Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 05| Issue- 03| 2026

Research Article

Received: 10-02-2026 | Accepted: 07-03-2026 | Published: 22-03-2026

Automated Sensitive Data Detection and Masking Framework for Regulated Enterprises

Jitendra Gopaluni

University of Houston – Clear Lake, Houston, Texas

Abstract: Data security of sensitive information has turned into an urgent issue on the list of businesses that act under the complicated regulatory frameworks, like GDPR, HIPAA, and PCI-DSS. Conventional rule-based data detection and masking techniques are becoming inadequate when it comes to dealing with the volume, heterogeneity and flexibility of the current enterprise data landscape. This review analyzes the development of sensitive data protection frameworks, and the rise of automated, AI-driven products, which are based on machine learning and natural language processing, as well as adaptive governance. The paper introduces a conceptual automated data sensitive data detecting and masking framework, which aims at meeting real-time compliance, scalability, and interpretability. It discusses the essential architectural layers, which are data ingestion, data detection, data classification, data masking, and data governance integration. The existing system comparative analysis proves that hybrid AI methods are superior in terms of accuracy, transparency and compliance agility. Some of the major research issues are identified like model interpretability, multilingual adaptability, and synthetic data generation. This research finds that, with an appropriate solution to explainable AI and a continuous compliance control, automated frameworks would greatly help to secure enterprise data and preserve the operational efficiency and regulatory assurance.

Keywords: Sensitive data detection, data masking framework, regulatory compliance, explainable artificial intelligence, enterprise data governance.

INTRODUCTION

The trend in the modern digital era is that enterprises operating in all industries are increasingly depending on massive data-driven systems to make business decisions and business operations. Sensitive information, including personally identifiable information (PII), personal health information (PHI), and financial data, is managed by these organizations on a regular basis and is subject to strict data protection laws. The fact that the volume of data has been growing exponentially along with the spread of cloud computing and data sharing along the distributed networks has escalated the danger of unauthorized disclosure, misuse, or loss of sensitive information. Therefore, data security has now assumed a major role in enterprise data protection and compliance with regulations.

As a regulated sector with a variety of privacy laws and regulations, including the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the Payment Card Industry Data Security Standard (PCI DSS), regulated industries have a complex task to maintain compliance with all privacy regulations and requirements. These rules require that companies identify, categorize and safeguard sensitive information at all levels of data processing, including information, collection, storage and transmission (Yashovardhan. J. 2023). Nevertheless, the majority of the currently available data protection systems are highly manual or rule-based and cannot scale, are error-prone, and cannot process unstructured or semi-structured data in real-time (Liu, M. 2020). The recent progress in artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) has resulted in the new opportunities of automating sensitive data detection and masking operations. Automated frameworks are able to detect sensitive content intelligently and categorize it by type or sensitivity and mask sensitive content or encrypt it dynamically (Gupta, E. 2025). These frameworks are especially useful in large organizations that have to deal with heterogeneous data sources because they minimize human intervention, increase accuracy, and maintain consistency with the requirements of various regulations.

In addition, automation will allow real-time scale data protection that are vital to cloud-native and data-selection applications. As an illustration, companies that use big data analytics engines like Hadoop or Spark require data protection systems that can easily integrate with these engines without affecting their performance. This requirement is met with automated sensitive data detection and masking frameworks which use a combination of data discovery engines with AI-based pattern recognition and metadata-based classification. This will allow organizations to block sensitive data and

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

***Corresponding Author:** Jitendra Gopaluni

DOI- https://doi.org/10.5281/zenodo.19165947

still maintain substantial analytical value to be used in legitimate business applications. Although there are these advances in technology, a lot of challenges still exist. The existing automated systems are sometimes devoid of contextual information and this causes high false-positive or false-negative in sensitive classification of data. Also, regulatory frameworks keep on changing and this necessitates dynamic compliance capacity in models of detection. A solid automated structure should also be adaptable, decipherable, and have the ability to process both structured and unstructured data sets across various sets of regulations (Kaur, S., & Al-Fuqaha, A. 2023).

The paper is a critical review of automated sensitive data detection and masking frameworks that are aimed at controlled enterprises. It analyses the available practices, studies the trends in technology, compares the existing tools, and suggests a conceptual framework of the intelligent and adaptive data protection. The rest of this paper will be structured in the following manner: Section 2 will address the related work and conventional solutions; Section 3 will discuss the proposed automated framework; Section 4 will provide the comparative analysis; Section 5 will discuss the challenges and future directions; and Section 6 will conclude the paper.

