https://doi.org/10.58086/r7pn-1f89

SECURITY AND PRIVACY IN EXPLAINABLE AI: A BIBLIOMETRIC ANALYSIS OF EMERGING LEAKAGE RISKS

Mafalda Matos[1]✉, **Mário Dias Lousã**[1,2], **José Carlos Morais**[1,3]

[1] Instituto Superior Politécnico Gaya (ISPGAYA), Portugal.
[2] Insight - Piaget Research Center for Ecological Human Developement, Portugal.
[3] CEOS.PP, ISCAP, Polytechnic of Porto, Portugal.
✉ Corresponding authors**:** ispg2024103142@ispgaya.pt

Abstract

Explainable Artificial Intelligence (XAI) has gained increasing attention as a means of improving the transparency and trustworthiness of machine learning algorithms, particularly in domains where security and privacy concerns are relevant. This study presents a bibliometric analysis of research at the intersection of explainable artificial intelligence, security, and privacy. The aim was to characterize publication trends, thematic structures, and keyword relationships within the field. Scholarly records were retrieved from the Lens database using a structured search strategy based on the PRISMA protocol and analyzed using bibliometric tools, including Bibliometrix and VOSviewer. The total number of studies analyzed was 8,099, and the analyzed time frame was 2010–2025. The analysis examined general publication information, annual scientific production, leading publication venues, and keyword co-occurrence networks. Results indicate a rapid growth in XAI-related publications in recent years and reveal several major thematic clusters, including deep learning–driven medical imaging applications, foundational machine learning and data science concepts, explainability methods in security and distributed learning contexts, and governance-oriented themes related to ethics, privacy, and trust. Overall, the findings highlight the application-driven and interdisciplinary nature of explainable AI research, while showing that security and privacy topics, although present, remain relatively peripheral within the broader XAI literature.

Keywords: Explainable Artificial Intelligence; Bibliometric Analysis; Cybersecurity; Information Leakage; Machine Learning; Data Science; Privacy-Preserving ML.

1. Introduction

The increasing adoption of machine learning in security-sensitive and high-stakes domains has intensified interest in explainable artificial intelligence (XAI) as a means of improving transparency, trust, and accountability in automated decision-making systems (Doshi-Velez & Kim, 2017). In areas such as cybersecurity and related application domains, explainability is frequently promoted to support analyst decision-making, facilitate regulatory compliance, and enhance the interpretability of complex detection models (Capuano et al., 2022). As a result, research at the intersection of XAI, security, and privacy has expanded rapidly over the past decade.

At the same time, previous studies have revealed potential conflicts between explainability and security. By revealing information about model behavior or internal decision logic, explanation mechanisms may inadvertently expose additional information that could be exploited by adversaries, raising concerns related to privacy and information leakage (Shokri et al., 2017). While these issues are increasingly acknowledged, existing research remains dispersed across multiple disciplines and application domains, complicating efforts to obtain a consolidated view of how security and privacy concerns are positioned within the broader XAI literature.

In this context, bibliometric analysis offers a systematic and quantitative approach to mapping publication patterns, thematic structures, and research trends across large and interdisciplinary bodies of literature (Donthu et al., 2021).

Accordingly, this study conducts a bibliometric analysis of scholarly research at the intersection of explainable artificial intelligence, security, and privacy using data retrieved from the Lens database. The analysis examines publication growth, source distribution, and keyword co-occurrence patterns to characterize the conceptual structure of the field, with results presented in section four and discussed in section five. The objective of this work is not to evaluate specific algorithms or security mechanisms but to provide a high-level overview of research trends and thematic emphasis that can support future systematic reviews and empirical investigations.

To guide the bibliometric analysis, this study is structured around the following research questions:

RQ1: How has scientific production related to explainable artificial intelligence, security, and privacy evolved over time?

RQ2: Which journals and publication venues have contributed most prominently to research at the intersection of explainable AI, security, and privacy?

RQ3: How are security- and privacy-related themes positioned within the broader explainable AI research landscape?

