortages and improves model performance [5]. Furthermore, GANs are the best at creating realistic datasets while striking a balance between data utility and privacy to reduce the danger of re-identification, according to a study assessing synthetic data generation techniques [6]. In order to improve healthcare analytics and maintain regulatory compliance, GANs are also utilized to generate synthetic time-series medical data that finds patterns in patient records for predictive modeling [7]. Studies show that, while preserving data privacy, GAN-generated data increases predictive accuracy, decreases overfitting, and improves generalization. In order to improve data production and representation beyond conventional methods and ensure superior privacy-preserving synthetic data generation, further studies introduce Masked Auto-Encoders with GANs [8].

Through secure AI frameworks like ATLAS, GANs also contribute to multi-party data exchange, allowing businesses to work together safely while maintaining privacy [9]. Additionally, GANs facilitate AI-driven analysis of medical data, improving data accessibility while safeguarding actual patient records [10].

III. PROPOSED APPROACH

The suggested strategy describes a methodical way to handle documents that contain sensitive data. The process is started by the user uploading a document via an interface. Following the upload, the system carries out document identification, recognizing and categorizing the file type to guarantee that it may be processed further. This stage is essential for expediting the text extraction and analysis stages that follow. Following detection, the algorithm moves on to preprocessing and text extraction. This entails standardizing the text for precise analysis, eliminating noise, and transforming the document's content into a machine-readable format. Sensitive data detection is the next stage after text extraction, during which sophisticated algorithms or artificial intelligence models search the document for private information like financial information, personal information, or other sensitive components.

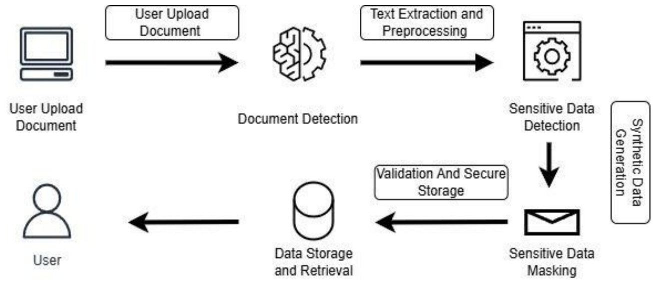

User Upload Document
Text Extraction and Preprocessing
User Upload Document
Document Detection
Sensitive Data Detection
Synthetic Data Generation
Validation And Secure Storage
User
Data Storage and Retrieval
Sensitive Data Masking

Figure 1: Architecture for proposed Approach

Fig.1 Demonstrates the safe data the masking system starts by ingesting documents in different formats, then extracting the text using OCR. NER is used to identify sensitive data, whereas synthetic counterparts produced by ML-based synthetic data generation (e.g., GANs) are used to mask sensitive data. Validation tests preserve compliance while guaranteeing data usefulness. Sensitive information is protected and privacy is guaranteed by the secure storage of the processed data.

A. Document Identification

A wide range of file formats, including text files (TXT, DOCX), PDFs, and image-based documents (JPG, PNG, and TIFF), are supported by the document ingestion system. Because different formats call for distinct handling methods, this step is essential for getting documents ready for additional processing. Users are given freedom by having the option to upload documents directly through the system or through an internet portal. Standardizing file structures, fixing discrepancies, and guaranteeing precise data extraction from each kind are all part of managing these diverse formats.

Managing the variety of file formats is one of the main obstacles in document intake. Because text-based files already include information in a machine-readable format, processing them is comparatively simple. However, various extraction techniques are needed for PDFs that contain either scanned images or embedded text. While parsing libraries such as PyPDF2 or PDF Miner can be used to handle text-based PDFs, optical character recognition (OCR) is necessary for scanned PDFs and photos in order to transform text from an image into a readable format. Furthermore, for improved accuracy, handwritten documents require sophisticated deep-learning OCR algorithms.

B. Document Ingestion

The system is built to support a wide range of document types, including text files, PDFs, and pictures, during the document ingestion stage.

Managing these various file formats and getting them ready for additional processing is the first step. Documents are gathered and sent to the following phases for analysis, either directly through the system or via an online interface. File format inconsistencies must be fixed during the ingestion process, and accurate data extraction from each format must be guaranteed.

A preprocessing procedure that includes scaling, de-noising, and transforming the file to a consistent format is necessary for image-based documents in order to guarantee the highest level of accuracy in the recovered text. Textual material, the basis for additional analysis and masking procedures, is extracted from PDFs and text documents using parsing libraries such as PyPDF2 or other specialized tools.

