

A Systematic Review and Taxonomy for Privacy Breach Classification: Trends, Gaps, and Future Directions

Clint Fuchs

*The Beacom College of Computer and Cyber Sciences
Dakota State University
Madison, USA
clint.fuchs@trojans.dsu.edu*

John D. Hastings

*The Beacom College of Computer and Cyber Sciences
Dakota State University
Madison, USA
john.hastings@dsu.edu*

Abstract—In response to the rising frequency and complexity of data breaches and evolving global privacy regulations, this study presents a comprehensive examination of academic literature on the classification of privacy breaches and violations between 2010-2024. Through a systematic literature review, a corpus of screened studies was assembled and analyzed to identify primary research themes, emerging trends, and gaps in the field. A novel taxonomy is introduced to guide efforts by categorizing research efforts into seven domains: breach classification, report classification, breach detection, threat detection, breach prediction, risk analysis, and threat classification. An analysis reveals that breach classification and detection dominate the literature, while breach prediction and risk analysis have only recently emerged in the literature, suggesting opportunities for potential research impacts. Keyword and phrase frequency analysis reveal potentially underexplored areas, including location privacy, prediction models, and healthcare data breaches.

Index Terms—data breach, data privacy, privacy breach, taxonomy.

I. INTRODUCTION

Cybersecurity breaches continue to be a global problem with 2023 seeing the highest volume of breaches of any of the previous 17 years and the number of attacks exploiting vulnerabilities as the main initiator increasing 180% from the previous year [1]. [2] supported those findings, reporting an overall increase in the volume of breaches as the threat landscape continues to evolve. Considering the massive amounts of data that are collected about each and every person [3], the findings should be a concern for every individual or entity that provides or collects sensitive data.

Privacy laws such as the European General Data Protection Regulation (GDPR), the California Privacy Rights Act (CPRA), and other emerging global regulations are reshaping how organizations must handle data. Hefty fines are incurred by data collectors who do not abide by the appropriate laws [4]. In 2023 alone, Meta was fined USD 1.3 billion (Facebook) and USD 413 million (Instagram) for GDPR infractions [5]. This is just the tip of the iceberg. With the growing global adoption of privacy laws, organizations no longer have a choice; they are required to emphasize the protection of sensitive personal information that they collect, process, and store [6].

As regulatory requirements and the volume of collected data increase, so does the need for a comprehensive understanding of how privacy breaches occur and are classified in academic literature. In response, this study aims to develop a taxonomy to shed light on existing research involving the classification of privacy breaches and violations. The study is guided by the following research questions:

- RQ1: What are the trends in existing research related to the classification of privacy breaches (violations)?
- RQ2: Concerning privacy breach (violation) classification, what are the under-researched areas and emerging fields of study?
- RQ3: At what frequency do select keywords (privacy, breach, violation, etc.) occur within the surveyed literature?

The remainder of this paper is presented as follows. Section II provides background related to privacy and privacy breaches. Section III presents the methodology of the literature review. Section IV presents the findings including the resulting novel taxonomy, followed by extended findings in Section V. Future work appears in Section VI with conclusions in Section VII.

II. BACKGROUND

A. Taxonomy

A taxonomy is defined strictly as the classification of organisms, both living and extinct, or more succinctly, biological classification [7]. In a broader sense, taxonomy refers to the science of classification by establishing a hierarchy of superior and subordinate groupings [7]. Taxonomies give structure and organization to various fields of knowledge, allowing researchers to investigate relationships between concepts, thus allowing researchers to form hypothesis about these relationships [8]–[10]. In addition, taxonomies are helpful for the advancement of discussion, research, and pedagogy [10], [11]. [12] noted that the cybersecurity field is fragmented, with a broad scope. Many fields including cybersecurity benefit from the use of taxonomies to provide order where it may be lacking or nonexistent [10], [13].

B. Privacy

In the late 1800s, in response to newspaper circulation of photographs of private individuals, [14] distilled privacy into “the right to be let alone.” Today, however, the boundaries of privacy have shifted dramatically. The rise of digital platforms has transformed privacy concerns from simple physical intrusions into issues of data surveillance and personal information management. As noted by [15], practically every business and organization with which we interact collects large amounts of personal data from all aspects of our lives, from the purchases we make, to the web sites we visit, and more. Individuals or data providers trade personal information as currency with service providers or data collectors in exchange for services [16]. To get an idea of the scope, Facebook [17] collects over 52,000 unique attributes in each individual account holder. Personal information is the currency Facebook and other providers take as payment for the use of the services they provide.

C. Privacy Breaches

A general definition of a data breach is given by [18] as “an illegal disclosure or use of information without authorization”. The Department of Health and Human Services adds privacy and security to the definition of confidential health information [18], [19]. [20] posited that legal scholars merge privacy and security, while asserting that these are separate concerns. Data security and privacy are undoubtedly intertwined, and privacy is often viewed through the lens of security. Despite network security and privacy policies being in place, privacy is not guaranteed and can still be violated [21]. [22] noted that studies tend to focus on the development of threat and harm taxonomies as they relate to security. These taxonomies are then applied to the classification of security breaches with a focus on mitigating the risk the classified breaches pose or the outcome specific security breaches incur.