2. Conceptualization

2.1. Lightweight Machine Learning in Security Applications

Lightweight machine learning models are widely adopted in security-sensitive and resource-constrained environments due to their low computational overhead and relatively transparent decision structures. In application domains such as intrusion detection, malware analysis, and Internet of Things (IoT) security, models including logistic regression, decision trees, random forests, and support vector machines are commonly employed to balance predictive performance with deployment feasibility on edge and embedded systems (Buczak & Guven, 2016; Khraisat et al., 2019). Their continued prevalence is driven by practical constraints related to processing power, memory, energy consumption, and the need for timely operational decisions (Shi et al., 2016). From an explainability perspective, the structural simplicity of these models facilitates post-hoc interpretation and transparency, making them frequent baselines in explainable artificial intelligence research within applied security and privacy contexts (Molnar, 2022).

2.2. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) encompasses methods and techniques aimed at making machine learning model outputs more transparent and interpretable to human users, particularly as such systems are increasingly deployed in high-stakes domains (Floridi et al., 2018; Gunning et al., 2019). Existing approaches are commonly distinguished between model-intrinsic methods and post-hoc explanation techniques applied to black-box models, with explanations further characterized as global or local in scope (Guidotti et al., 2018). Widely adopted post-hoc methods include feature attribution and surrogate modeling approaches that approximate model behavior to support human understanding (Ribeiro et al., 2016). In applied contexts such as healthcare, finance, and cybersecurity, explainability is

typically framed as a mechanism for supporting human oversight, regulatory compliance, and operational trust rather than as a replacement for automated decision-making (Samek et al., 2021). From a bibliometric perspective, the diversity of definitions, methods, and application domains has contributed to a fragmented and interdisciplinary research landscape.

2.3. Explainable Artificial Intelligence in Security and Privacy Contexts

Explainable artificial intelligence has attracted growing attention in security and privacy-sensitive domains, where understanding and justifying automated decisions is often as critical as predictive accuracy (Adadi & Berrada, 2018). In cybersecurity applications, XAI techniques are frequently employed to support intrusion detection, malware classification, anomaly detection, and fraud analysis by providing interpretable insights into model behavior for analysts and decision-makers (Amershi et al., 2014; Tramèr et al., 2016). While explainability is primarily motivated by practical considerations such as alert validation, trust-building, and operational transparency, its integration into machine learning pipelines also raises potential security and privacy concerns. Prior work has shown that explanation outputs may inadvertently expose information about model internals or training data, particularly in adversarial or untrusted environments (Tramèr et al., 2016). From a bibliometric standpoint, this dual role of XAI, as both a transparency-enhancing mechanism and a potential source of additional risk, contributes to the dispersion of security and privacy-related themes across multiple research clusters.

2.4. Conceptual Positioning and Scope of the Bibliometric Analysis

The combined proliferation of lightweight machine learning models, the methodological diversity of explainable artificial intelligence, and the increasing use of XAI in security and privacy-sensitive contexts have resulted in a rapidly expanding and heterogeneous body of research spanning multiple disciplines and application domains (Adadi & Berrada, 2018; Samek et al., 2021). While prior studies provide valuable conceptual frameworks and application-specific insights, the fragmentation of literature complicates efforts to identify dominant themes and broader research trajectories. In this context, bibliometric analysis offers a

complementary, high-level perspective by enabling systematic mapping of publication patterns, thematic structures, and the relative positioning of security and privacy concerns within the broader explainable AI landscape (Donthu et al., 2021). Rather than evaluating individual methods, this approach supports the identification of research trends and emerging gaps across interdisciplinary bodies of work.

3. Methodology

A bibliometric analysis was conducted to map the structure and evolution of research at the intersection of explainable artificial intelligence, security, and privacy. The analysis followed standard bibliometric procedures, including corpus construction, descriptive indicator analysis, and keyword co-occurrence mapping, and was conducted in December of 2025.

More specifically, the parameters used were:

- Years: 2010–2025.
- Language: English.
- Document types: Articles, Reviews, Conference papers.
- Database: The Lens.
- Documents analyzed: All that fit the previous criteria.

As Lens does not provide a dedicated language metadata field, English-language publications were identified through abstract-based filtering, followed by manual validation of a random sample. This approach is consistent with prior bibliometric studies using Lens.

These were the exact search parameters used:

Scholarly Works (8,198) = (("explainable AI" OR (XAI OR ("interpretable machine learning" OR "model interpretability"))) AND (“security” OR (“cybersecurity” OR (privacy OR "information leakage"))))

Filters: Year Published = (2010 -)

Publication Type = (journal article , review , journal issue , conference proceedings , journal , conference proceedings article) Field of Study = (Computer science , Artificial intelligence , Machine learning , Data science , Computer security)

A PRISMA-inspired workflow (Figure 1) was employed to transparently report the identification and preprocessing of records for bibliometric analysis. The workflow reflects

automated filtering and metadata-based preprocessing steps. No further steps were taken for manual verification.