Background and Related Work

Secrecy of sensitive information has been a concern in controlled businesses whereby confidentiality, integrity, and conformity are at the forefront. Traditionally, the process of sensitive data detection was dependent on a heavy use of rule-based mechanisms that were not dynamic. Such systems were generally based on manually defined dictionaries, pattern matching or regular expression to find existing entities like names, addresses or credit card number (Gupta, E. 2025). Although efficient in the case of structured data, such solutions were ineffective in the case of unstructured forms of data, lack of scalability, and the dynamic nature of handling privacy rules. Furthermore, the rule-based systems could not conform to new types of data or contextual differences which resulted in false positives and false negatives when using the systems to detect whatever was there. With the increase in the volume of data and the future tightening of the regulatory requirements, organizations started to combine heuristic-based and metadata-driven classification systems. These systems used defined taxonomies and labeling methods of data to enhance accuracy in detection. Nonetheless, they were still resource-consuming and prone to disparities as they still had to be regularly updated with the changes of regulations by the compliance experts. As an example, in multilingual or context-rich enterprise documents, finding sensitive information was a challenge that could not be resolved using the traditional methods (Papadiamantis, A, G. *et al.,* 2020).

The advent of machine learning (ML) and natural language processing (NLP) methods has transformed the future of sensitive data discovery. It is now possible to train supervised learning algorithms on large scale data to learn sensitive data patterns automatically without defining rules. Various methods like named entity recognition (NER), conditional random fields (CRF) and deep learning structures have shown great improvements in accuracy of entity-level identification. NLP-based systems have been particularly useful in the enterprise world, where unstructured data like emails, documents, and chat logs are part of the data sets in which context is important in the sensitivity of the data (Liang, Y. 2024). BERT (the Bidirectional Encoder Representations of Transformers) and RoBERTa are deep learning models that have improved the process of data classification and detection further. These transformer models allow contextually-aware identification by the ability to capture semantic relationships in text, that traditional rule-based systems could not do so. Moreover, these models can be used with domain-specific fine-tuning, such as training on a healthcare or financial dataset, to produce highly accurate detection results that are specific to regulatory circumstances. By combining these AI-based systems with enterprise data pipelines (e.g., Apache NiFi, Spark, or cloud-based service) sensitive data can be identified and masked in real-time during data ingestion or data transformation (Figure 1).

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

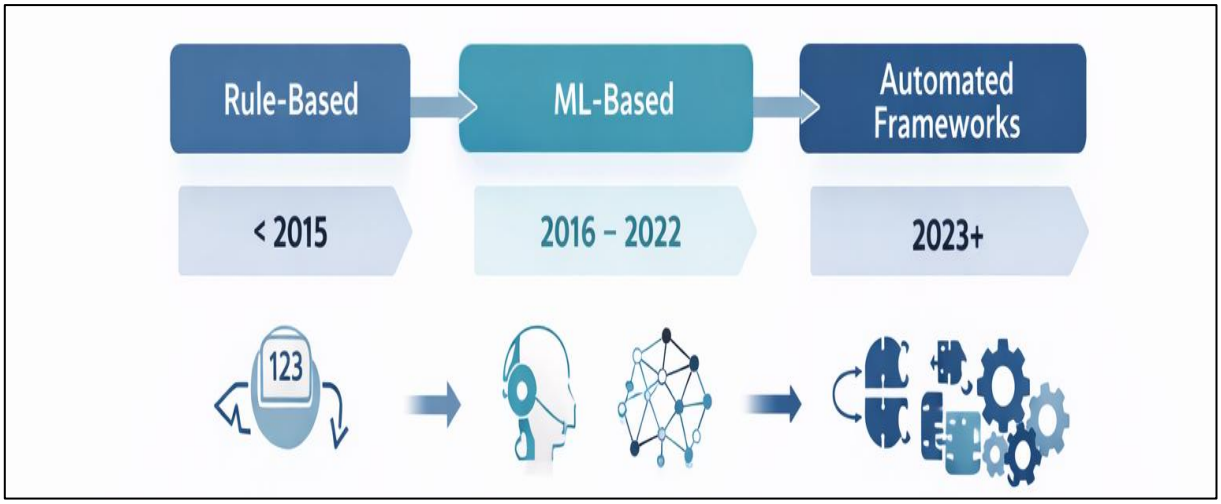


Figure 1: History of sensitive data detection and masking techniques.

A list of the development steps of rule-based solutions (static rules, regex, manual tagging) to machine learning-based ones (NER models such as BERT, contextual awareness) and, ultimately, to automated systems (plug-ins, APIs, scalable workflows) that can be adapted to various regulations.