D. Privacy Legislation

In 2018, California introduced the California Consumer Privacy Act of 2018 (CCPA) [23]. In the same year, the European Union (EU) adopted the General Data Protection Regulation (GDPR), which is considered by many to be the gold standard [23]. These laws serve to protect the citizenry of California and EU respectively. This protection goes beyond the territorial limits of each entity and has caused organizations all over the world to update data security and compliance due to the potentially large fines that can be incurred [23].

GDPR Privacy Principles [4], [23]:

- 1) Processing must be lawful, fair, and transparent to the EU citizens as data subjects.
- 2) Data can only be processed for clearly stated legitimate reasons to the data subject.
- 3) Data minimization must be in practice.
- 4) Data accuracy and timeliness must be adhered to.
- 5) Storage limitation and purpose limitation on information in that it can only be kept for the least possible time.
- 6) All processing of data and data flows must ensure security, confidentiality, and integrity.

- 7) The data controller is accountable for compliance with GDPR protection privacy laws.

III. STUDY DESIGN

A. Methodology

The objective of this study is to discover existing and emerging trends and research gaps by conducting a systematic review of the literature on the classification of privacy breaches. To answer RQ1, RQ2, and RQ3, this study employed a quantitative approach in the implementation of a systematic literature review that was based on quantitative descriptive research methodology. This methodology was chosen due to it being non-experimental, as opposed to experimental designs, which focus on the measurement of selected variables. The non-experimental approach was a fit due to its ability to identify causality and that it is also possible to identify associations and relationships [24].

B. Literature Search

To conduct the literature search, the terms of the primary research question (RQ1) were used to generate queries based on the terms and synonyms, as shown in Table I. A corpus of literature was established through a systematic review of the literature utilizing several databases. These included the ACM Digital Library, IEEE Xplore, and Google Scholar. The ACM Digital Library and IEEE Xplore were selected on the basis of reputation and subject relevancy. Google Scholar was also included to increase the scope of the search and to include additional sources, including Elsevier and Springer. The search was conducted in November 2024.

TABLE I
LITERATURE SEARCH QUERY TERMS

security breach classification	cyber breach
security breach classification	breaches categorization
data security classification	privacy violations
data privacy classification	threat classification
privacy taxonomy breach	breach classification
privacy framework violation	privacy classification
privacy framework data breach	privacy systematic

C. Literature Selection

The set of articles found in the search was then narrowed using the inclusion and exclusion criteria listed in Table II. The screening process was conducted sequentially from the first result, until the studies resulting from the search were deemed unrelated according to the inclusion/exclusion criteria. This process was repeated for each search query term. The first round of the screening process eliminated candidates that met the exclusion criteria. The remaining candidates were manually selected by reviewing the abstract according to the inclusion criteria. The goal was to select papers that engaged in some form of classification related to data/privacy breaches or violations. This result was the exclusion of research whose focus did not match RQ1, such as studies involving data classification instead of breaches or violations.

TABLE II
STUDY INCLUSION/EXCLUSION CRITERIA

Inclusion Criteria	Exclusion Criteria
Studies involve classification/categorization and some combination of the terms data, privacy, breach, and/or violation	Study does not give a clear connection to RQ1
In English	Not in English
PDF downloadable	PDF unconvertible for analysis
Relevant to topic (RQ1)	Secondary sources such as magazines or reviews
Found in conference/journal	No duplicates

D. Taxonomy Construction

After the selection process, a thematic analysis was performed on the final set of studies. From this analysis, a novel taxonomy was constructed to classify the focus areas of existing research related to privacy breach and violation classification.

E. Extended Corpus Analysis

To complement the thematic taxonomy, additional corpus-wide analysis was conducted to uncover trends over time and frequency-based insights. This included keyword frequency counting, temporal distribution analysis of publication dates, and TF-IDF scoring across the selected literature set. These steps were taken to address RQ1–RQ3 in greater depth and are detailed in the extended findings section.

IV. FINDINGS

The literature search produced a large number of matches ranging from 400,000 to more than a million results for each query. From those, the targeted selection process produced a total of 29 articles published between 2010 and 2024: 23 in Google Scholar and six in IEEE Xplore with two of those articles appearing in (Table III). Google Scholar produced results from: ACM, AIS Electronic Library, Elsevier, HICSS, Little Lion Scientific, MDPI, Oxford University Press, River Publishers, Springer, and two that could not be determined. No papers were produced from the ACM Digital Library as the results focused primarily on classifying the privacy level of data or on aspects of differential privacy and thus did not meet the inclusion criteria.

TABLE III
DATABASE AND SELECTED STUDIES COUNT

Database	Selected Study Count
ACM Digital Library	0
Google Scholar	23
IEEE Xplore	6

A. Taxonomy

The purpose of this study was to explore the classification of privacy breaches, but only a limited number of studies [25]–[27] from the selected corpus that met expectations for RQ1. During the analysis of the selected studies, the areas of focus found in the corpus revealed themselves, as shown in Table IV. This resulted in a novel taxonomy for research related

to privacy breach/violation classification that warrants future exploration. The related categories were further distilled into a macro set of privacy research focus categories for academic and potential future work in this field.