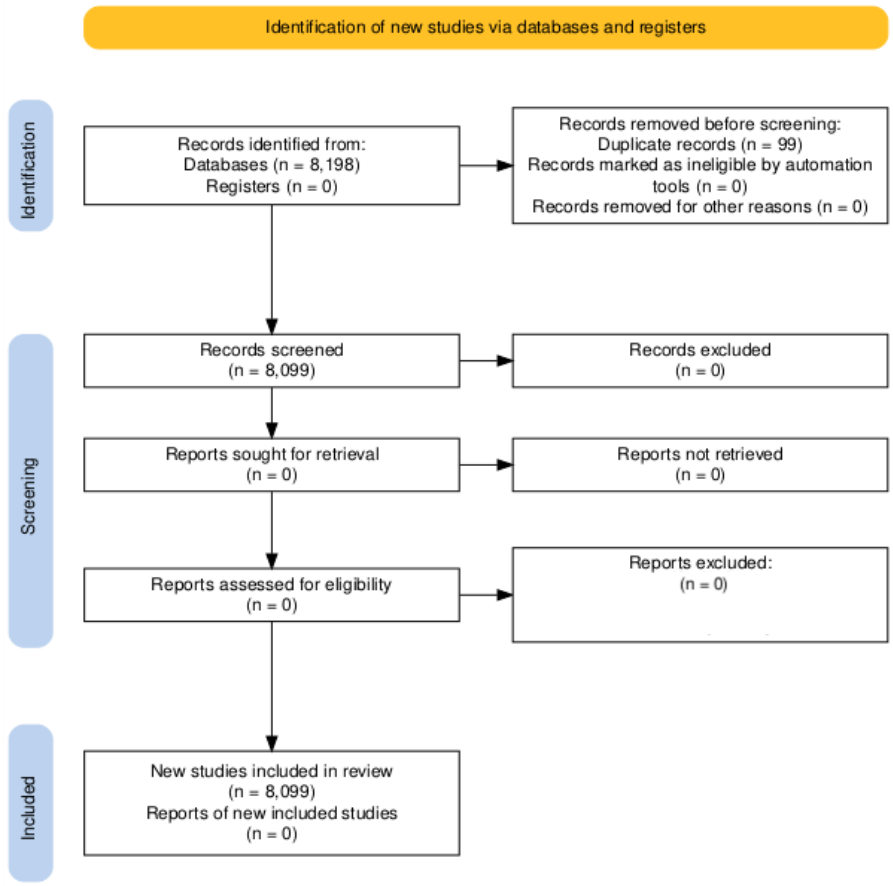

Fig 1. PRISMA-inspired workflow schema
Source: Own generation using https://estech.shinyapps.io/prisma_flowdiagram/

4. Results and Discussion

4.1. Overall Dataset Characteristics

The final bibliometric corpus comprises 8,099 publications indexed in the Lens database, spanning the period 2010–2025. These publications are distributed across 2,669 distinct sources, reflecting the highly multidisciplinary nature of research at the intersection of explainable artificial intelligence, security, and privacy. The dataset exhibits a very high annual growth rate (59.7%). The average document age is 1.57 years, highlighting the strong recency bias of literature.

4.2. Temporal Evolution of Publications

Analysis of annual scientific production (Figure 2) reveals a slow emergence phase between 2010 and 2016, followed by a sharp acceleration starting in 2018. Publication counts increase significantly after 2019, with particularly strong growth observed from 2021 onward. The most pronounced rise occurs in 2024 and 2025, which together account for a substantial proportion of the total corpus.

