International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



PII Exfiltration and Detection: Threat Vectors, Detection Techniques and a Layered Defence Framework

Arpitha B M, Bharath Akki ,Kalpana J and Kavana K R, Lavanyashree DP

Akshaya institute of technology /CSE, Tumkuru, India
Email: arpithamallikrajun8@gmail.com
Email: {bharathakkinkl@gmail.com , kalpana05j@gmail.com, ranganathdivakarashat@gmail.com}

How to Cite this Article:
M, A., Akki, B., J, K. & R, K. K. (2026). PII Exfiltration and Detection: Threat Vectors, Detection Techniques and a Layered Defence Framework. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.163

License:
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i11.163

***Abstract*—Personally identifiable information (PII) is among the most frequently targeted assets in modern data breaches. This paper surveys how PII is exfiltrated, from insider misuse and compromised credentials to misconfigured cloud storage and covert network channels, and reviews the main techniques used to detect it, including pattern matching, machine-learning-based entity recognition, behavioural analytics, egress monitoring and deception. We then propose a five-layer defence framework that combines these techniques and discuss their trade-offs in accuracy, coverage and operational cost. The paper concludes that no single control is sufficient and that layered, context-aware detection tied to data classification gives the most dependable protection.**

***Index Terms*—PII, data exfiltration, data loss prevention, anomaly detection, named entity recognition, insider threat, privacy.**

I. Introduction

Personally identifiable information (PII) is any data that can identify a person, either alone or when combined with other data. Names, national identity numbers, email addresses, financial account numbers and health records all qualify [1]. Because PII can be sold, used for fraud or leveraged for extortion, it is a consistent target for attackers, and its loss exposes organisations to regulatory penalties under frameworks such as the GDPR [2] as well as reputational damage.

Exfiltration is the unauthorised transfer of data out of an organisation's control. It may be carried out by external attackers who have gained a foothold, by malicious insiders, or accidentally by well-meaning employees. Industry breach reports repeatedly show that stolen credentials, human error and misconfiguration are leading contributors to PII exposure [3].

This paper makes three contributions: (i) a taxonomy of common PII exfiltration vectors, (ii) a review of detection techniques and their limitations, and (iii) a layered framework that combines them. Section II gives background, Section III reviews detection methods, Section IV presents the framework, Section V discusses results and trade-offs, and Section VI concludes.

II. Background

A. What Counts as PII

PII is commonly divided into direct identifiers (for example a passport number or full name) and quasi-identifiers (for example date of birth, postcode and gender) that identify a person only in combination [1]. Detection systems must handle both. Sensitivity also varies: a name alone carries less risk than a name linked to a bank account or diagnosis.

B. Exfiltration Vectors

Table I summarises the most common routes by which PII leaves an organisation. Fig. 1 later shows where each is monitored in the proposed framework.

TABLE I. COMMON PII EXFILTRATION VECTORS

Vector	Typical Actor	Example
Web and cloud upload	Insider or malware	Upload to personal cloud storage
Email and messaging	Insider or error	Mis-sent spreadsheet with customer records
Removable media	Insider	Copy of database export to USB drive
Covert network channels	External attacker	Data hidden in DNS or HTTPS traffic
Misconfigured storage	Process failure	Publicly readable cloud bucket
Compromised credentials	External attacker	Bulk API or database queries using stolen login

Covert channels deserve special mention. Attackers who already hold access often split data into small chunks, compress or encrypt it, and send it over protocols that are normally permitted, such as DNS or HTTPS, so that no single transfer looks unusual.

III. Detection Techniques

A. Pattern Matching and Validation

The most widely deployed approach uses regular expressions and checksum validation to find structured identifiers such as payment card numbers (validated with the Luhn algorithm), national ID formats and email addresses. It is fast and transparent, but it misses unstructured PII such as names and addresses in free text, and it generates false positives when numeric strings resemble identifiers.



B. Machine-Learning Entity Recognition

Named entity recognition (NER) models label people, locations and organisations in free text and can be trained to find PII that fixed patterns miss [4]. Combining NER with pattern matching typically improves coverage, though it requires labelled data, adds compute cost and can degrade on domain-specific text such as clinical notes.