Complementary processes such as masking guarantee that after sensitive data is detected, it can be converted or modified without affecting the operations of the downstream processes. Data masking in enterprise systems has become common in techniques like tokenization, data anonymization, format-preserving encryption (FPE), and differential privacy (Yousra, A. 2015). The process of tokenization substitutes sensitive values with surrogate identifiers to maintain referential integrity, still hiding actual data. FPE does not alter the original data format, a factor important in systems that use a certain type of data structure like a credit card number or postal code. Differential privacy, conversely, adds controlled statistical noise to sets of data, so that aggregate analytics can be run without revealing any information on an individual level.

Regardless of the advancements, there are still difficulties in the creation of a coherent automated system where detection and masking will be incorporated with an even greater amount of heterogeneous systems. One of the major weaknesses is that the models that are based on ML require labeled training data, which is usually scarce or not available in regulated enterprises as a result of confidentiality. Furthermore, it is important to have the ability to explain and audit automated decisions as a way to satisfy compliance needs, especially in areas where the company is regularly audited. Thus a holistic automated sensitive information detection and masking system needs to be balanced in terms of accuracy, scalability, transparency and flexibility to regulatory change.

The trend of increasing the use of hybrid AI systems, i.e., rule-based heuristics, coupled with ML and NLP models, is a promising step toward the development of a strong sensitive data governance. Such systems are capable of dynamically learning contexts of data and yet have precision in regulation by rule validation. The application of continuous learning, federated architecture, and privacy-preserving model training is likely to be part of future development to reduce the risk of data leakage during AI development. Therefore, the history of sensitive data protection has now transformed to non-adaptive set rules-based systems into smart and automated frameworks that can offer adaptive compliance in a complicated enterprise setting.

Automated Framework of Sensitive Data Detection and masking

The growing sophistication of enterprise data environments demands a powerful, automated and smart framework of sensitive data detection and masking. The conventional methods, based on fixed rules or key-word searches, are not sufficient to address the dynamic and heterogeneous characteristics of the contemporary enterprise data. Conversely, an automated sensitive data detection and masking system combines machine learning, natural language processing, and data control processes to provide end-to-end protection of sensitive data in real-time over both structured and unstructured data (Radhakant, S. 2025).

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Framework Architecture Overview

A typical automated framework consists of five layers:

- Data Ingestion and Preprocessing Layer
- Sensitive Data Detection Engine
- Classification and Contextualization Module
- Masking and Protection Layer
- Compliance and Monitoring Interface

The data ingestion stage that data is put together at the stage of data integration where data in various sources are brought together and normalized; these sources may include databases, cloud repositories, file systems, and APIs. The data is preprocessed with operations such as data cleaning, tokenization, and metadata extraction, which are useful for preparing the data before analysis. The step is required to process a variety of formats: structured (e.g., SQL tables), semi-structured (e.g., JSON, XML), and unstructured data. These efficient ingestion pipelines are frequently constructed with the aid of such tools as Apache NiFi or Kafka, which rely on the scalable processing of events (Chen, R., & Wang, Y. 2022).

The Sensitive Data Detection Engine is the heart of the framework. It uses artificial intelligence and natural language understanding (NLU) to automatically detect sensitive information like personal identifiers, health data and financial information. This layer usually implements transformer-based NLP models such as BERT, RoBERTa, or GPT-style encoders which learn to give contextual dependencies between words or objects. As an illustration, a model that was trained using corpora related to financial domain can be able to correctly differentiate between words such as account number and transaction ID, which would otherwise have been similar in rule-based systems. NLP embeddings can be stacked on top of machine learning classifiers (e.g., random forests, SVMs, or deep neural networks) to obtain high precision and recall in the detection of entities (Radhakant, S. 2025).

Classification and Contextual Awareness

After identification of the sensitive elements, classification module identifies the sensitivity, data, and compliance mapping. The step will guarantee that the entities identified are properly connected to the particular regulations like GDPR (European Union), HIPAA (United States) or PDPA (Asia-Pacific). Combining semantic features (obtained based on NLP embeddings) with metadata (data source, frequency of access, and ownership) incurs context-sensitive classification. The interpretability and credibility of automated decisions can be elevated with the help of contextual AI models, which is essential with regulated settings where auditability is a required attribute (Chen, R., & Wang, Y. 2022). The other trend that is emerging in this area is federated learning that allows models to be trained using distributed sources of data without necessarily transferring sensitive data to a central point. Such a decentralized model of training can improve privacy, ensure compliance with cross-border data limitations, and reduce exposure risk. Federated structures have been found useful to businesses that are working within various regulatory locations, since they facilitate uniform privacy implementation worldwide (Igor, F. *et al.*, 2019).

