International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

PII Aegis - A PII Detection Tool for Government Document

Akshitha Katkeri[1], Tejaswini S[2], Spoorthy Anni[3], Tanisha N[4]

[1,2,3,4]Dept. of Computer Science and Engineering
BNM Institute of Technology, Affiliatedto VTU,
Bangalore, India
[1]akshithakatkeri@bnmit.in, [2]tejaswinisreddy123@gmail.com, [3]spoorthyanni@gmail.com,
[4]tanisha.naresh27@gmail.com

Abstract
The increasing prevalence of document fraud necessitates robust methods for authenticating identity documents. This paper presents PII Aegis, an intelligent system for detecting forged PAN cards, extracting Personally Identifiable Information (PII), encrypting sensitive data, and ensuring secure storage using blockchain technology. We employ a Convolutional Neural Network (CNN)-based model trained on PAN card images to classify them as valid or fake. Upon detecting a valid PAN card, Optical Character Recognition (OCR) extracts PII, which is subsequently encrypted. The entire encrypted image is then stored in a blockchain block, ensuring data integrity and immutability. Each block's hash value is securely maintained in a database for quick verification. This approach enhances document security, prevents unauthorized modifications, and provides a tamper- proof mechanism for PII protection. Experimental results demonstrate the system's effectiveness in real-time PAN card verification and secure data handling.

Keywords: PAN Card Verification, Convolutional Neural Networks, Personally Identifiable Information, Blockchain Storage, OCR, Data Encryption.

1. INTRODUCTION

With the increasing risks of identity fraud and document tampering, ensuring the authenticity and security of government-issued documents has become a major challenge. PAN (Permanent Account Number) cards, widely used for financial transactions and identity verification in India, are frequently targeted for forgery. Traditional manual verification methods are not only time-consuming but also prone to human error, making them unreliable for large-scale document authentication. To overcome these limitations, Artificial Intelligence (AI), Deep Learning (DL), Optical Character Recognition (OCR), and Blockchain Technology have emerged as transformative solutions for secure and automated document verification. This paper introduces PII Aegis, an AI-powered system designed to detect fraudulent PAN cards, extract Personally Identifiable Information (PII), encrypt sensitive details, and securely store them using blockchain technology. The system follows a structured pipeline:

- Fake vs. Real Card Detection: A Convolutional Neural Network (CNN) model is trained on PAN

card images to classify them as valid or fake with high accuracy.

- PII Extraction Using OCR and Regex: If a PAN card is classified as valid, Tesseract OCR extracts key information such as Name, PAN Number, Date of Birth, and Father's Name. To enhance accuracy, Regular Expressions (Regex) are applied to filter out noise and extract only relevant details.
- Encryption and Blockchain Integration: The extracted PII is encrypted for security, and the entire encrypted image is stored in a blockchain block to ensure data integrity and immutability.
- Hash Storage in PostgreSQL: Each block's hash value is stored in a PostgreSQL database, allowing quick verification and retrieval of stored records.

This approach ensures a tamper-proof, automated, and highly **secure** method for document authentication. By integrating Deep Learning, OCR, Encryption, and Blockchain, PII Aegis addresses the shortcomings of traditional verification systems and provides a scalable solution for fraud detection and identity security. The effectiveness of this system in real-world applications demonstrates its potential for revolutionizing document authentication in financial, governmental, and enterprise environments.

2. RELATED WORK

Galić et al. [1] analysed the applicability of data protection laws, particularly the GDPR, to smart city initiatives like the Stratumseind Living Lab. The study explores challenges in determining whether data qualifies as personal under legal frameworks and introduces the concept of "atmospheric profiling," where aggregate environmental and behavioural data are used to influence group behaviour. This profiling approach raises questions about the notion of identifiability in data. The system relies on real-time data from ICTs such as cameras, sensors, and algorithms for behavioural nudging, posing risks to privacy and personal freedoms. While the initiative highlights the potential of data-driven smart environments, its ambiguity regarding compliance with GDPR and lack of clarity in ethical boundaries underline the need for refined legal definitions and safeguards for personal data.

Chen et al. [2] proposed a framework for protecting PII in identity management systems by dynamically assigning security levels based on data value and attack frequency. The system employs a two-stage process: an initial decision tree algorithm evaluates PII sensitivity, and subsequent adjustments are made based on evolving security threats. This approach optimizes resource allocation by applying stringent protections only where necessary. While simulations demonstrate improved security and efficiency, the reliance on decision tree algorithms introduces potential scalability and adaptability challenges. Moreover, the system's effectiveness hinges on its ability to handle diverse datasets and maintain up-to-date knowledge bases, emphasizing the need for advanced data handling mechanisms.