C. Smart Data Anonymization

To protect sensitive information, smart data anonymization dynamically masks data based on its type and usage. Unlike fixed masking methods, this approach adapts in real-time, ensuring privacy while keeping data useful. Machine learning helps detect and anonymize personal details effectively, improving security while maintaining accuracy in data processing.

D. OCR Processing

When it comes to extracting text from scanned photos or non-searchable PDFs, optical character recognition, or OCR, is essential. The system can process text-containing images and transform them into machine-readable representations using programs like PyTesseract. Pre-processing methods including picture binarization, noise reduction, and contrast enhancement are used to increase accuracy because raw OCR data may contain errors.

Post-processing techniques are also used, such as combining broken words or fixing misread characters. In order to ensure that the extracted data is as clean and useable as possible, the system makes use of algorithms created to increase OCR accuracy. This guarantees that the sensitive data identification phase receives accurate data from the OCR stage.

Because erroneous text might result in misclassification or missed detections in Named Entity Recognition (NER) and regex-based pattern matching, it is imperative to provide high-quality OCR output for the sensitive data identification phase. The system ensures that the extracted text is as clear, precise, and useable as feasible by incorporating cutting-edge OCR improvement techniques. This enables efficient sensitive data identification and privacy-preserving modifications in later processing stages.

E. Sensitive Data Identification

Together, Named Entity Recognition (NER) and regular expressions (regex) are used to identify sensitive data, including Personally Identifiable Information (PII), such as Aadhaar numbers, PAN cards, or other personal identifiers. While regex patterns aid in identifying common PII formats like phone numbers, emails, and dates, NER models trained on domain-specific data are used to detect things like names, addresses, and identification numbers.

The system makes sure that sensitive data is marked for masking by employing these strategies. Following identification, this data is separated to facilitate the subsequent masking stage. This procedure is essential for protecting data privacy and making sure that no private information is misused or revealed during the creation of synthetic data.

A thorough and dependable PII detection mechanism is achieved by the system by using both regex and NER. Once sensitive data is identified, the system makes sure that it is flagged for masking.

The detected information is then separated and categorized, allowing for efficient processing in the subsequent masking stage. This process is crucial to protect data privacy and prevent unauthorized exposure of personal details when generating synthetic data.

F. Data Masking Mechanism

The system applies a structured data masking process to safeguard sensitive information while preserving its usability. Optical Character Recognition (OCR) extracts text from documents, identifying crucial details such as PAN and Aadhaar numbers. Named Entity Recognition (NER) then detects personally identifiable information (PII), which is either masked or replaced with generated synthetic data. This method ensures that essential data remains usable for analysis without compromising privacy.

To further enhance security, Generative Adversarial Networks (GANs) generate synthetic data that closely resembles real datasets while eliminating actual sensitive records. This approach prevents unauthorized exposure while maintaining the statistical integrity of the data. Additionally, differential privacy techniques introduce minor modifications, reducing the chances of re-identification and strengthening overall data protection.

By permanently changing sensitive values before storing or distributing data, static masking ensures that unauthorized. Users cannot access the original data while it is at rest. Dynamic masking, on the other hand, implements masking rules in real time according to user roles and permissions keeping masked values visible to others and allowing only authorized people to examine the original data.

The system incorporates validation and compliance mechanisms to ensure data accuracy and regulatory adherence. Audit logs monitor all masking processes, while automated security checks enforce compliance with GDPR, HIPAA, and PCI-DSS standards. Finally, masked and synthetic data are securely stored in encrypted databases, ensuring restricted access while keeping the data functional for analytics and testing.

G. Synthetic Data Generation

Generative Adversarial Networks (GANs) play a crucial role in generating realistic synthetic data while preserving privacy. In this project, GANs are trained using sensitive data to learn its statistical properties without retaining actual identities. By leveraging adversarial training, where a generator creates synthetic data and a discriminator evaluates its authenticity, the model continuously improves, ensuring that generated data closely resembles real datasets without exposing confidential information.

To strengthen privacy, differentially private techniques such as Differentially Private Stochastic Gradient Descent (DP-SGD) are integrated, preventing GANs from memorizing and reproducing exact sensitive records. This ensures compliance with data protection regulations while maintaining high data utility for analytics and testing. Additionally, incorporating Variational Autoencoders (VAEs) alongside GANs enhances feature retention, enabling the generation of high-quality synthetic data that accurately represents real-world distributions.

To This refined explanation can replace existing sections discussing synthetic data generation and machine learning-based privacy protection in the document. It should be inserted where GANs are introduced, ensuring that the role of adversarial training and privacy-preserving techniques is highlighted.