TABLE IV
PRIVACY RESEARCH FOCUS TAXONOMY

Derived Categories	Distilled Categories
Breach Classification	Classification [Breaches, Reports, Threats]
Detect Breaches	Detect [Breaches, Threats]
Detect Threats	Predict [Breaches, Threats, Outcomes]
Predict Breaches	Risk [Breaches, Violations, Threats]
Report Classification	
Risk	
Threat Classification	

1) *Classification of Breaches*: This category is important because security professionals learn mitigation strategies from documented breaches. To accomplish this task, researchers used machine learning to classify breaches according to a corresponding taxonomy. [25], [28]–[30] focused on the classification of breach incidents. While [18], [29] focused on the classification of IoT and IoMT data breaches, respectively. [31] study focused on predicting the types of cyber-breach outcomes, a major portion of the study focused on developing a taxonomy and framework.

2) *Classification of Reports*: This category involved machine learning for the classification of breach reports produced by organizations after breach events. The research conducted by [27], [32] focused on the analysis of data breach reports and classification to improve internal controls.

3) *Detect Breaches*: This category consists of studies that leverage classification for the purpose of detecting breaches. [33] defined four categories of privacy violations to focus on privacy challenges, privacy violations, and privacy preserving techniques. In relation to data stored on database systems, and [34] leveraged NLP to personally identifiable information (PII) in data that was publicly available. [26] defined a four-dimensional taxonomy to record privacy violations, the severity of individual privacy violations, and the probability that a data provider would stop providing data due to violations and their severity. [35] developed a tool to detect privacy violations found on social networks. While [36] developed a tool to detect privacy violations in Android applications. Studies focused on associated privacy violations, such as those associated with location-based services on mobile devices [37] or via proximity-based applications [38]. [39] developed detection models to identify self-disclosures of personal information.

4) *Detect Threats*: Detecting potential threats to sensitive data or security gives security professionals direction for mitigation strategies. Studies in this category utilize classification for the purpose of identifying threats to sensitive data or security. [40] created a tool for the purpose of identification and classification to validate alerts and incidents. Similarly, [41] developed a tool to detect and classify cyber-attacks. [42] instead focus on classification of unstructured data to conduct threat analysis. Insider threat detection was presented by [43].

[44] focused on classification of location privacy attacks and approaches.

5) *Predict Breach*: Predicting data breaches is an important undertaking that allows organizations to prepare for an event. If one is aware of potential security threats, defenses can be implemented. [45] used classification algorithms to predict network security threats by gathering pertinent attributes, including network protocol and service. [46] instead looked to predict privacy leakages from user-generated content.

6) *Risk*: Determining the risk associated with a particular activity or application provides security professionals with insight into which areas to focus resources. [47] analyzed android apps by leveraging machine learning techniques to classify and identify Android applications that intentionally or unintentionally leak information to third parties.

7) *Threat Classification*: This category is important because it quantifies threats, giving security professionals direction for the application of resources to combat. [22] focus on threat class classification. The authors present a detailed multidimensional classification model that includes both internal and external threats.

B. Data Extraction

After the creation of the novel taxonomy depicted in the previous section, the corpus was manually reviewed by the author. This process involved identifying keywords from each study's title, abstract, and methodology. Then each study was placed in a category based on fit. The first phase classified each study according to the first set of more fine-grained categories shown in Table IV to facilitate RQ2. The next phase began by distilling the fine-grained categories into a macro set, also shown in Table IV to get a broader overview of the corpus and trend identification for RQ1.

C. Taxonomic Classification

Table V shows the classification of the articles in the corpus according to the taxonomy. [48] note that taxonomies for specific domains are highly useful for a multitude of applications, and a taxonomy can be extracted from the key words and phrases found in a corpus. More importantly, [48] stated that beyond simple keywords and phrases, additional knowledge and context is required to build a taxonomy. This serves as inspiration for the method used in this research. The selected studies that comprise the corpus, each provide key words and phrases that state the focus of the research. The context resulted in the categories that were derived from the corpus shown in Table IV. The derived categories were distilled into macro categories that share high-level keywords and themes that match one of the distilled categories shown in Table IV, allowing a high-level view of the research focus contained within the selected corpus.

V. EXTENDED FINDINGS AND ANALYSIS

With the corpus in hand, a deeper analysis was conducted to comprehensively address the research questions and provide a broader understanding of current research. This extended

TABLE V
STUDY CORPUS CLASSIFIED BY MACRO CATEGORIES

Literature	Classification	Detect	Predict	Risk
[18], [25], [28]–[31], [33], [49], [50]	Breach			
[27], [32], [51]	Report			
[22]	Threat			
[26], [34]–[39], [52]		Breaches		
[40]–[44]		Threats		
[45], [46]			Breach	
[47]				Risk

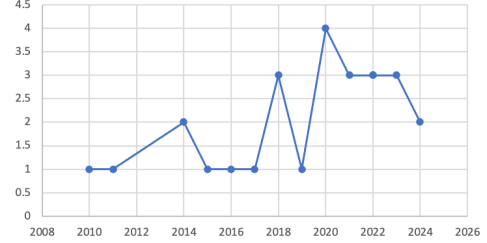


Fig. 1. Corpus Published by Year

analysis includes temporal trends, keyword frequency, and TF-IDF scoring across the selected literature.