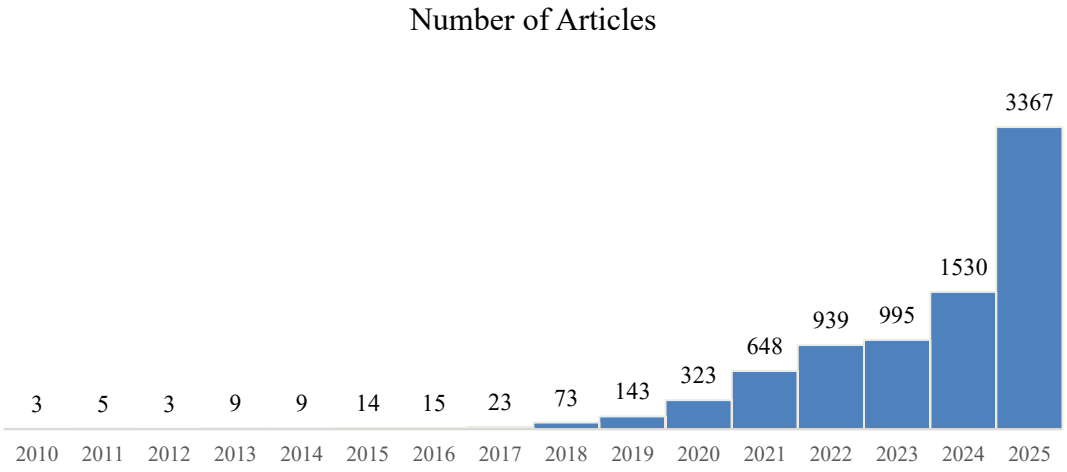

Fig 2. Evolution of the number of publications on explainable artificial intelligence, security, and privacy (2010–2025)
Source: Authors' elaboration based on Bibliometrix and python.

This surge coincides with the widespread adoption of deep learning systems and the increasing demand for explainability in high-stakes domains, suggesting that concerns related to transparency, accountability, and privacy have become central drivers of recent research activity.

4.3. Source Distribution and Publication Venues

The literature is highly dispersed across venues, with no single dominant outlet, further reinforcing the multidisciplinary character of the field (Table 1 and Figure 3). The sources with the highest number of publications address applied artificial intelligence and data-driven decision-making.

In terms of publications, the strong prevalence of biomedical, healthcare, and sensing-oriented journals among the top-ranked sources (Table 1 and Figure 3). This suggests that security and privacy issues related to XAI are most frequently investigated within data-sensitive application domains, such as medical diagnosis, health monitoring, and clinical decision support, rather than in traditional cybersecurity venues.

Table 1. Top 20 publication venues by number of documents in the analyzed corpus

Source	Number of Publications
Scientific Reports	606
Sensors	359
PLOS ONE	151
Diagnostics	146
IEEE Access	128
Frontiers in Artificial Intelligence	120
Applied Sciences	100
Journal of Medical Internet Research	95
PeerJ Computer Science	93
Bioengineering (Basel, Switzerland)	89

Source: Authors' elaboration based on data from The Lens database (2025)

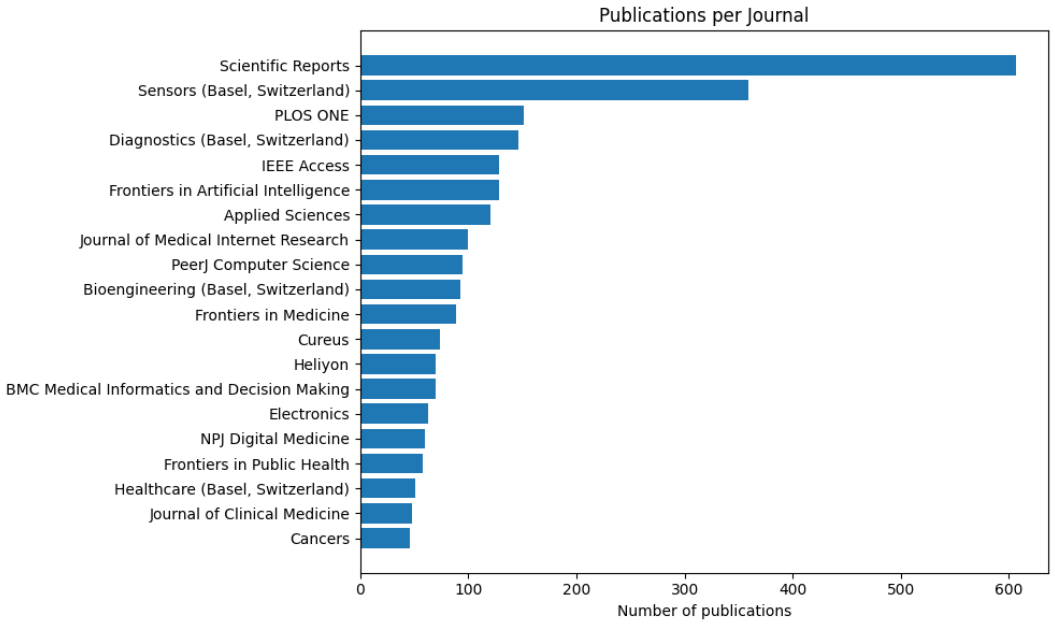

Fig 3. Top 20 publication venues by number of documents in the analyzed corpus
Source: Bibliometrix and python