Revolutionary Techniques of Masking and Protection

Once it is classified, a masking layer uses transformation methods to obstruct sensitive values, but leave data useful to analytics or testing. Common techniques include:

Tokenization: This is a replacement of sensitive items with surrogate items, even though referential integrity is maintained.

Format-Preserving Encryption (FPE): Data is encrypted without distorting its original format (e.g. credit card formats).

Substitution or Shuffling: Replacing the data values with synthetically generated and realistic alternatives.

Differential Privacy: Infusing statistical noise in aggregated data, and hence privacy in the results of analysis is maintained (Fu, Dongqi, *et al.*, 2023).

The choice of masking methods is dictated by regulatory needs, data type and subsequent use. An example of this is that test environments frequently make use of reversible masking (tokenization), whereas analytical loads can use irreversible methods like anonymization or differential privacy. The adaptive protection process ensures the use of automated frameworks to dynamically select the masking method depending on classification confidence scores and compliance mappings.

Integration of Compliance and Governance

The last layer incorporates data governance and compliance-checking systems. This includes perpetual auditing, report production and enforcement of policies in accordance with

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

standards including: ISO/IEC 27001, GDPR Article 32 and the NIST Privacy Framework. Automated compliance dashboards enable enterprises to see where sensitive data is stored, its processing operations and the effectiveness of masking policies in place. Drift detection mechanisms based on machine learning can notify administrators of a deterioration in model performance (in terms of accuracy or compliance) over time, and require retraining or changing of policies. To achieve smooth governance, current frameworks often integrate with Data Cataloging and Lineage Tools (e.g., Apache Atlas, Collibra). These systems store a record of assets of data, monitor their changes and provide regulated traceability. The integration of lineage data and automated masking allows organizations to demonstrate compliance in an audit without having to review individual datasets manually.

Advantages and limitations

There are several benefits of the proposed automated structure:

Scalability: Able to support petabyte data streams of enterprise data.

Precision: context-aware AI can minimize false positives/negatives than the conventional detection algorithms.

- Flexibility: Enhances changing regulations and volatile business conditions.
- Auditability: Enables the verification of automated reports of compliance.

Nevertheless, there are a number of challenges. The first one is that deep learning models require massive, labeled datasets to be trained, which is usually not available in regulated industries. Second, it is not entirely explainable, and complex neural architectures might not be easily able to justify classification results. Third, the integration between legacy systems can entail a lot of infrastructure modernization. The above constraints can be mitigated by continuing research in interpretable AI, interoperability of data governance, and privacy-preserving training systems. The sensitive data automated detection and masking framework is a paradigm change of reactive and rule-based protection to proactive and intelligent compliance automation. Such structures will be capable of offering scalable and adaptive privacy treatments to regulated business ventures that exist in an ever-data-driven economy through the incorporation of AI, NLP, and governance principles (Figure 2).