H. Secure Validation, Compliance, and Data Storage

To ensure data security and compliance, AES-256 encryption protects stored data, while TLS encryption secures data during transmission. Audit logs track all access attempts, ensuring continuous monitoring and early threat detection. Regular risk assessments and automated compliance audits help maintain adherence to GDPR, HIPAA, and PCI-DSS standards.

Privacy is enhanced using Generative Adversarial Networks (GANs) with Differentially Private Stochastic Gradient Descent (DP-SGD) to prevent real data exposure. Named Entity Recognition (NER) identifies and masks sensitive fields, while k-anonymity, l-diversity, and t-closeness techniques ensure privacy protection. These measures safeguard data while preserving its usability for analysis and testing.

IV. EXPERIMENTAL RESULT

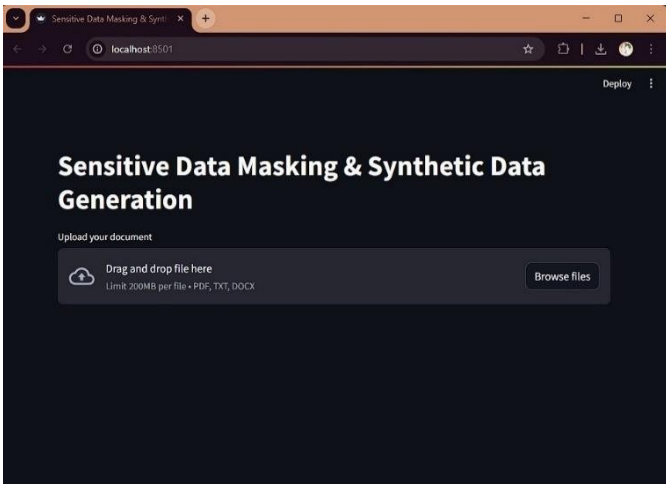
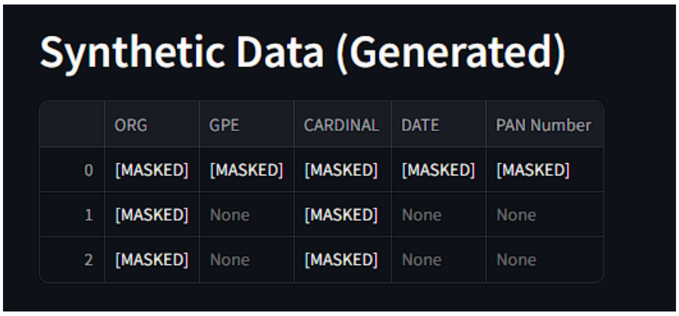

Sensitive Data Masking & Synt
localhost:8501
Deploy
Sensitive Data Masking & Synthetic Data Generation
Upload your document
Drag and drop file here
Limit 200MB per file • PDF, TXT, DOCX
Browse files

Figure 2. User Interface

For document processing, our streamlit-based user interface (UI) is made as shown in figure 2, to be easy to use, intuitive, and effective. It ensures wide compatibility by enabling users to submit TXT, DOCX, and PDF files. The technology preserves the document's structure while extracting and displaying text after it has been uploaded. OCR technology is used to transform scanned PDFs from images into legible text. Sensitive data, including phone numbers, email addresses, PAN numbers, and Aadhaar numbers, can be found using natural language processing (NLP) and pattern matching. For privacy protection, users can then conceal or swap out identified data with artificial counterparts. The solution guarantees safe storage and adherence to India's data protection regulations as well as the GDPR. Confidentiality is maintained while processed documents are still useful. streamlit and NLP work together to offer a solution that is both interactive and privacy-focused.



Synthetic Data (Generated)

	ORG	GPE	CARDINAL	DATE	PAN Number
0	[MASKED]	[MASKED]	[MASKED]	[MASKED]	[MASKED]
1	[MASKED]	None	[MASKED]	None	None
2	[MASKED]	None	[MASKED]	None	None

Figure 3. Masked Data

Our project's synthetic data table is made to display masked sensitive data in an organized manner, protecting privacy while preserving data usefulness. The table, which was constructed with Pandas and streamlit, dynamically shows fake numbers created to take the place of real sensitive data. Every detected category—including phone numbers, email addresses, PAN numbers, and Aadhaar numbers—is masked, and tabular synthetic equivalents are produced.