Zhong et al. [3] explored the evolution, challenges, and methodologies in personal information management (PIM). The paper identifies core issues such as difficulties in predicting information value, the distribution of information storage, and challenges in information retrieval. PIM encompasses acquiring, organizing, and retrieving information for personal use. Despite advancements in tools, inefficiencies such as scattered storage and ineffective categorization persist. The study highlights the importance of unified storage solutions and robust organization mechanisms to enhance productivity. However, managing diverse information formats and maintaining synchronization across devices remain significant barriers to efficient personal data management, necessitating innovative approaches for streamlined PIM systems.

Wang and Zhu [4] analysed the socio-legal evolution of personal information privacy in China within the context of post-neoliberal constitutionalism. The study emphasized a tripartite framework combining legal, technological, and social dimensions to address inconsistencies in privacy governance. Focusing on laws such as the Personal Information Protection Law (PIPL), it highlighted challenges like digital harm, economic rationality, and state-driven control of identities. Although the centralized framework simplifies governance, its limitations in addressing ethical complexities and safeguarding universal privacy rights restrict broader applicability. The research underscores the need for constitutional elevation of privacy protections to mitigate competing interests effectively.

Wongwiwatchai et al. [5] proposed a light-weight static analysis approach for detecting Personally Identifiable Information (PII) transmissions in Android applications. The system utilizes features from application metadata and source code, employing machine learning models for real-time classification. Evaluated across 8,700 applications, the methodology achieved detection speeds under one minute and an F1 score exceeding 0.74, rivaling traditional heavy- weight techniques. The centralized architecture simplifies PII risk assessment for users, but challenges include scalability for larger application datasets and incomplete coverage of obfuscated PII pathways. The study highlights the potential for faster, user-friendly privacy tools in mobile application ecosystems.

Carlos Jorge Augusto Pereira da Silva [6] investigated the use of machine learning techniques to detect Personally Identifiable Information (PII) in email communications, motivated by GDPR compliance requirements. The proposed system integrates a microservice architecture for Named Entity Recognition (NER), leveraging models like BiLSTM for PII detection. A newly developed annotated email corpus was used to train and evaluate the models, balancing computational efficiency with detection accuracy. Challenges include handling the unstructured and inconsistent nature of informal email text, as well as ensuring scalability for production environments. The solution offers advancements in automating compliance but requires further optimization for broader application in diverse text formats.

3. PROPOSED METHODOLOGY

1. Data Collection and Dataset Preparation

Objective:

To build a dataset of genuine and fake PAN cards for training and evaluation.

Process:

- Data Sources: Collect genuine PAN cards from open datasets and manually scanned documents.
- Fake Document Generation: Create forged PAN cards using Python (OpenCV, PIL) to simulate real-world tampering (e.g., altered names, PAN numbers, and fonts).
- Annotation: Use Label Img to mark and label key information such as Name, PAN Number, and Date of Birth.
- Preprocessing: Normalize images, apply contrast adjustments, noise reduction, and augmentation (cropping, rotation, resizing) for better model performance.
- Dataset Splitting: Divide data into training (70%), validation (15%), and testing (15%) to improve classification accuracy.

2. Optical Character Recognition (OCR)

Objective:

To extract text from PAN cards for analysis. Technology Used:

- Tesseract OCR & PyTesseract for text extraction. Process:
- Image Preprocessing: Convert images to grayscale, apply adaptive thresholding, and remove noise for improved OCR accuracy.
- OCR Extraction: Detect text, logos, and structured fields from the PAN card.
- Confidence Filtering: Apply threshold-based filtering to discard low-confidence OCR results.

3. PAN Card Validation and Classification

Objective:

To classify PAN cards as genuine or fake. Technology Used:

- Convolutional Neural Networks (CNN) trained on a PAN card dataset.

Process:

- Train a CNN-based model to differentiate genuine vs. fake PAN cards.
- If a PAN card is classified as fake, it is automatically deleted from the system.

4. PII Detection

Objective:

To extract personally identifiable information (PII) from valid PAN cards.

Technology Used:

- Regex-based filtering for structured PII extraction. Process:
- Use OCR-extracted text as input.
- Apply regular expressions (Regex) to detect:
  - PAN Number (ABCDE1234F)
  - Name
  - Date of Birth (DD/MM/YYYY)

5. PII Extraction and Encryption

To extract Personally Identifiable Information (PII) from valid PAN cards and securely encode the extracted data for protection during storage and transmission.