4.4. Authorship and Collaboration Patterns

The dataset includes contributions from 35,922 unique authors, with an average of 5.08 co-authors per document, indicating a predominantly collaborative research culture. Single-authored publications account for only 887 documents, reflecting the interdisciplinary nature of XAI-related research.

The most productive authors are distributed across diverse institutional and disciplinary backgrounds, reflecting the absence of a small, centralized research community. Instead, contributions emerge from a broad range of applied AI, machine learning, and domain-specific research groups.

4.5. Keyword Coverage and Metadata Characteristics

Figure 4 presents the keyword co-occurrence network, with a minimum occurrence limit of 30 to balance thematic scope and readability, resulting in 78 keywords.

The analysis reveals several dominant application-driven clusters, particularly in medical imaging, healthcare analytics, and data-intensive biomedical domains. Explainable AI methods form a distinct cluster connected to anomaly detection, IoT, and intrusion detection, while security and privacy-related terms appear comparatively smaller and more peripheral. This suggests that, although security and privacy concerns are present in XAI research, they are not yet central organizing themes within the broader literature.

Given the highly fragmented and application-driven nature of the field, the analysis focuses on thematic and venue-level structures rather than individual author productivity.

A thematic analysis (not shown) further supports the findings of the keyword co-occurrence network. Dominant themes are centered on applied machine learning and explainability in healthcare and sensing contexts, while security and privacy-related themes appear as less central and less developed. This thematic configuration reinforces the observation that security considerations in XAI research remain secondary to application-driven concerns.