To understand how related research has progressed over time, Fig. 1 shows the number of studies found in the selected corpus by year. Initially, a 10-year window was selected for analysis, but with minimal results garnered from the past ten years, two outliers from 2010 and 2011 were included. It is interesting to note that the trend is in an upward trajectory. The data suggest that the increase in studies related to privacy classification corresponds to the adoption of GDPR in 2018 and COVID-19 in 2020, as noted by [28], [34], [45]. However, an investigation of causality is outside the scope of this study.

A. RQ1: Classification Trends

To address RQ1, the distilled taxonomy shown in Table IV was used to classify the corpus. Fig. 2 reveals the results of classifying the corpus by year and macro categories. The data show that over the past 10+ years, the Classify and Detect categories of privacy or data breaches dominate the landscape. Fig. 2 shows that these two categories have been active throughout this period of time. However, Detect failed to make an appearance over the last two years, indicating that this area has likely been saturated. Classify, on the other hand, remains active. Predict and Risk do not appear until 2021 and 2023 respectively, indicating recent interest by researchers.

Another quantifiable measure lies in the number of citations per category shown in Table VI. Threat Classification, Detect Threats, and Breach Classification are the dominant fields of study in the previous 10+ years. However, several of the articles in the corpus have exceptionally high citation counts, which skew the numbers. [22] (434), [18] (532), and [44] (408) affect Threat Classification, Detect Threats, and Breach Classification categories. After removing these three articles from the citation count and looking at the remaining 26 articles (with individual

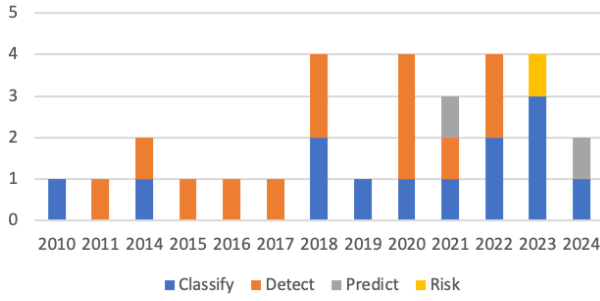


Fig. 2. Publications Classified by Macro Categories by Year

citation counts of less than 100), the final column in Table VI paints a different picture. Threat Classification is affected the most, going to 0. The low counts for Predict Breaches (2) and Risk (0) may be attributed to the fact that they are relatively newer studies.

TABLE VI
CORPUS CLASSIFICATION AND CITATION METRICS

Classification	Citation Count	<100 Citation Count
Risk	0	0
Threat Classification	430	0
Report Classification	83	83
Predict Breaches	2	2
Detect Threats	537	120
Detect Breaches	94	94
Breach Classification	547	15

B. RQ2: Underexplored & Emerging Areas

To address RQ2, the finer-grained taxonomy from Table IV was employed. Looking at Fig. 3, Breach Classification remained an area of interest, with the majority of studies appearing in 2023. The Predict Breaches category did not appear until 2021 and again in 2024, indicating that it is an emerging and understudied area of research. Risk also made its first (and only) appearance in 2023, indicating an emerging field of study. Table VI reinforces the idea that Predict Breaches and Risk represent emerging trends in their respective fields.

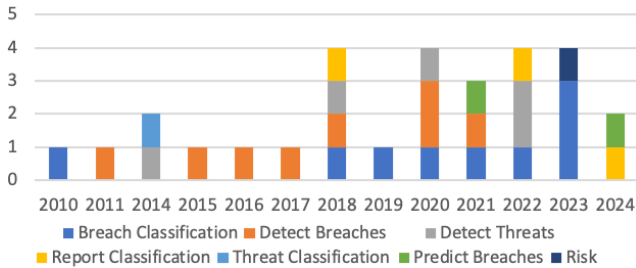


Fig. 3. Publications Classified by Derived Categories by Year

C. RQ3: Keyword Frequency

To address RQ3, word frequency analysis was performed on the selected corpus. Word frequency analysis refers to the number of times words are found in each document. TF-IDF



Fig. 4. Corpus Word Frequency Cloud

provides a measure of potential gaps in words as shown in Table VII, or phrases as shown in Table VIII. The analysis of TF-IDF is important because calculated scores can uncover themes by identifying major words or phrases. Potential gaps can also be identified by the absence of particular topics from the analysis. Table VII reveals the most important keywords, while Table VIII reveals the important keyword phrases

The word Privacy occurred 1689 times in 27 of 29 documents in the corpus. The term Breach occurred 339 times across 21 of 29 documents. The word Violations occurred 375 times in 13 of 29 documents. Fig. 4 represents the word frequency cloud for the study corpus. The results are not surprising considering the field and search terms that were used to identify the corpus. It is interesting to note that the terms Model and Threat appear more frequently, with Threat being more meaningful for future investigation.