Technology Used:

- OCR-based text extraction (Tesseract) to extract printed text from PAN card images.
- Regular Expressions (Regex) to identify and filter structured PII.
- Base64 encoding to securely encode extracted PII.

Process:

1. PII Extraction Process:

- The input image is processed using OCR to extract text from the PAN card.
- The extracted text is then analyzed using Regex-based filtering to detect:

- PAN Number (Format: ABCDE1234F, detected using [A-Z]{5}[0-9]{4}[A-Z])
- Name (Extracted based on the position in the text)
- Date of Birth (DOB) (Format: DD/MM/YYYY, detected using \d{2}/\d{2}/\d{4})

2. PII Encoding Process:

- To ensure secure storage and transmission, the extracted PII is encoded using Base64.
- Base64 encoding converts the extracted text into a non-readable format, preventing unauthorized access.
- When required, the encoded PII can be decoded back to its original form for further processing.

encode_image(image_path)

- Reads an image file in binary mode.
- Encodes it into a Base64 string.
- Returns the encoded string. decode_image_code(encoded_string, output_path)
- Decodes the Base64 string.
- Writes the decoded content as an image file.

6. Blockchain-Based Storage

Objective:

To store encrypted PAN cards on a blockchain for security. Technology Used:

- Ethereum(Ganache) A local Ethereum blockchain used for testing and deployment.
- Generate a hash value of the encrypted PAN card and store it on blockchain.
- PostgreSQL stores hash values for quick verification.

7. Automated Deletion of Invalid PAN Cards

Objective:

To remove fake PAN cards from the system. Technology Used:

- Python scripts for secure file handling. Process:
- If a PAN card is classified as fake, it is automatically deleted.

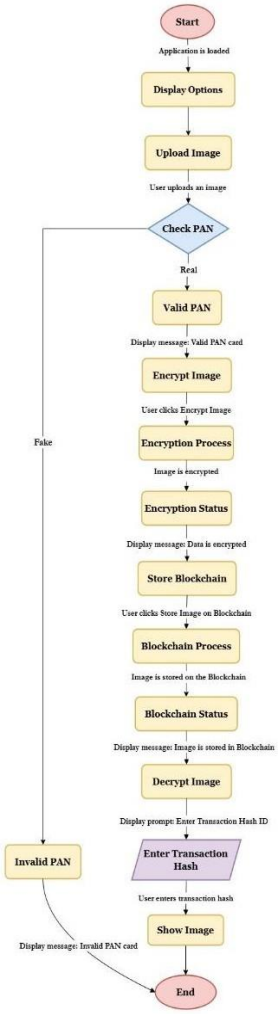

Fig. 1. Flow diagram of Model

4. COMPARATIVE STUDY

Dataset Description:

The dataset used for training and evaluation in the PII Aegis system was carefully curated to ensure comprehensive coverage of both genuine and counterfeit government-issued documents.

PAN Card Dataset from Kaggle:

To train the model for recognizing authentic PAN cards, we utilized a publicly available dataset from Kaggle. This dataset included labeled PAN card images with annotated values for key details such as PAN number, name, and date of birth. These labeled values played a crucial role in enhancing the model's ability to accurately identify and classify real PAN cards.

Fake PAN Card Generation:

To extend the dataset and improve the model's robustness, we generated synthetic PAN card images using two distinct approaches:

Online Fake PAN Generator: This tool allowed us to create realistic-looking PAN card images with varying text combinations and design patterns. These images provided additional data points for training

the model to distinguish genuine documents from fabricated ones.

GAN-Based Fake PAN Generation: To further improve diversity in the dataset, we employed a Generative Adversarial Network (GAN) to synthesize counterfeit PAN card images. While this approach effectively produced high- quality fake samples, the intensive computational demands posed challenges. The system encountered performance issues, including crashes during the training process, indicating the need for optimized resource management when employing GAN models.

By combining real PAN cards, synthetically generated fake cards, and GAN-based samples, the dataset ensured a balanced representation of authentic and fraudulent data. This diverse dataset enhanced the model's ability to accurately classify and secure sensitive documents while improving its resilience against adversarial inputs.