C. Behavioural and Anomaly Detection

User and entity behaviour analytics build a baseline of normal activity for each account and flag deviations, for example an analyst suddenly querying ten times more customer records than usual or accessing data at unusual hours [5]. This approach can reveal exfiltration that uses legitimate credentials, but it needs a training period, and legitimate changes in behaviour produce false alarms.

D. Egress and Network Monitoring

Data loss prevention (DLP) gateways inspect email, web and cloud traffic for sensitive content and enforce policy by blocking or quarantining transfers [6]. Traffic-analysis methods look for tunnelling signatures such as high-entropy DNS queries, unusually long hostnames or sustained low-volume uploads. Encrypted traffic limits content inspection, so TLS interception or endpoint agents are often needed.

E. Deception and Honeytokens

Honeytokens are fake records, credentials or documents seeded among real data. Because no legitimate process should touch them, any access is a high-confidence alert. They are cheap to deploy and have very low false-positive rates, but they only detect actors who interact with the decoys.

IV. Proposed Layered Framework

We propose combining the techniques above into five layers, shown in Fig. 1. Discovery and classification first locate PII and label its sensitivity. A policy engine then uses sensitivity, user role and context to decide what is allowed. Behavioural analytics watch for abnormal access patterns, egress monitoring inspects data leaving the environment, and a final alerting and audit layer supports response and compliance evidence.

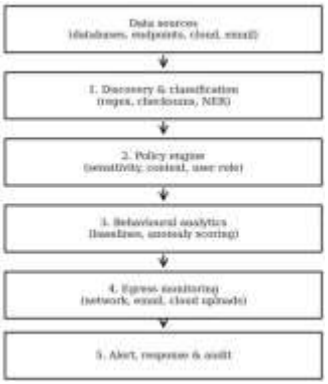

Fig. 1. Layered framework for PII exfiltration detection.

The layers are designed to compensate for each other's weaknesses. Classification gives the later layers context, so that a large transfer of public data is treated differently from a small transfer of health records. Behavioural analytics catch misuse of valid credentials that content inspection alone would pass, while content inspection catches accidental leaks that exhibit no behavioural anomaly.

V. Results and Discussion

Table II compares the techniques qualitatively against the criteria most relevant to deployment. The ratings reflect the general characteristics reported in the literature [4]–[6] rather than measurements from a specific deployment.

TABLE II. Qualitative Comparison of Techniques

Technique	Coverage	False Positives	Cost
Pattern matching	Structured PII only	Medium	Low
NER / ML	Free text	Medium	Medium
Behavioural analytics	Credential misuse	Medium to high	High
Egress / DLP	Known channels	Low to medium	Medium
Honeytokens	Intruders only	Very low	Low

Three observations follow. First, no technique covers every vector, which supports the layered design. Second, false positives are the main operational burden: alert fatigue leads analysts to ignore warnings, so tuning and risk-based prioritisation matter as much as detection accuracy. Third, encryption and cloud adoption move the control point from the network perimeter to endpoints, identity and the data itself.

Limitations of this work should be noted. The framework is conceptual and has not been validated experimentally here; future work should evaluate it on labelled datasets and measure detection rate, false-positive rate and time to detect. Privacy of the monitoring itself also needs care, since behavioural monitoring processes employee data and must be proportionate and lawful.

VI. Conclusion

PII exfiltration occurs through many routes, and attackers frequently use legitimate access or permitted channels to avoid simple controls. Pattern matching, machine-learning classification, behavioural analytics, egress monitoring and deception each address part of the problem. A layered framework anchored in data classification and context-aware policy offers the most dependable protection, and its effectiveness depends on careful tuning to keep false positives manageable. Future work will focus on empirical evaluation and on reducing the privacy impact of the monitoring itself.



REFERENCES

[1] E. McCallister, T. Grance and K. Scarfone, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," NIST Special Publication 800-122, 2010.

[2] European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, 2016.

[3] Verizon, "Data Breach Investigations Report," annual report series.

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition," in Proc. NAACL-HLT, 2016, pp. 260–270.

[5] V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[6] A. Shabtai, Y. Elovici and L. Rokach, A Survey of Data Leakage Detection and Prevention Solutions, Springer, 2012.