TABLE VII
CORPUS WORD FREQUENCY TOP HITS

Word	Frequency	% Shown	% Processed	% Total	No. Cases	% Cases	TF-IDF
Data	2097	3.34%	1.90%	1.06%	29	100%	0.0
Privacy	1689	2.69%	1.53%	0.86%	27	93.10%	52.4
Information	1108	1.76%	1.00%	0.56%	29	100%	0.0
Security	659	1.05%	0.60%	0.33%	28	96.55%	10.0
User	581	0.92%	0.53%	0.29%	25	86.21%	37.5
Based	534	0.85%	0.48%	0.27%	29	100%	0.0
Users	533	0.85%	0.48%	0.27%	26	89.66%	25.3
Model	510	0.81%	0.46%	0.26%	27	93.10%	15.8
Threat	494	0.79%	0.45%	0.25%	16	55.17%	127.6
Cyber	435	0.69%	0.39%	0.22%	17	58.62%	100.9
Breach	404	0.64%	0.37%	0.20%	18	62.07%	83.7
Time	387	0.62%	0.35%	0.20%	27	93.10%	12.0
Violations	375	0.60%	0.34%	0.19%	13	44.83%	130.7
Breaches	339	0.54%	0.31%	0.17%	21	72.41%	47.5

The frequency of phrases extracted from the corpus illuminates themes and potential areas for future attention, as shown in Table VIII. High TF-IDF scores, e.g., Privacy Violations (106.9), indicate phrases that occur frequently in specific documents but are relatively rare across the broader corpus. Such findings highlight their significance within individual studies and their underrepresentation elsewhere, making them potential research gaps worthy of further exploration. A threshold of 60 for TF-IDF was used to identify the phrases Location Based (101.5), Healthcare Data (96.4), Data Provider (89.2), Prediction Models (67.4), Healthcare Data Breaches (64.3), and Android Apps (62.7), and Privacy Policy (67.1) stand out as possible gaps within the corpus.

TABLE VIII
CORPUS PHRASE FREQUENCY TOP HITS

Phrase	Frequency	No. Cases	% Cases	TF-IDF
Privacy Violations	279	12	41.38%	106.9
Data Breaches	153	14	48.28%	48.4
Data Breach	148	13	44.83%	51.6
Location Based	119	4	13.79%	101.5
Privacy Violation	115	9	31.03%	58.4
International Conference	112	22	75.86%	13.4
Machine Learning	91	17	58.62%	21.1
Big Data	84	12	41.38%	32.2
Healthcare Data	83	2	6.90%	96.4
DOI Org	82	13	44.83%	28.6
Privacy Policy	78	4	13.79%	67.1
Location Privacy	73	3	10.34%	71.9
Personal Information	72	16	55.17%	18.6
Data Provider	61	1	3.45%	89.2
Cyber Breach	60	3	10.34%	59.1
Physical World	60	2	6.90%	69.7
Prediction Models	58	2	6.90%	67.4
Cyber Security	57	10	34.48%	26.4
Data Collection	57	11	37.93%	24.0
Social Network	55	6	20.69%	37.6
Android Apps	54	2	6.90%	62.7
Threat Intelligence	54	4	13.79%	46.5
Social Networks	53	8	27.59%	29.6
Online Social	50	7	24.14%	30.9
Information Security	49	9	31.03%	24.9
Location Based Services	48	2	6.90%	55.7
Data Providers	46	1	3.45%	67.3
Decision Tree	46	6	20.69%	31.5
Threat Data	46	3	10.34%	45.3
Breach Reports	45	6	20.69%	30.8
D HTTPS	45	3	10.34%	44.3
Healthcare Data Breaches	44	1	3.45%	64.3
Computer Science	43	14	48.28%	13.6
Privacy Concerns	43	12	41.38%	16.5
Information Systems	42	15	51.72%	12.0
Privacy Preserving	41	6	20.69%	28.1
Cyber Insurance	40	4	13.79%	34.4
Social Media	40	14	48.28%	12.7

VI. FUTURE WORK

Opportunities exist to learn more about privacy or data breach classification and how it applies to the overall protection of privacy, and to investigate new and underrepresented areas in the domain. Beyond the emerging trends and identified gaps, this study could be expanded to garner deeper insights into privacy breach classification. The novel privacy breach classification taxonomy developed by this research provides additional avenues for discovery. For example, each category could be expanded into additional subcategories such as study type, population, and controls. Conducting additional systematic reviews of the literature by category would afford a more focused delineation of word and phrase frequencies. This could provide a more focused view of emerging trends and potential gaps in specific categories of interest.