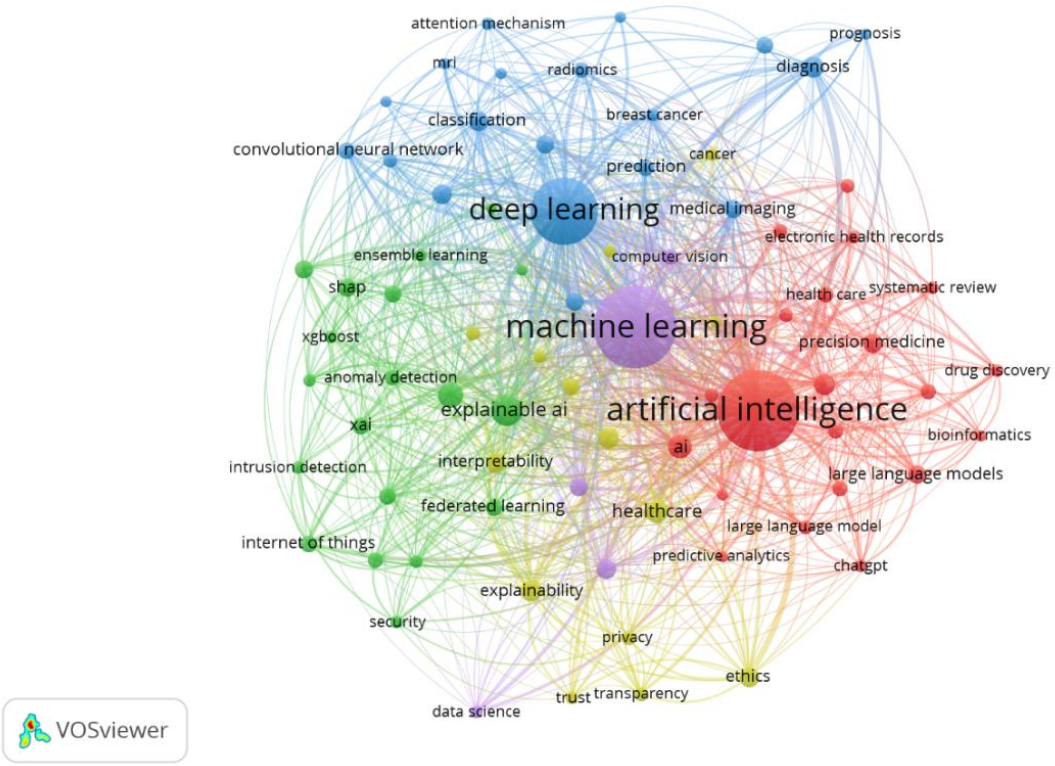

Fig 4. Keyword Co-occurrence Network Figure
Source: Own generation in VOSViewer

Purple Cluster: Core Machine Learning and Data Science Foundations

The purple cluster represents the methodological core of the explainable AI literature and is characterized by foundational terms such as machine learning, computer vision, and data science. The centrality and strong interconnectivity of these keywords indicate their role as shared technical foundations across diverse application areas rather than as a distinct domain of inquiry. The prominence of computer vision reflects the importance of visual data processing tasks within the corpus, particularly in contexts where model performance and interpretability intersect. Overall, this cluster functions as a structural backbone linking application-driven research with more specialized methodological and thematic clusters, including explainability techniques and security-related topics.

Red Cluster: General Artificial Intelligence and Data-Driven Application Domains

The red cluster captures a broad, application-oriented body of research centered on general artificial intelligence concepts and data-driven methodologies. Core keywords such as artificial intelligence and machine learning co-occur with terms associated with contemporary

AI paradigms, including large language models, predictive analytics, and generative systems. The strong presence of healthcare-related applications—such as precision medicine, bioinformatics, drug discovery, and electronic health records—highlights the dominant role of biomedical domains within the broader AI literature captured in the dataset. The inclusion of systematic review further suggests a maturing research stream characterized by rapid growth and increasing synthesis activity.

Blue Cluster: Deep Learning–Driven Medical Imaging and Diagnostic Applications

The blue cluster is dominated by deep learning–based medical imaging research, with prominent keywords including deep learning, convolutional neural networks, classification, and attention mechanisms. The strong association with clinical terms such as breast cancer, radiomics, MRI, diagnosis, and prognosis indicates a concentration on imaging-based decision-support systems in healthcare settings. Within this cluster, advanced neural architectures are commonly applied to high-dimensional visual data to support disease detection and outcome prediction. The presence of attention mechanisms reflects ongoing efforts to enhance both predictive performance and interpretability in clinically oriented deep learning applications.

Green Cluster: Explainable AI Methods in Security and Distributed Learning Contexts

The green cluster centers on explainable artificial intelligence methods applied in security and infrastructure-oriented contexts. Key terms such as explainable AI, XAI, and SHAP co-occur with intrusion detection, anomaly detection, and security, indicating the use of explanation techniques to support transparency and analyst interpretation in detection systems. The presence of ensemble methods and models such as XGBoost reflects a balance between predictive performance and post-hoc interpretability. In addition, keywords related to IoT and federated learning highlight the relevance of distributed and resource-constrained environments, where explainability is used to support trust and insight without centralizing sensitive data.

Yellow Cluster: Interpretability, Trust, and Ethical Considerations in Applied AI

The yellow cluster encompasses interpretability and governance-oriented themes, including explainability, privacy, ethics, trust, and transparency. These keywords reflect a body of literature concerned with the societal, regulatory, and operational implications of deploying

machine learning systems in sensitive contexts. The frequent association with healthcare-related terms underscores the prominence of ethical and trust-based discussions in clinical and biomedical applications, where regulatory compliance and patient safety are central concerns. Rather than emphasizing specific technical methods, this cluster highlights normative and governance perspectives that cut across application domains and connect explainability research with broader discussions of responsible and trustworthy AI.

4.6. Security and Information Leakage Implications of Explainable AI

The results indicate that research on explainable artificial intelligence at the intersection with security and privacy is predominantly application-driven and dispersed across multiple domains and publication venues. Although security- and privacy-related concerns are present throughout the analyzed corpus, they are primarily embedded within applied contexts—particularly in the healthcare and Internet of Things domains—rather than being articulated from explicitly adversarial or threat-model–centric perspectives. The bibliometric analysis reinforces this observation by showing that keywords associated with security and privacy do not play a structuring role in the thematic organization of the explainable artificial intelligence literature. Notably, terms related to concrete information leakage mechanisms and attack models occupy peripheral positions within the keyword co-occurrence network, suggesting that the potential of explainability mechanisms to introduce new attack surfaces has not yet been systematically addressed at the landscape level. This peripheral positioning highlights an emerging research gap at the intersection of explainability, security, and privacy, in which information leakage risks induced by explanation methods remain underexplored.