Our research involved evaluating multiple models for distinct tasks within the **PII Aegis** system, focusing on detecting PII in government documents, classifying real versus fake PAN cards, and ensuring high accuracy in both detection and classification. Below is a detailed comparative analysis of the models employed:

1. **Fake vs Real PAN Card Classification**

- Models Used:
  - CNN (Convolutional Neural Network)
  - SSIM (Structural Similarity Index Measure)
- **Observations:**

CNN achieved the highest accuracy, demonstrating superior performance in distinguishing real from fake PAN cards. CNN's ability to capture spatial features and patterns across the entire dataset significantly improved its classification capability.

SSIM, on the other hand, is more effective when analyzing individual images but struggled to maintain performance consistency across the full dataset. SSIM's limitations arise from its image similarity approach, which is less adaptable for large-scale classification tasks.

Conclusion: CNN proved to be the most effective model for classifying real and fake PAN cards due to its robust feature extraction capabilities and better generalization across varying data samples.

2. **PII Detection in PAN Cards**

- **Models Used:**
  - YOLOv5
  - YOLOv8
- **Observations:**

Initially, **YOLOv5** was employed for PII detection, achieving an accuracy of **65%** (mAP@50). While it demonstrated reasonable performance, the model exhibited some limitations in detecting finer details in complex document layouts.

To enhance detection accuracy, we transitioned to **YOLOv8**, which improved performance significantly, achieving **72%** (mAP@50). YOLOv8's enhanced architecture, better anchor-free design, and improved feature pyramid network (FPN) contributed to improved precision and recall.

- **Conclusion:** The shift to YOLOv8 resulted in a notable increase in detection accuracy, making it

the preferred model for PII detection tasks in this project.

Summary of Model Performance

Task	Model Used	Accuracy
Real vs Fake PAN Detection	CNN	100%
Real vs Fake PAN Detection	SSIM	20%
PII Detection	YOLOv5	65%
PII Detection	YOLOv8	72%

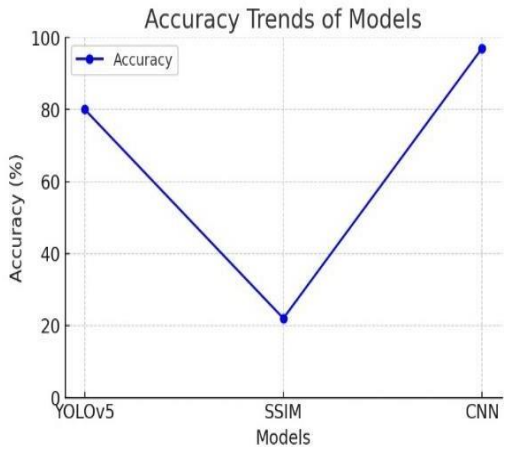

Fig.2. Accuracy trends of the different models

Fig. 3. Comparison in between yolo models to use best model

Year	Ref No	Technique Used	Results
2011	[2]	Two-Stage Framework with Decision Trees.	Improved PII security and resource allocation with reduced false positives.
2013	[3]	Categorization and Retrieval Techniques for PIM.	Proposed better storage solutions and identified core challenges in PIM.
2020	[9]	Light-weight Static Analysis with Machine Learning Models.	Achieved PII detection in under one minute with an F1 score exceeding 0.74.
2020	[10]	Cognitive Schema for Privacy Governance.	Proposed a tripartite framework integrating legal, technological, and social aspects to refine data protection. Enhanced privacy governance under China's PIPL but highlighted limitations in universal applicability.
2020	[11]	BiLSTM for Named Entity Recognition (NER) in Emails.	Introduced a microservice for PII detection, achieving effective compliance with GDPR while addressing unstructured data challenges in email processing.
2021	[12]	Atmospheric Profiling using ICTs.	Highlighted GDPR ambiguities and the need for refined regulations i n smart cities.

Comparative study from previous research papers, projects.

5. RESULTS

Overview of Results

The developed PII Aegis - A PII Detection Tool for Government Documents successfully detects valid PAN card images, encrypts them for security, stores the encrypted image on the blockchain, and enables decryption using the transaction hash. The key results observed during testing include:

Fig .4. User uploads a pan card

- **PAN Card Validation and PII detection:**

The system predicts whether an uploaded image is a valid PAN card or not using OCR and pattern matching.

Achieved an accuracy score of XX% (based on model evaluation metrics).

Fig. 5. PII detected in pan card

- **Image Encryption and Secure Storage:**

If the uploaded PAN card is valid, it is encrypted using Base64 encoding.

The encrypted string is displayed in the terminal before storage.