VII. CONCLUSION

With the adoption of privacy laws, identifying emerging trends and gaps related to the classification of privacy breaches (violations) is a meaningful exercise to guide discussions, pedagogy, and future research efforts. The application of a topical systematic review of the literature has created a knowledge base that can help researchers identify areas

ripe for future research. The research presented a systematic review of the literature to build a corpus of 29 research articles that resulted in a knowledge base related to privacy breach classifications. A novel taxonomy for privacy breach classification was also introduced to quantify related research trends. Research in this domain is on an upward trend, with a boost occurring in 2018. Analysis reveals that classifying and detecting data breaches has dominated research efforts over the last ten years. Detecting Breaches has remained consistent throughout this time frame, while Breach Classification has received more recent attention. Predicting Breaches and Risk analysis were revealed as emerging trends. In addition, word and phrase frequency analysis has identified the following gaps: Android Apps, Data Provider Violations, Location Privacy, Prediction Models, HealthCare Data Breaches, and Privacy Concerns. Emerging trends and identified gaps provide many avenues and opportunities for future work.

REFERENCES

- [1] C. D. Hylender, P. Langlois, A. Pinto, and S. Widup, *2024 data breach investigations report*, 2024. [Online]. Available: <https://www.verizon.com/business/resources/T22a/reports/2024-dbir-data-breach-investigations-report.pdf> (visited on 04/01/2025).
- [2] R. Page, *5 key takeaways from verizon's 2024 data breach investigations report*, May 2024. [Online]. Available: <https://www.csoononline.com/article/2096991/5-key-takeaways-from-verizons-2024-data-breach-investigations-report.html> (visited on 04/01/2025).
- [3] A. L. Allen, "Protecting one's own privacy in a big data economy," *Harv. L. Rev. F.*, vol. 130, p. 71, 2016.
- [4] B. Wolford, *What is GDPR, the EU's new data protection law?* [Online]. Available: <https://gdpr.eu/what-is-gdpr/> (visited on 04/01/2025).
- [5] M. Hill and S. Sharma, *The biggest data breach fines, penalties, and settlements so far*, Mar. 2025. [Online]. Available: <https://www.csoononline.com/article/567531/the-biggest-data-breach-fines-penalties-and-settlements-so-far.html> (visited on 04/01/2025).
- [6] M. Pedrosa, C. Costa, and J. Dorado, "GDPR impacts and opportunities for computer-aided diagnosis guidelines and legal perspectives," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2019, pp. 616–621. DOI: [10.1109/CBMS.2019.00128](https://doi.org/10.1109/CBMS.2019.00128).
- [7] A. J. Cain, *Taxonomy*, Feb. 2025. [Online]. Available: <https://www.britannica.com/science/taxonomy> (visited on 04/01/2025).
- [8] K. D. Fiedler, V. Grover, and J. T. Teng, "An empirically derived taxonomy of information technology structure and its relationship to organizational structure," *Journal of Management Information Systems*, vol. 13, no. 1, pp. 9–34, 1996. DOI: [10.1080/07421222.1996.11518110](https://doi.org/10.1080/07421222.1996.11518110).
- [9] R. L. Glass and I. Vessey, "Contemporary application-domain taxonomies," *IEEE Software*, vol. 12, no. 4, pp. 63–76, 1995. DOI: [10.1109/52.391837](https://doi.org/10.1109/52.391837).
- [10] R. C. Nickerson, U. Varshney, and J. Muntermann, "A method for taxonomy development and its application in information systems," *European Journal of Information Systems*, vol. 22, no. 3, pp. 336–359, 2013. DOI: [10.1057/ejis.2012.26](https://doi.org/10.1057/ejis.2012.26).
- [11] J. G. Miller and A. V. Roth, "A taxonomy of manufacturing strategies," *Management Science*, vol. 40, no. 3, pp. 285–304, 1994. DOI: [10.1287/mnsc.40.3.285](https://doi.org/10.1287/mnsc.40.3.285).
- [12] D. Klapser and E. Hovy, "A taxonomy and a knowledge portal for cybersecurity," *Proceedings of the 15th Annual International Conference on Digital Government Research*, pp. 79–85, Jun. 2014. DOI: [10.1145/2612733.2612759](https://doi.org/10.1145/2612733.2612759).
- [13] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, vol. 13, no. 3, pp. 334–359, 2002. DOI: [10.1287/isre.13.3.334.81](https://doi.org/10.1287/isre.13.3.334.81).
- [14] S. Warren and L. Brandeis, "The right to privacy," *Harvard Law Review*, vol. 4, no. 5, pp. 172–183, 1890.

- [15] D. J. Solove, "The meaning and value of privacy," *Social dimensions of privacy: Interdisciplinary perspectives*, pp. 71–81, 2015. DOI: [10.1017/CBO9781107280557.005](https://doi.org/10.1017/CBO9781107280557.005).
- [16] B. Auxier and L. Rainie, "Key takeaways on americans' views about privacy, surveillance and data-sharing," Pew Research Center, 2019.
- [17] B. C. Boatwright and C. White, "Is privacy dead? does it matter? how facebook frames its data policy through public communication," *The Journal of Public Interest Communications*, vol. 4, no. 1, pp. 78–78, 2020. DOI: [10.32473/jpic.v4.i1.p78](https://doi.org/10.32473/jpic.v4.i1.p78).
- [18] A. H. Seh, M. Zarour, M. Alenezi, *et al.*, "Healthcare data breaches: Insights and implications," *Healthcare*, vol. 8, no. 2, 2020. DOI: [10.3390/healthcare8020133](https://doi.org/10.3390/healthcare8020133).
- [19] S. B. Wikina, "What caused the breach? an examination of use of information technology and health data breaches," *Perspectives in health information management*, vol. 11, no. Fall, 1h, 2014.
- [20] D. E. Bambauer, "Privacy versus security," *J. Crim. L. & Criminology*, vol. 103, no. 3, pp. 667–683, 2013.
- [21] J. Bhattacharya, R. Dass, V. Kapoor, and S. Gupta, "Utilizing network features for privacy violation detection," in *2006 1st International Conference on Communication Systems Software & Middleware*, IEEE, 2006, pp. 1–10. DOI: [10.1109/COMSWA.2006.1665184](https://doi.org/10.1109/COMSWA.2006.1665184).
- [22] M. Jouini, L. B. A. Rabai, and A. B. Aissa, "Classification of security threats in information systems," *Procedia Computer Science*, vol. 32, pp. 489–496, 2014. DOI: [10.1016/j.procs.2014.05.452](https://doi.org/10.1016/j.procs.2014.05.452).
- [23] L. de la Torre, *GDPR matchup: The California consumer privacy act 2018*, IAPP, Jul. 2018. [Online]. Available: <https://iapp.org/news/agdpr-matchup-california-consumer-privacy-act/> (visited on 04/01/2025).
- [24] K. Eich Drummond and A. Alison Reyes, in *Nutrition Research: Concepts and Applications*, 2nd ed. Jones & Bartlett Learning, 2018, ch. Quantitative research designs: Experimental, quasi-experimental, and descriptive, pp. 155–183.
- [25] L. T. Ayres, C. M. Curtin, and T. A. Ng, "Standardizing breach incident reporting: Introduction of a key for hierarchical classification," in *2010 Fifth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, IEEE, 2010, pp. 79–83. DOI: [10.1109/SADFE.2010.19](https://doi.org/10.1109/SADFE.2010.19).
- [26] M. Banerjee, R. Karimi Adl, L. Wu, and K. Barker, "Quantifying privacy violations," in *Secure Data Management: 8th VLDB Workshop (SDM 2011)*, Springer, Seattle, WA, 2011, pp. 1–17. DOI: [10.1007/978-3-642-23556-6_1](https://doi.org/10.1007/978-3-642-23556-6_1).
- [27] P. Rani, A. K. Nag, and R. Shahriyar, "A data-driven classification framework for cybersecurity breaches," *IT Professional*, vol. 26, no. 2, pp. 39–48, 2024. DOI: [10.1109/MITP.2024.3374096](https://doi.org/10.1109/MITP.2024.3374096).
- [28] L. S. Flak, Ø. Sæbø, and P. Spagnoletti, "Privacy violations in light of digital transformation: Insights from data breaches in norway," in *Workshop on Information Security and Privacy (WISP 2019)*, 2019.
- [29] R. Gore, "Privacy breach: A concern of cloud integrated iot framework for smart city," in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 2018, pp. 26–27. DOI: [10.2139/ssrn.3170523](https://doi.org/10.2139/ssrn.3170523).
- [30] V. Shah, Z. Peng, G. Malla, and N. Niu, "Towards norm classification: An initial analysis of HIPAA breaches," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, IEEE, 2021. DOI: [10.1109/REW53955.2021.00074](https://doi.org/10.1109/REW53955.2021.00074).
- [31] J. R. Goh, S. S. Wang, Y. Harel, and G. Toh, "Predictive taxonomy analytics (LASSO): Predicting outcome types of cyber breach," *Journal of Cybersecurity*, vol. 9, no. 1, 2023. DOI: [10.1093/cybsec/tyad015](https://doi.org/10.1093/cybsec/tyad015).
- [32] B. Blakely, J. Kurtenbach, and L. Nowak, "Exploring the information content of cyber breach reports and the relationship to internal controls," *International Journal of Accounting Information Systems*, vol. 46, 2022. DOI: [10.1016/j.accinf.2022.100568](https://doi.org/10.1016/j.accinf.2022.100568).
- [33] T. A. L. Ali, M. H. Khafagy, and M. Farrag, "Big data challenges: Preserving techniques for privacy violations," *J Theor Appl Inf Technol*, vol. 100, no. 8, pp. 2505–2517, 2022.
- [34] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, and M. Curado, "Using NLP and machine learning to detect data privacy violations," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, 2020. DOI: [10.1109/INFOCOMWKSHPS50562.2020.9162683](https://doi.org/10.1109/INFOCOMWKSHPS50562.2020.9162683).