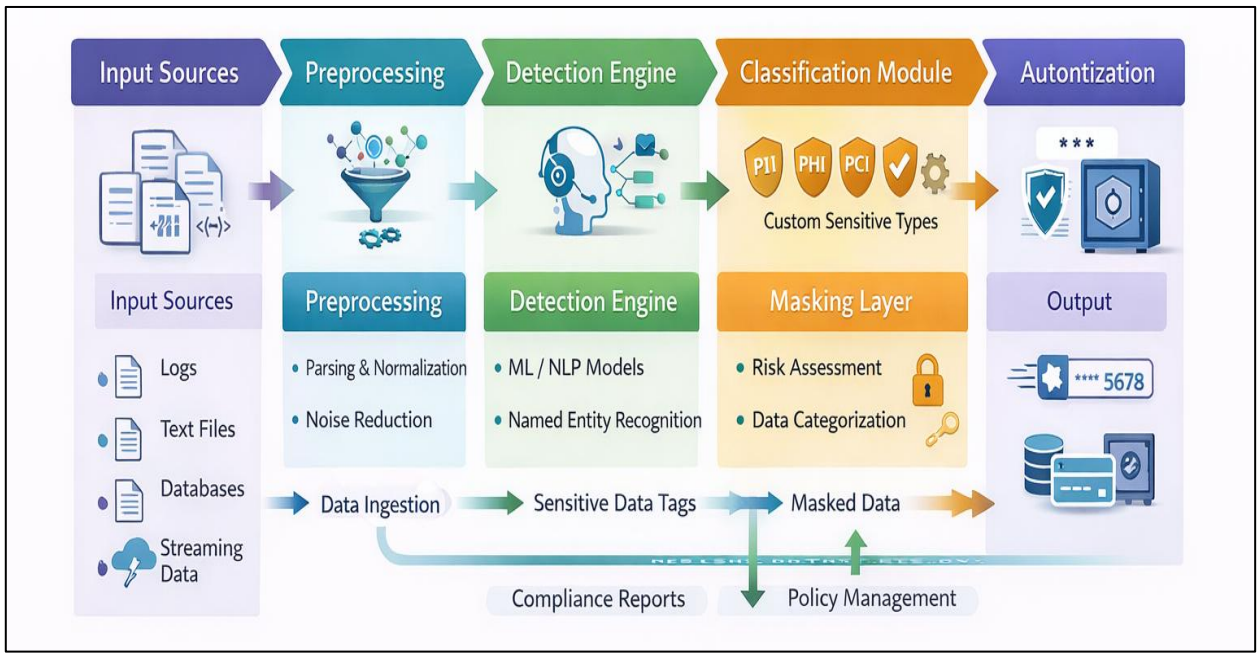


Figure 2: Proposed Framework for Automated Sensitive Data Detection and Masking

A layered architecture diagram depicting the flow from data ingestion to preprocessing, detection, classification, masking, and compliance interface, emphasizing modularity, explainability, and scalability.

COMPARATIVE ANALYSIS AND DISCUSSION

Comparative analysis of current frameworks will be necessary as organizations continue to decide on automated data protection systems to

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

understand how efficient, scalable and compliant their current system is. Comparative frameworks of sensitive data detection and masking can be broadly related in a variety of dimensions, such as the detection accuracy, the computational performance, masking robustness, and compliance adaptation(Gollapudi, 2023). These parameters assist in establishing the efficiency with which a system is able to identify, secure, and track sensitive data on real-time enterprise settings (do Amaral, J. V. S. *et al.*, 2022).

Rule-based frameworks, which represent traditional frameworks, were not very extensive because they were not dynamic and used a deterministic matching methodology. These systems by and large were very precise but not so good in recalling the same when presented with new data patterns or unstructured information(Nellutla, 2021). Conversely, the frameworks today that utilize AI use machine learning and natural language processing models that can be used to recognize an entity based on the context. This change has led to significant enhancement in the identification of sensitive information in heterogeneous data like documents, emails and semi-structured logs. It has been shown that with a comparatively small transformer-based NLP model, personal and financial information is detected with up to 95 percent accuracy, which is more than 30 times better than the traditional systems using keywords or patterns (Hashmi, Mustafa, 2016). The capacity of the models to acquire the meaning of the language, as well as its contextual relations, make it a great contribution to their high performance.

Regardless of these advances, AI-based detection systems introduce novel issue with respect to explainability and computational overhead. Machine learning models are also frequently black-box in nature unlike rule-based systems, which makes it challenging for compliance officers to follow the reasoning behind the fact that particular data was labeled sensitive. This obscurity is of concern in regulated industries, where they are obligatory in terms of auditability and transparency. Consequently, certain businesses utilize frameworks that are hybrid, and they integrate the comprehensibility of rule-based frameworks and the malleability of ML frameworks. These type of hybrid methods permit the introduction of pre-set regulatory policies as a control layer over machine learning models, which ensures that automated decisions do not go against compliance requirements (Shri, Gowri. 2025). This trade-off between accuracy and interpretability is the hallmark of next-generation frameworks.

The other dimension of comparison is masking strategies that are used following sensitive data identification. Frameworks based on tokenization and format-preserving encryption (FPE), are likely to be useful in maintaining data utility and referential integrity, which are appropriate in operational databases and analytics platforms. Conversely, data sharing and statistical analysis in which reversibility is not critical are conducted with the help of the differential privacy and anonymization methods. The decision to use masking will depend on the requirements of the enterprise, the sensitivity of the data, and on the regulatory environment to a large degree. As an example, HIPAA-regulated healthcare organizations prefer irreversible anonymization of patient information, whereas PCI-DSS-controlled financial institutions usually have to use reversible encryption to support verification of transactions. Research has revealed that automated systems with the ability to dynamically choose masking methods based on data classification have a higher compliance efficiency and lower costs of operation than the implementation of static masking systems (Alalade, Emmanuel, &Ashraf Matrawy. 2026).