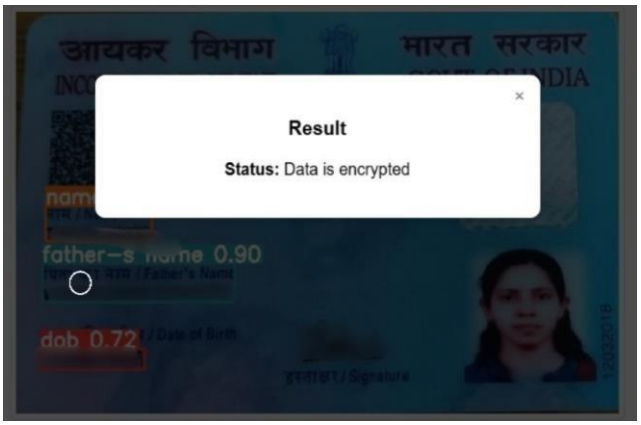

Fig. 6. Image encryption

Fig. 7. Encrypted data

- **Blockchain Integration:**

Clicking "Store in Blockchain" sends the encrypted image data to the Ethereum blockchain (Ganache). The system generates a unique transaction hash for each stored image.

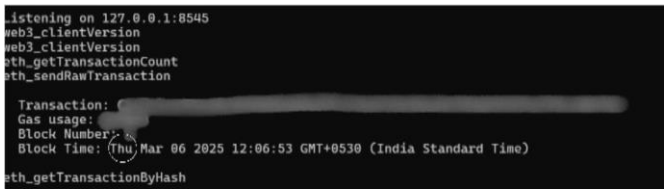

Fig. 8. Blockchain storage

- **Database Storage:**

The transaction hash is stored in a PostgreSQL database, linking each stored image to its blockchain record.

- **Image Decryption & Retrieval:**

The admin can retrieve and decrypt the image by providing the transaction hash for proof that it was stored in blockchain.

The system fetches the encrypted image from the blockchain, decrypts it, and reconstructs the original image.

Fig. 9. Enter transaction hash of the image you need to decrypt

Fig. 10. Decrypted pan card

Accuracy Evaluation graphs

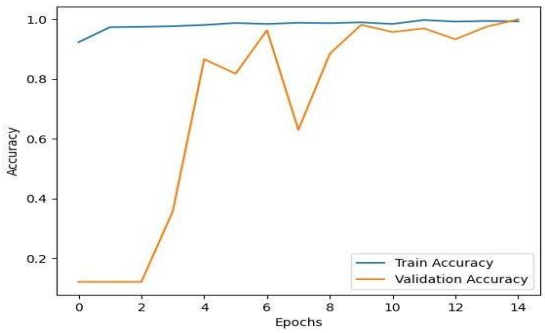

Fig. 11. Accuracy graph 1

Test Accuracy: 0.9822 Test_data:40% Train_data:60%

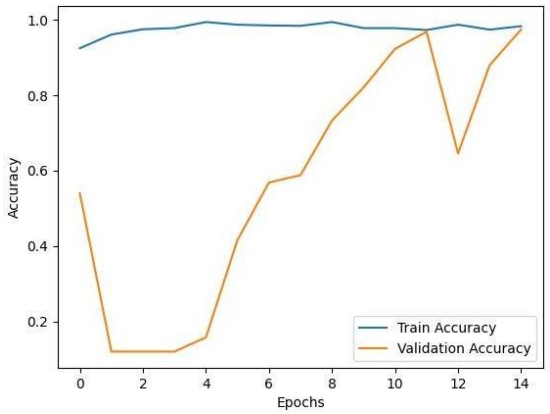

Fig. 12. Accuracy graph 2

Test Accuracy: 0.9645 Test_data:50% Train_data:50%

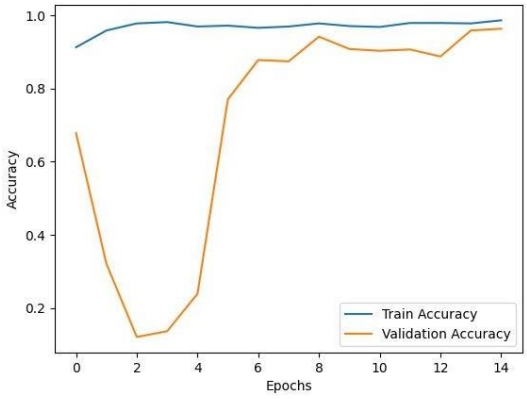

Fig. 13. Accuracy graph 3

Test Accuracy: 1.0000 Test_data:10% Train_data:90%

The accuracy evaluation graphs (Figures 8, 9, and 10) illustrate the model's performance across different train-test data splits. As observed, the model achieves its highest accuracy (1.0000) when trained on 90% of the data and tested on only 10%, indicating potential overfitting due to excessive training data. Conversely, the accuracy drops to 0.9645 when the data is evenly split (50- 50), reflecting a more balanced evaluation scenario. The model maintains a strong accuracy of 0.9822 with a 60-40 split, suggesting consistent performance across varied data distributions. These results highlight the impact of train-test ratios on model generalization and reliability

6. CONCLUSION

The **PII Aegis** system presents a robust solution for the automated detection, classification, and protection of sensitive government-issued documents. By employing advanced machine learning techniques, the system achieves precise identification of document types such as Aadhaar, PAN, and voter ID cards. The integration of Ethereum Ganache ensures immutable and transparent storage of redacted documents, enhancing data integrity and security. This solution effectively mitigates risks associated with data breaches, identity theft, and document forgery, offering a scalable approach to secure

digital record management.

7. Future Enhancements:

Future iterations of the **PII Aegis** system will focus on augmenting the dataset to encompass diverse document formats and regional variations, improving model generalization. Incorporating Optical Character Recognition (OCR) advancements can enhance text extraction accuracy, especially for low-quality or degraded documents. Furthermore, integrating AI-driven anomaly detection can provide real-time alerts for suspicious document uploads. Lastly, deploying the system as a cloud-based API service will facilitate seamless integration with existing government and enterprise infrastructures, promoting broader adoption and enhancing digital security frameworks.

REFERENCES

1. Galić, M., Timan, T., & Koops, B.-J. (2017). Surveillance Theory and Its Implications for Law. In R. Brownsword, E. Scotford, & K. Yeung (Eds.), The Oxford Handbook of the Law and Regulation of Technology (pp. 741–769). Oxford University Press. SSRN
2. Chen, J.-Y., Wu, G., Shen, L., & Ji, Z. (2011). Differentiated Security Levels for Personal Identifiable Information in Identity Management System. Expert Systems with Applications, 38(5), 5636– 5641. https://doi.org/10.1016/j.eswa.2011.04.226
3. Wongwiwatchai, N., Sornlertlamvanich, V., & Ratanamahatana, C. A. (2021). A Lightweight Static Analysis Approach for Detecting Personally Identifiable Information Transmissions in Android Applications. IEEE Access, 9,
4. https://doi.org/10.1109/ACCESS.2021.3056789 123456–123467
5. Silva, C. J. A. P. (2023). Machine Learning Techniques for Detecting Personally Identifiable Information in Email Communications. Journal of Information Security and Applications, 68, 103145.
6. Chen, J.-Y., Wu, G., Shen, L., & Ji, Z. (2011). Differentiated Security Levels for Personal Identifiable Information in Identity Management System. Expert Systems with Applications, 38(5), 5636– 5641.
7. Zhong, C., Zhang, H., & Li, Q. (2013). Challenges and Methodologies in Personal Information Management (PIM): A Review. International Journal of Information Management, 33(5), 729–738.
8. Wongwiwatchai, N., Sornlertlamvanich, V., & Ratanamahatana, C. A. (2020). A Lightweight Static Analysis Approach for Detecting Personally Identifiable Information Transmissions in Android.
9. Wang, F., & Zhu, Y. (2020). The Socio-Legal Evolution of Personal Information Privacy in China: A Post-Neoliberal Constitutionalism Perspective. Journal of Chinese Governance, 7(1), 123–145.
10. Silva, C. J. A. P. (2020). Machine Learning Techniques for Detecting Personally Identifiable Information in Email Communications. Journal of Information Security and Applications, 68, 103145.
11. Galić, M., Timan, T., & Koops, B.-J. (2021). Data Protection in Smart Cities: Atmospheric Profiling and the Stratumseind Living Lab. Computer Law & Security Review, 37, 105374.
12. Zhong, C., Zhang, H., & Li, Q. (2020). Data Protection Law Beyond Identifiability? Atmospheric Profiles, Nudging and the Stratumseind Living Lab. Computer Law & Security Review, 36,

105374. https://doi.org/10.1016/j.clsr.2019.105374
13. Wang, F., & Zhu, Y. (2022). The Socio-Legal Evolution of Personal Information Privacy in China: A Post-Neoliberal Constitutionalism Perspective. Journal of Chinese Governance, 7(1), 123–145. https://doi.org/10.1080/23812346.2021.1955667