- [35] N. Kökciyan and P. Yolum, "PriGuardTool: A web-based tool to detect privacy violations semantically," in *Engineering Multi-Agent Systems: 4th International Workshop, (EMAS 2016)*, Springer, 2016. DOI: [10.1007/978-3-319-50983-9_5](https://doi.org/10.1007/978-3-319-50983-9_5).
- [36] Q. Luo, Y. Yu, J. Liu, and A. Benslimane, "Automatic detection for privacy violations in Android applications," *IEEE Internet of Things Journal*, vol. 9, no. 8, 2021. DOI: [10.1109/JIOT.2021.3109785](https://doi.org/10.1109/JIOT.2021.3109785).
- [37] M. Poikela and S. Möller, "The role of privacy violations in privacy calculus," in *European Workshop on Usable Security (EuroUSEC-2018)*, London, England, 2018. DOI: [10.14722/eurosec.2018.23015](https://doi.org/10.14722/eurosec.2018.23015).
- [38] S. Puglisi, D. Rebollo-Monedero, and J. Forné, "Potential mass surveillance and privacy violations in proximity-based social applications," in *2015 IEEE Trustcom/BigDataSE/ISPA*, IEEE, vol. 1, 2015, pp. 1045–1052. DOI: [10.1109/Trustcom.2015.481](https://doi.org/10.1109/Trustcom.2015.481).
- [39] P. Umar, A. Squicciarini, and S. Rajtmajer, "A study of self-privacy violations in online public discourse," in *International Conference on Big Data (Big Data-20)*, IEEE, 2020. DOI: [10.1109/BigData50022.2020.9378163](https://doi.org/10.1109/BigData50022.2020.9378163).
- [40] C. Islam, M. A. Babar, R. Croft, and H. Janicke, "SmartValidator: A framework for automatic identification and classification of cyber threat data," *Journal of Network and Computer Applications*, vol. 202, 2022. DOI: [10.1016/j.jnca.2022.103370](https://doi.org/10.1016/j.jnca.2022.103370).
- [41] K. K. Raghunath, V. V. Kumar, M. Venkatesan, K. K. Singh, T. Mahesh, and A. Singh, "Xgboost regression classifier (XRC) model for cyber attack detection and classification using inception v4," *Journal of Web Engineering*, vol. 21, no. 4, 2022. DOI: [10.13052/jwe1540-9589.21413](https://doi.org/10.13052/jwe1540-9589.21413).
- [42] M. H. M. Pakhari, N. Jamil, M. E. Rusli, and A. A. A. Rahim, "Implementation of token parsing technique for regex based classification of unstructured data for cyber threat analysis," in *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*, IEEE, 2020, pp. 395–398. DOI: [10.1109/ICIMU49871.2020.9243415](https://doi.org/10.1109/ICIMU49871.2020.9243415).
- [43] F. Meng, F. Lou, Y. Fu, and Z. Tian, "Deep learning based attribute classification insider threat detection for data security," in *2018 IEEE third international conference on data science in cyberspace (DSC)*, IEEE, 2018, pp. 576–581. DOI: [10.1109/DSC.2018.00092](https://doi.org/10.1109/DSC.2018.00092).
- [44] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and ubiquitous computing*, vol. 18, 2014. DOI: [10.1007/s00779-012-0633-z](https://doi.org/10.1007/s00779-012-0633-z).
- [45] A. Kumari, P. Prakash, and M. Umadevi, "Prediction of data breaches using classification algorithms," in *2021 Fifth International Conference on IoT in Social, Mobile, Analytics and Cloud(I-SMAC)*, IEEE, 2021. DOI: [10.1109/I-SMAC52330.2021.9640844](https://doi.org/10.1109/I-SMAC52330.2021.9640844).
- [46] X. Li, C. Zhao, Y. Hu, H. Xie, Y. Wang, and J. Zhao, "Precursor of privacy leakage detection for individual user," *Computers & Security*, vol. 142, 2024. DOI: [10.1016/j.cose.2024.103879](https://doi.org/10.1016/j.cose.2024.103879).
- [47] M. M. Uddin, R. Surendran, G. Gokul, and T. Thomas, "A data flow-based approach for classification and risk estimation of Android apps," in *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, IEEE, 2023, pp. 1–6. DOI: [10.1109/RASSE60029.2023.10363478](https://doi.org/10.1109/RASSE60029.2023.10363478).
- [48] X. Liu, Y. Song, S. Liu, and H. Wang, "Automatic taxonomy construction from keywords," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1433–1441.
- [49] M. Windl, V. Winterhalter, A. Schmidt, and S. Mayer, "Understanding and mitigating technology-facilitated privacy violations in the physical world," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16. DOI: [10.1145/3544548.3580909](https://doi.org/10.1145/3544548.3580909).
- [50] C. Tran, B. Reich, and H. Yuan, "When and why consumers respond to online privacy violations," in *Hawaii International Conference on System Sciences 2023 (HICSS-56)*, 2023.
- [51] G. Ayoade, S. Chandra, L. Khan, K. Hamlen, and B. Thuraishingham, "Automated threat report classification over multi-source data," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, IEEE, 2018. DOI: [10.1109/CIC.2018.00040](https://doi.org/10.1109/CIC.2018.00040).
- [52] A. Panou, C. Ntantogian, and C. Xenakis, "RiSKI: A framework for modeling cyber threats to estimate risk for data breach insurance," in *Proceedings of the 21st Pan-Hellenic Conference on Informatics*, 2017, pp. 1–6. DOI: [10.1145/3139367.313942](https://doi.org/10.1145/3139367.313942).