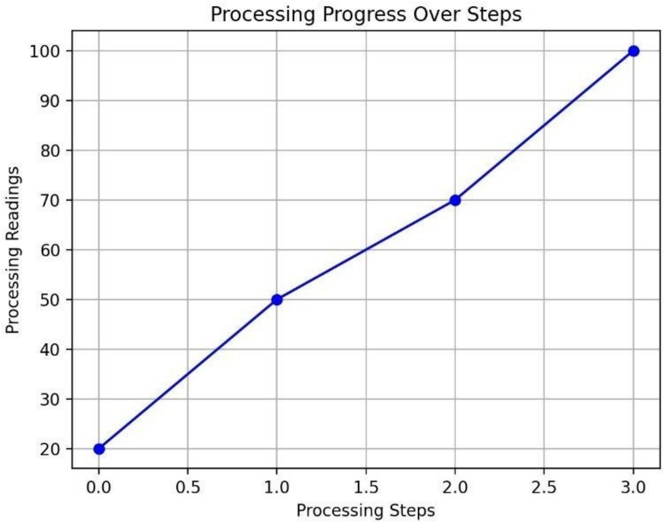


Figure 4. Synthetic data generation progress

The graph figure 4 shows the relationship between processing steps (x-axis) and processing readings (y-axis). The data points, marked in blue, indicate a steady and linear increase in readings as the number of processing steps advances. This suggests that as the system progresses through different stages, it achieves higher processing efficiency or accuracy. The consistent upward trend implies a structured workflow where each step contributes significantly to the overall progress. The visualization is useful for monitoring step-wise improvements in data processing, ensuring that performance increases proportionally with each phase.

V. CONCLUSION

The Secure Data Masking through Synthetic Data Generation Using ML effort offers a creative approach to safeguard sensitive data while preserving data usefulness. By incorporating Generative Adversarial Networks (GANs), the system generates realistic synthetic data that mimics the original dataset while hiding any real-world identifiers. Because it ensures compliance with privacy rules like GDPR and HIPAA, it is suitable for industries managing sensitive or personal data. By employing OCR algorithms for text extraction and Named Entity Recognition (NER) to identify personally identifiable information (PII), the system is able to accurately detect and conceal sensitive data. The masking techniques used anonymize important information such as names, phone numbers, and email addresses, but the data retains its structure and usefulness for further activities like testing and analysis. This achieves a balance between security and data utility. In conclusion, this research provides a comprehensive and secure approach to data masking and synthetic data creation. By addressing both data security and usability, it offers a practical solution for industries like healthcare, finance, and education where protecting sensitive data is crucial. Businesses can handle and share data in a secure, privacy- compliant way without compromising confidentiality or legal obligations by implementing this system.

VI. REFERENCES

[1]. Gaddam, Narayana. "AI-Powered Data Masking for Privacy-Preserving Cloud Data Sharing." International Journal of Advanced Research in Cloud Computing 5, no. 2 (2024): 12-22.

[2] Damer, Naser, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. "Privacy-friendly synthetic data for the development of face morphing attack detectors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1606-1617. 2022.

[3] Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger. "Privacy and synthetic datasets." *Stan. Tech. L. Rev.* 22 (2019): 1.

[4] Lu, Yingzhou, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. "Machine learning for synthetic data generation: a review." *arXiv preprint arXiv:2302.04062* (2023).

[5] Uddin, Md Ashraf, Md Naimul Ahsan, and Mrinmoy Das. "A Comparative Study on Various ML Models Using Synthetic Data for Privacy Preservation." In *2024 4th Interdisciplinary Conference on Electrics and Computer (INTCEC)*, pp. 1-6. IEEE, 2024.

[6] Tekchandani, Prakash, Ashok Kumar Das, and Neeraj Kumar. "Large-scale secure model learning and inference using synthetic data for IoT-based big data analytics." *Computers and Electrical Engineering* 119 (2024): 109565.

[7] Soufleri, Efstathia, Gobinda Saha, and Kaushik Roy. "Synthetic dataset generation for privacy-preserving machine learning." *arXiv preprint arXiv:2210.03205* (2022).

[8] Lewis, Tyler. "Extending Synthetic Data and Data Masking Procedures Using Information Theory." Master's thesis, Purdue University, 2023.

[9] Das, Hari Prasanna, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-Vincentelli, and Costas J. Spanos. "Conditional synthetic data generation for robust machine learning applications with limited pandemic data." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 11792-11800. 2022.

[10] Innocent, Ede Kelechukwu. "Enhancing Data Security in Healthcare with Synthetic Data Generation: An Autoencoder and Variational Autoencoder Approach." Master's thesis, Oslo Metropolitan University, 2024.