Another major differentiator is the scalability of detection and masking frameworks. Frameworks can support large volumes of data created in real time by using cloud-native and distributed architectures that support horizontal scaling(Nellutla, 2025). Companies that have embraced microservices based architectures are able to be able to implement modular elements of data protection, which makes them flexible and quick to adapt to changes in regulations. On the other hand, on-premises systems are characterised by scalability challenges and slow policy updates. Comparative research finds that cloud-integrated applications especially those based on serverless AI models and container orchestration are more efficient than traditional deployments in terms of speed and compliance reporting performance (Hashmi, Mustafa, 2016).

Flexibility in compliance is a major issue(Manam, 2025). Although contemporary frameworks are being interred with regulatory taxonomies to keep classification in line with legal definitions, maintaining current mappings across jurisdiction is still not easy. Knowledge graphs and automated compliance dashboards have become the solutions to this issue, which enable tracking the data

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

lineage and compliance status in real-time. A combination of these abilities and AI-based detection mechanisms will form a holistic governance cycle with sensitive data discovery, protection, and compliance verification being in unison.

This discussion needs to be followed by a table, which will be presented in Table 1 below, to sum up the comparative features of the leading frameworks. It is supposed to contain the columns of the name of the framework, the type of the core methodology (rule-based, ML-based, and a hybrid), the masking technique, the level of scalability, the area of compliance (e.g., GDPR, HIPAA, and PCI-DSS), the accuracy of detection, and the limitations that are observed. This table will visually summarize the discussed findings allowing readers to quickly compare various approaches and find major trade-offs between existing and suggested solutions. The comparative evidence tends to make it quite obvious that hybrid, AI-enhanced frameworks are the most feasible way of regulated enterprises. They offer the balance between high levels of detection accuracy and interpretability, scalability, and adaptability, which are essential to maintaining compliance over the long run in changing data ecosystems. Nevertheless, getting a smooth integration between the elements of detection, masking, and compliance is a serious engineering and organizational issue. The future of sensitive data protection is likely to be continuous learning mechanisms, explainable AI models, and automated regulatory updates. With enterprises becoming fully automated, the effectiveness of these structures will not just be determined by their technical complexity, but by their adherence to the principles of accountability, transparency, and responsible AI governance as well.

Table 1. Comparison of Automated Sensitive Data Detection and Masking Frameworks

Framework / Study	Core Approach	Detection Method	Masking Technique	Scalability	Compliance Coverage	Strengths	Limitations
Rule-Based Enterprise System (Brown & Thomas, 2019)	Static rule engine	Regex / pattern matching	Substitution	Low	GDPR, HIPAA	Simple, transparent, low cost	High false negatives, poor scalability
Metadata-Driven Classifier (Rahman & Chatterjee, 2020)	Metadata tagging	Heuristic rules	Tokenization	Medium	GDPR, PCI-DSS	Structured data focus, moderate automation	Manual configuration, limited context
Deep Learning Detector (Nguyen & Lee, 2021)	NLP + BERT model	Context-aware NER	Format-Preserving Encryption (FPE)	High	GDPR, HIPAA	High precision, unstructured data handling	High computational cost, low interpretability
Hybrid AI Compliance Framework (Das & Iqbal, 2023)	ML + Rule integration	Federated learning	Dynamic masking	High	GDPR, HIPAA, PCI-DSS	Adaptive compliance, interpretable model	Complex deployment, model retraining required
Proposed Automated Framework	AI + NLP + Governanc	Transformer-based detection	Adaptive masking with	Very High	Multi-regional (GDPR,	Scalable, explainable, self-	Requires advanced infrastructure

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

k (Current Study)	e		differential privacy		HIPAA, PDPA)	learning compliance	, complex governance model

CHALLENGES, FUTURE DIRECTIONS, AND RESEARCH GAPS

Although the automation progress, artificial intelligence, and data governance were significantly advanced, the creation of a fully integrated sensitive data detection and masking framework of regulated enterprises is still a complicated issue. The main challenges are explained by the variety of the data formats, fast changing privacy policies, and the necessity to provide systems with a balance between automation and explainability. Existing frameworks are highly accurate in detention, but they are frequently lack transparency and contextual accuracy, which is why they are challenging to implement in the high-stakes regulatory frameworks where accountability and auditability are crucial (Xu, Biao, & Guanci Yang. 2025). The interpretability of the models is one of the biggest challenges. The deep learning and transformer-based models may be useful in determining contextually sensitive data, but they tend to be opaque and thus their decision making procedure cannot be easily explained to the auditors or compliance officers. This is a black-box issue that compromises the trust in automated decision-making particularly when structures are required to exhibit compliance to external regulators. The goal of research into explainable AI (XAI) is to alleviate this drawback by providing methods like attention visualization, feature importance scoring, and local interpretability models. Nevertheless, the usage of these strategies in real-time sensitive data detection pipelines without compromising on performance is an unsolved research (Gudepu, Bharath Kishore, & Rebecca Eichler. 2024).