5. Conclusions, Limitations and Future Work

This study presented a bibliometric analysis of research intercepting explainable artificial intelligence, security, and privacy, aiming to map publication trends, thematic structures, and keyword relationships within this rapidly expanding field. By analyzing a large corpus of publications indexed in the Lens database, the study provides a structured overview of

how explainability-related research has evolved over time and how security and privacy-oriented topics are situated within the broader XAI literature.

The results indicate a sharp growth in scientific output in recent years, reflecting the increasing relevance of explainable AI across multiple application domains. Keyword co-occurrence and thematic analyses reveal that explainable AI research is strongly intertwined with general machine learning and AI concepts, with deep learning playing a central role. At the same time, security- and privacy-related topics appear as more peripheral but increasingly connected themes, often embedded within applied contexts rather than forming standalone research clusters.

Overall, this bibliometric mapping highlights the interdisciplinary and application-driven nature of explainable AI research while providing a high-level quantitative overview that can support more focused systematic reviews and empirical investigations.

This study has several limitations that should be considered when interpreting the results. First, the analysis relies on metadata retrieved from a single bibliographic database. Although the Lens platform offers broad coverage across disciplines, some relevant publications indexed exclusively in other databases may not be included.

Second, the bibliometric approach is inherently dependent on the quality and consistency of bibliographic metadata, particularly author keywords. Variations in terminology, keyword granularity, and indexing practices may influence the structure of the identified clusters and co-occurrence networks.

Third, bibliometric methods capture patterns of publication activity and thematic relationships but do not assess the methodological quality, empirical validity, or practical effectiveness of individual studies. As a result, the findings should be interpreted as indicative of research trends rather than as evaluative judgments about specific approaches or techniques.

Finally, the dominance of certain application domains, such as healthcare, reflects the distribution of the existing literature and does not imply that explainable AI research is inherently limited to these contexts. The scope of this work was kept general on purpose to capture the full landscape of this field, which necessarily means it emphasizes coverage over depth, and future work should explore domain-specific nuances in greater detail.

The findings also point to differentiated implications for public and private stakeholders. From a regulatory perspective, the prominence of ethics and governance-related themes suggests that transparency and explainability requirements should be coupled with explicit consideration of security and information leakage risks, particularly in adversarial or high-stakes

contexts. From a product development perspective, explainability mechanisms should be treated as potential risk vectors rather than purely trust-enhancing features, requiring integration into security and privacy-by-design practices. This distinction highlights the need to align ethical and governance-oriented discussions of explainable AI with concrete deployment and threat considerations.

Future research could extend this bibliometric analysis in several directions. First, integrating multiple bibliographic databases may provide a more comprehensive representation of the explainable AI research landscape and reduce potential database-specific biases.

Second, the descriptive insights presented in this study could serve as a foundation for systematic literature reviews focusing on specific subtopics, such as explainability techniques in adversarial settings, privacy-preserving explanations, and information leakage risks introduced by explainability mechanisms, particularly in security- and privacy-sensitive deployment contexts. Such reviews would enable deeper qualitative analyses that go beyond the scope of bibliometric mapping.

Third, longitudinal analyses examining the evolution of security and privacy-related themes over shorter time intervals could provide finer-grained insights into emerging research directions.

Finally, future work may explore the relationship between explainability research and regulatory or governance frameworks, particularly as AI becomes an increasingly prominent policy concern.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in CyberSecurity: A survey. *IEEE Access: Practical Innovations, Open Solutions*, *10*, 93575–93600. https://doi.org/10.1109/access.2022.3204171

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity, 2*(1), Article 20. https://doi.org/10.1186/s42400-019-0038-7

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Samek, W., Wiegand, T., & Müller, K.-R. (2021). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *IEEE Signal Processing Magazine, 38*(3), 40–48. https://doi.org/10.48550/arXiv.1708.08296

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal, 3*(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 3–18). https://doi.org/10.1109/SP.2017.41

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In *Proceedings of the 25th USENIX Security Symposium* (pp. 601–618).