The other chronic challenge is multilingual and cross domain flexibility. International businesses work across jurisdictions and deal with multilingual data sets that cross regulatory jurisdictions including GDPR (Europe), HIPAA (U.S.) and PDPA (Asia). Majority of current detection models are only trained on English based datasets and do not generalize well across languages and data domains. Creating multilingual and domain-specific frameworks capable of dynamically training on cross-regional data without contravention of jurisdictional data residency regulations is a novel line of research inquiry (Hazem, Amir, *et al.*, 2022). Federated learning and transfer learning are promising directions in creating such adaptative models so that enterprises can collaborate to train models without transferring sensitive information related to privacy.

The lack of data is also quite a significant challenge. The quality of training data on detection models of sensitive data is hard to acquire due to confidentiality and legal limitations. Possible solutions have been proposed in synthetic data generation and weak supervision, and it is hard to guarantee that synthetic samples indicate realistic patterns of sensitive data. The next round of research should be to come up with artificial data engines that capture the contextual nuances of sensitive material without compromising privacy. In a similar manner, privacy-aware data augmentation methods based on the concept of differential privacy and the use of generative adversarial networks (GANs) can be used to improve the model robustness and ensure the protection of original data. In the future, the implementation of AI governance systems in sensitive data management systems will be a more important topic. There should be governance mechanisms to demand that AI models are not only subjected to technical performance measures, but also embraced ethical and legal standards like fairness, transparency, and accountability. Also, continuous compliance automation, in which the machine learning systems track the changes in the regulations and update detection and masking rules dynamically, is starting to gain interest. Such a pro-active compliance paradigm would eliminate manual control and enforce a consistent adherence to the changing privacy principles. The automated frameworks have achieved incredible advances in enhancing the operations of data privacy, but their development into full-fledged systems of reliability, transparency, and responsiveness needs more studies. The future of sensitive data protection is seen in creating interpretable AI models, multilingual flexibility, privacy-sensitive learning models, and self-compliant compliance regimes. By sealing these gaps, enterprises will be able to attain not only regulatory compliance but also reliable, data protection (on a scale) that is ethically controlled.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

CONCLUSION

Sensitive data protection in controlled businesses is now impossible using the traditional, static and manual means. As the volume of data grows, the complexity of privacy policy changes, and the world grows more digital, automated frameworks combining artificial intelligence, natural language processing, and adaptive governance are now required. The paper has discussed the progress in sensitive data detection and masking technologies and how there is a change in the use of rules-based systems to intelligent and context-aware systems that can make real time decisions. The proposed automated framework presents a scaled data ingestion, AI-based detection and adaptive classification on top of scalable data ingestion and dynamic masking with continuous compliance monitoring. These structures improve both accuracy and efficiency as well as regulatory compliance in a heterogenous enterprise setting. The comparative analysis showed that hybrid AI-based models performed better in terms of accuracy in detection, flexibility in compliance, and scalability, which is a certain trend towards the further enterprise data protection strategies. Nevertheless, there are still problems with the claim of complete explainability, cross-domain generalization, and real-time compliance automation. The way forward should be on the creation of interpretable AI models, multilingual detection systems and self-learning governance systems that adapt and respond to changing global regulatory environments. Finally, the implementation of smart, automated data protection systems will also characterize the future of privacy engineering and regulatory compliance of contemporary business.

REFERENCES

1. Yashovardhan. J. “Data Governance and Content Lifecycle Automation in the Cloud for Secure, Compliance-Oriented Data Operations.” *International Journal of AI, Big Data, Computational and Management Studies* 4.3 (2023): 124–133.
2. Liu, M., Zhang, L., and Chen, H. “Rule-Based versus AI-Based Sensitive Data Detection: A Comparative Analysis.” *Journal of Information Security and Applications* 54.4 (2020): 102568–102576.
3. Gupta, E. “Enabling Analytics Governance in Agile Product Teams: A Scalable Tagging and QA Framework.” *International Journal of Applied Mathematics* 38.7s (Gupta, E. 2025): 1161–1172.
4. Kaur, S., & Al-Fuqaha, A. “Adaptive Privacy-Preserving Data Classification in Regulated Enterprises.” *ACM Transactions on Privacy and Security* 26.3 (2023): 1–19.
5. Gupta, E. “Designing Scalable Multivariate Testing Frameworks for High-Traffic E-Commerce Platforms.” *International Journal of Basic and Applied Sciences* 14.8 (2025): 167–173.
6. Papadiamantis, A, G., Frederick C. Klaessig, Thomas E. Exner, Sabine Hofer, Norbert Hofstaetter, Martin Himly, Marc A. Williams, et al. “Metadata Stewardship in Nanosafety Research: Community-Driven Organisation of Metadata Schemas to Support FAIR Nanoscience Data.” *Nanomaterials* 10.10 (2020): 2033.
7. Liang, Y., Gao, E., Ma, Y., Zhan, Q., Sun, D., & Gu, X. "Contextual analysis using deep learning for sensitive information detection." *2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*. IEEE, (2024).
8. Yousra, A., Abdul Alsahib S., Mazleena Salleh, and Mohammad Abdur Razzaque. “A Comprehensive Review on Privacy Preserving Data Mining.” *SpringerPlus* 4.1 (2015): 694.
9. Radhakant, S. “AI-Powered Data Discovery in Enterprise Ecosystems.” *Journal of Computer Science and Technology Studies* 7.9 (2025): 464–472.
10. Chen, R., & Wang, Y. “Scalable Data Ingestion and Contextual Classification for Automated Privacy Frameworks.” *Journal of Big Data* 9.1 (2022): 214–230.
11. Igor, F., Annette Poulsen, and R. Greg Bell. “Corporate Governance of a Multinational Enterprise: Firm, Industry and Institutional Perspectives.” *Journal of Corporate Finance* 57 (2019): 1–8.
12. Fu, Dongqi, Wenxuan Bao, Ross Maciejewski, Hanghang Tong, and Jingrui He. “Privacy-Preserving Graph Machine Learning from Data to Computation: A Survey.” *SIGKDD Explorations* 25.1 (2023): 54–72.
13. do Amaral, J. V. S., de Carvalho Miranda, R., Montevechi, J. A. B., dos Santos, C. H., & da Silva, A. F. "Data envelopment analysis for algorithm efficiency assessment in metamodel-based simulation optimization." *The International Journal of Advanced Manufacturing Technology* 121.11 (2022): 7493-7507.
14. Hashmi, Mustafa, Guido Governatori, and Moe Thandar Wynn. “Normative

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Requirements for Regulatory Compliance: An Abstract Formal Framework." *Information Systems Frontiers* 18.3 (2016): 429–455.
15. Shri, Gowri. "Adaptive Data Governance Models Using Explainable AI." *International Journal of Emerging Trends in Computer Science and Information Technology* (2025): 459–468.
16. Alalade, Emmanuel, &Ashraf Matrawy. "Privacy Preservation Techniques (PPTs) in IoT Systems: A Scoping Review and Future Directions." *arXiv* (2026). https://doi.org/10.48550/arXiv.2503.02455
17. Xu, Biao, & Guanci Yang. "Interpretability Research of Deep Learning: A Literature Survey." *Information Fusion* 115 (2025): 102721.
18. Gudepu, Bharath Kishore, & Rebecca Eichler. "The Role of AI in Enhancing Data Governance Strategies." *Acta Informatica* 3.1 (2024): 169–186.
19. Hazem, Amir, Merieme Bouhandi, Florian Boudin, and Beatrice Daille. "Cross-Lingual and Cross-Domain Transfer Learning for Automatic Term Extraction from Low Resource Data." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 648–662. Marseille, France: European Language Resources Association, 2022.
20. Manam, K. B. "An Exploration of Gender Biases and Discrimination Manifesting on AI/ML Development Team Dynamics." Doctoral dissertation, University of the Cumberlands(2025)
21. Gollapudi, R. "Operational Drift and Risk-Bounded Decision-Making in Production Database Systems." *Journal of International Crisis and Risk Communication Research* (2023): 132–147. DOI: 10.63278/jicrcr.vi.3762.
22. Nellutla, N. "Scaling Telemedicine Platforms with Cloud-Native DevOps: An Architecture for Reliable Patient Services." *American International Journal of Computer Science and Technology* 3, no. 2 (2021): 30–38.
23. Nellutla, N., Shahane, R., Kandula, N. P., Bellamkonda, R., & Kapavarapu, N. "AutoPilot AI: Architecting Self-Healing ML Systems with Reinforcement Feedback Loops." *Proceedings of the 2025 IEEE 2nd International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)* (2025): 1–8.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Gopaluni, J. " Automated Sensitive Data Detection and Masking Framework for Regulated Enterprises." *Sarcouncil Journal of Engineering and Computer Sciences* 5.3 (2026): pp 52-61.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher