THE USE OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN THE CYBER SECURITY: A MINI SYSTEMATIC REVIEW

Dr. Esra SÖĞÜT
ORCID: 0000-0002-0051-2271
esrasogut@gazi.edu.tr
Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Türkiye

Abstract

Cyber attacks that may occur with developing and diversifying technology also develop and update themselves at regular intervals. The issue of cyber security must be handled meticulously to protect data, which is the most valuable asset in the systems, and ensure the continuity of operations. The effects of cyber attacks can cause devastating consequences beyond predictions. The methods and systems used to detect or prevent cyber attacks, which can vary widely, are also changing and being developed. There are various methods to detect attacks. XAI - Explainable Artificial Intelligence technology is used to obtain meaningful results in detecting attacks, minimize existing uncertainties, and guide experts. This study scanned academic publications such as books, book chapters, articles, and conference papers related to using XAI for cyber security. While performing these procedures, the mini systematic review method is used. This study, aimes to provide guidance on how XAI is evaluated in the field of cyber security, the situation of Turkey and other countries in this field, the trends of the studies carried out, and what studies can be done in the future using the data obtained.

Keywords: Explainable Artificial Intelligence, XAI, cyber security, systematic review, academic

AÇIKLANABİLİR YAPAY ZEKANIN (XAI) SİBER GÜVENLİK ALANINDA KULLANIMI: MİNİ SİSTEMATİK TARAMA

Özet

Gelişen ve farklılaşan teknoloji ile beraber meydana gelebilecek siber saldırılar da gelişmekte ve kendini belirli aralıklarla güncellemektedir. Sistemlerdeki en değerli varlık olan verinin korunması ve sistemlerin işleyişlerinin sürekliliğinin sağlanması için siber güvenlik konusu titizlikle ele alınmalıdır. Siber saldırıların etkileri tahminlerin ötesinde yıkıcı sonuçlara sebep olabilmektedir. Çeşitlilik gösterebilen siber saldırıların tespitinde veya engellenmesinde kullanılan yöntemler ve sistemler de farklılaşmakta ve geliştirilmektedir. Saldırıları tespit etmek için çeşitli yöntemler bulunmaktadır. Saldırıların tespitinde daha anlamlı sonuçlar elde edilmesi, var olan belirsizliklerin en aza indirilmesi ve uzmanlara yol gösterici olunması için Açıklanabilir Yapay Zeka (XAI - Explainable Artificial Intelligence - AYZ) teknolojisinden faydalanılmaktadır. Bu çalışmada AYZ'nın siber güvenlik için kullanımı ile ilgili olan kitap, kitap bölümü, makale ve bildiri olarak yapılan akademik yayınlar

taranmaktadır. Bu işlemler yapılırken mini sistematik tarama yönteminden faydalanılmaktadır. Bu çalışmada AYZ'nın siber güvenlik alanında nasıl değerlendirildiği, bu alanda Türkiye'nin ve diğer ülkelerin ne durumda olduğu, yapılan çalışmaların yönelimleri ve elde edilen verilerden faydalanılarak ilerisi için hangi çalışmaların yapılabileceği hakkında yol gösterici olmak hedeflenmektedir.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, AYZ, siber güvenlik, sistematik tarama, akademik

1. Introduction

The advancement of technology has accelerated the initiation of novel and comprehensive studies in the realm of cyber security. The data dimensions utilized, processed, and recorded in systems are expanding. Additionally, the diversity and significance of data undergo variations. Different security structures exist for all transactions occurring in sectors such as banking, commerce, and transportation; institutions like municipalities; and critical infrastructures like electricity and water, each with its distinct security framework. Ensuring the cyber security of data and all related processes is imperative. Otherwise, problems such as data theft, modification, or sale may occur. These problems can lead to consequences beyond predictions, such as a city experiencing a complete blackout throughout the day [1].

Diversifying and expanding of the scope of cyber attacks pose challenges in their detection. Detecting cyber attacks is one of the pivotal steps in ensuring security. Various methods are employed to detect attacks, including rule-based, signature-based, machine learning-based, and artificial intelligence-based approaches,. With a slight enhancement of the capabilities of artificial intelligence methods, Explainable Artificial Intelligence (XAI) has emerged. XAI is preferred to produce more meaningful results in detecting attacks and to facilitate a clearer understanding of methods conducted in the background but unknown. The utilization of XAI aims to minimize uncertainties in the existing attack detection system in real-time and to guide researchers by providing a clearer insight into the processes [2,3].

This study analyzed the status of cyber security and Explainable Artificial Intelligence (XAI) publications using the mini systematic review method. Publications, including books, book chapters, articles, and conference papers, were systematically reviewed within the last ten years. All targeted processes were executed within the Scopus environment, which offers various search features. Details about these processes will be provided in subsequent sections. This study aims to assess XAI's evaluation the cyber security field, examine the current global developments, and make predictions regarding future advancements.

2. Background

The study was based on Explainable Artificial Intelligence and Cyber Security topics. Enriched searches have been conducted using various keywords. The objectives of this study are elucidated in the Motivation section.

2.1.Explainable Artificial Intelligence

One of the methods employed to extract information from artificial intelligence datasets is artificial intelligence itself. Artificial intelligence-based methods are also utilized in the detection of cyber attacks. A new concept has emerged to verify the reliability and confirm the capability of artificial intelligence methods. Explainable Artificial Intelligence contributes to achieving explainability in artificial intelligence, making the processes conducted in the background and the ongoing procedures accessible for evaluation by experts and researchers [4].

2.2.Cyber Security

All systems containing data have the potential to be exposed to cyber attacks. The rapid advancement of technology is also altering attack techniques. Smarter, more comprehensive, and faster attacks are being developed. Consequently, the effects of these attacks are becoming more hazardous. The primary step in mitigating the effects of attacks is to detect and recognize the attack. Therefore, continuous improvement of attack detection systems is a crucial area of focus. Incorporating of artificial intelligence methods as cyber attack techniques has introduced new perspectives to the field of cyber security [5]. Adding Explainable Artificial Intelligence to detection method studies will accelerate the resolution and understanding of results, addressing uncertainties in these domains.

2.3.Motivation

This study addresses topics related to cyber security and Explainable Artificial Intelligence to track universal developments. The objectives identified within the scope of the study and the questions to be answered in line with the employed methodology are as follows:

- What are the numbers of publications based on keyword search results?
- What are the distributions of publications over the years?
- What is the geographical distribution of publications by country?
- How are the distributions of publications based on types?
- How are the distributions of publications based on research areas?

What is the impact of the keywords used during the publication search?

3. Methodology

In this study, publications in the type of books, book chapters, articles, and conference papers related to Explainable Artificial Intelligence and cyber security have been systematically reviewed. The systematic review was conducted to obtain a comprehensive overview of the general state of the relationship between Explainable Artificial Intelligence and cyber security. Systematic review involves a comprehensive evaluation of studies encompassing similar topics and methods according to a specific methodological approach. The stages include the selection of similar studies, investigation of these studies, data collection related to these studies, and, additionally, steps for data analysis and synthesis [6,7]. The stages involved in the study are detailed in this section.

3.1.Selecting Similar Studies

In this stage, characteristics possessing keywords such as Explainable Artificial Intelligence, cyber security, and cyber attack have been identified. Publications containing these characteristics, specifically in the types of books, book chapters, articles, and conference papers, were selected, and these were investigated.

3.2.Obtaining Data

The Scopus database, a widely used scientific publication indexing database, was chosen for data collection in the present study. Scopus provides various filtering, querying, and classification options, encompassing keywords and aspects such as the title, subject, authors, or abstract of a study. Additionally, the Scopus database presents visual information related to the obtained search data [8].

3.3.Data Analysis and Synthesis

The search was conducted on the Scopus database based on specific keywords, leading to the identification of relevant publications. The keywords are as follows:

- explainable artificial intelligence
- explainable artificial intelligence and cyber security
- explainable artificial intelligence and cyber security and cyber attack
- explainable artificial intelligence and cyber security and attack detection system
- explainable artificial intelligence and cyber security and intrusion detection system

For the publications selected based on these keywords, the publication year range has been set for the last decade (2014-the first half of the year 2024). Publication types have also been specified. Other filtering, querying, or classification options are set to default. The data obtained from the Scopus platform in the study is presented as graphs and tables. Additionally, analyses were carried out on them.

4. Findings

The findings obtained according to the objectives outlined in the study's purpose have been presented in subheadings in this section.

4.1.Finding 1: Scanning based on the "explainable artificial intelligence" keyword

It was scanned according to the keyword 'explainable artificial intelligence'. The findings obtained (Finding 1) are included here.

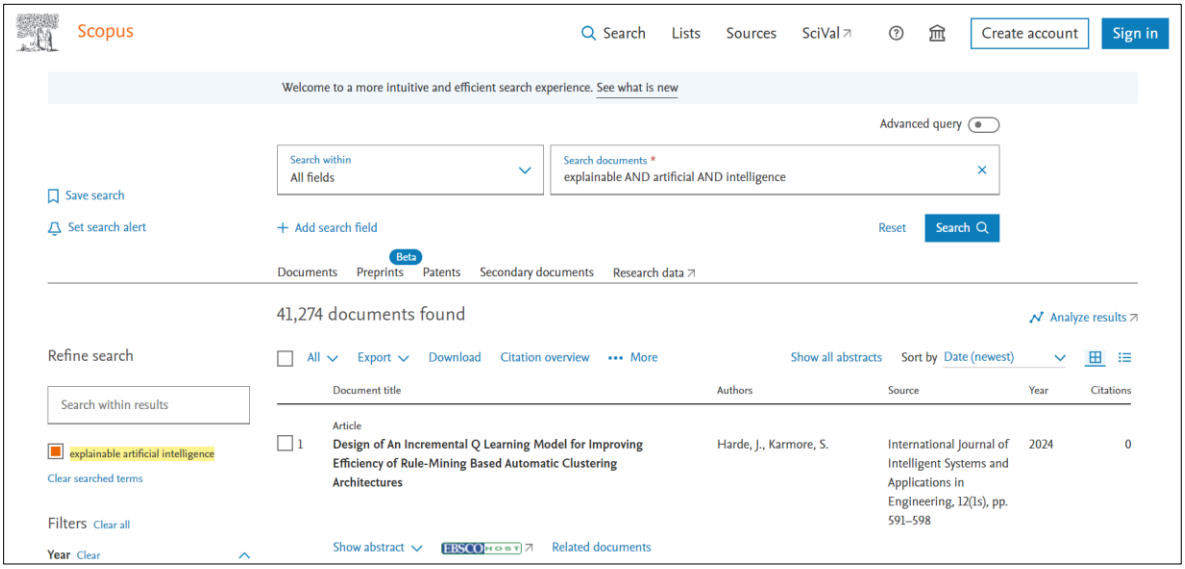


Figure 1. Finding 1: Total number of documents

When the scanning criteria were set, and a search was conducted on Scopus, the total number of documents on the topic of Explainable Artificial Intelligence was 41,274 (Figure 1).

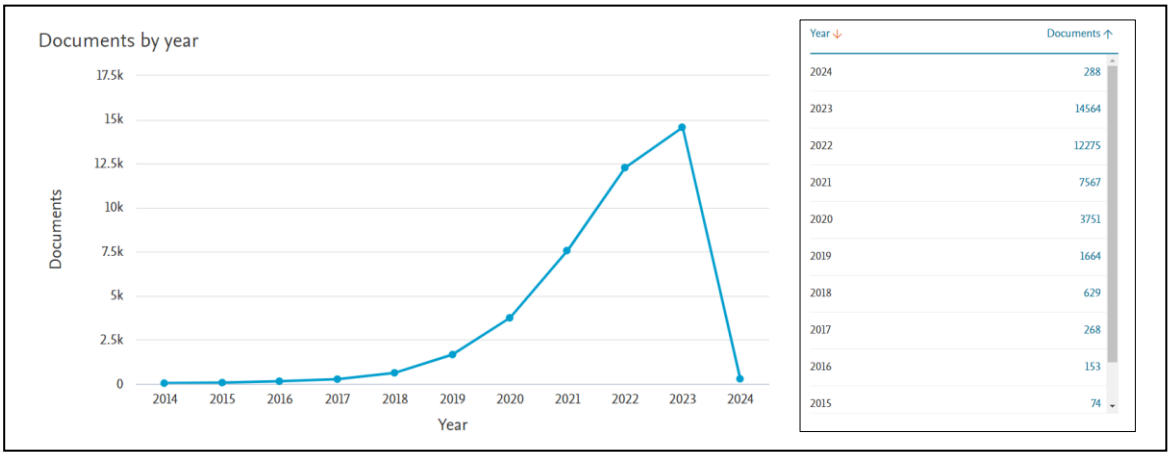


Figure 2. Finding 1: Distribution of documents by year

Accordingly, the number of documents has increased since 2018, reaching the highest number of publications (14,564) in 2023. Although 2024 has yet to begin, there are nearly 300 publications for that year. Academic studies on this subject have been conducted for the last six years.

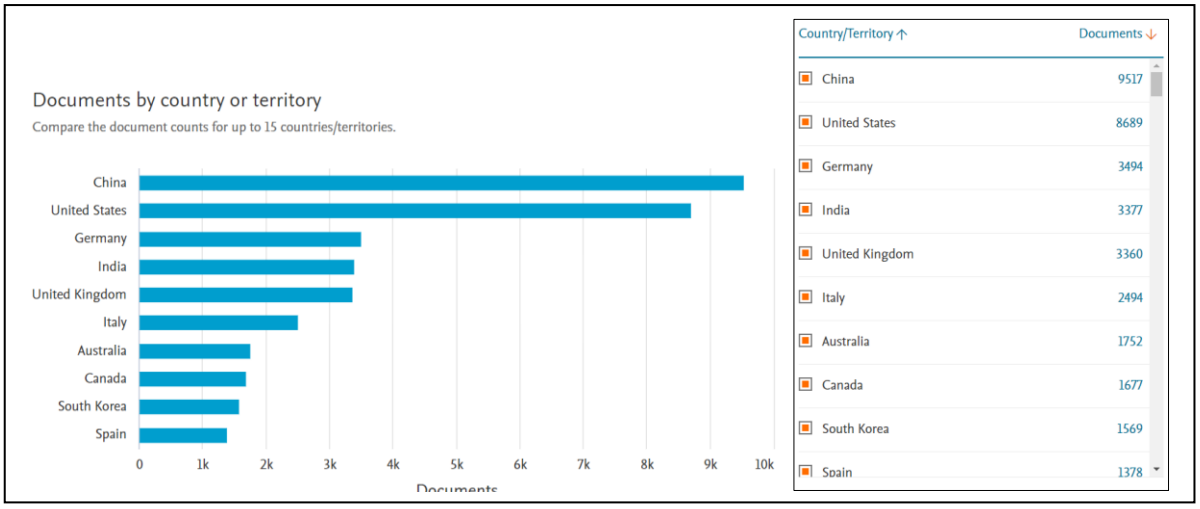


Figure 3. Finding 1: Distribution of documents by countries

The distribution of documents in this section according to the countries where they were conducted is presented in Figure 3. Upon examination of Figure 3, it is evident that China and the United States have a significantly higher number of studies than other countries. Germany, India, and the United Kingdom are observed following these two countries. The number of documents from Turkey needs to be improved, totaling 589.

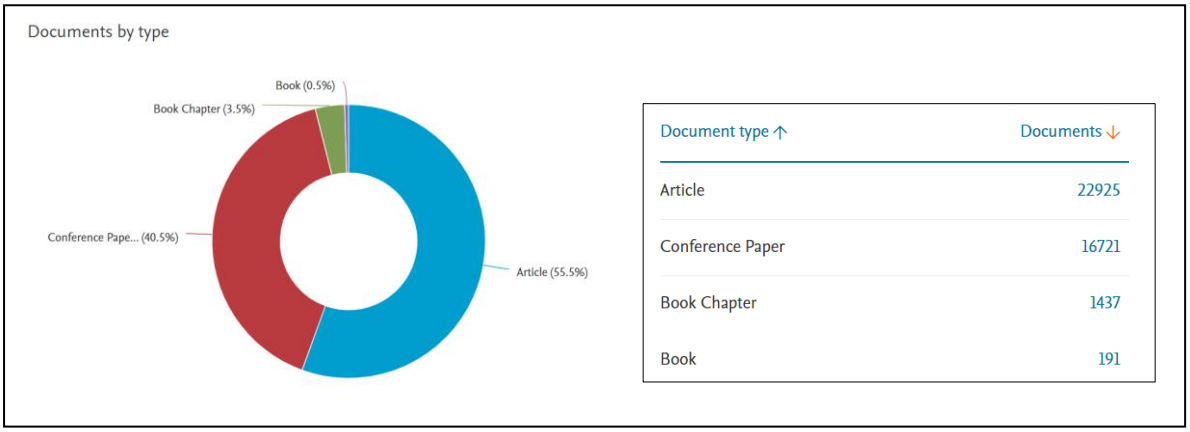


Figure 4. Finding 1: Distribution of documents by type

The types of documents to be scanned were specified in section 3.1, 'Selecting Similar Studies.' Upon examining Figure 4, it is observed that the 'Article' type of document is the most common. Following this, conference papers are the next most frequent. The least common type of study is book publications.

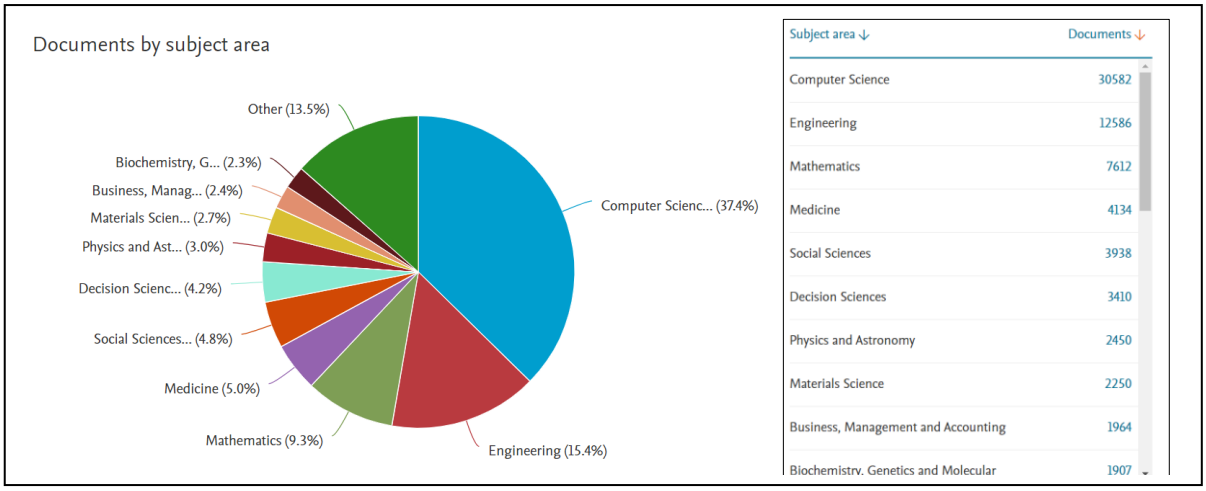


Figure 5. Finding 1: Distribution of documents by subject area

The final item related to Finding 1 is the determination of the fields or topics covered by the documents. Accordingly, Figure 5 has been obtained. The top three fields in terms of the number of documents are Computer Science, Engineering, and Mathematics.

4.2.Finding 2: Scanning based on "explainable artificial intelligence," and "cyber security " keywords

A search was conducted based on the keywords 'explainable artificial intelligence' and 'cyber security,' and the findings (Finding 2) obtained are presented here.

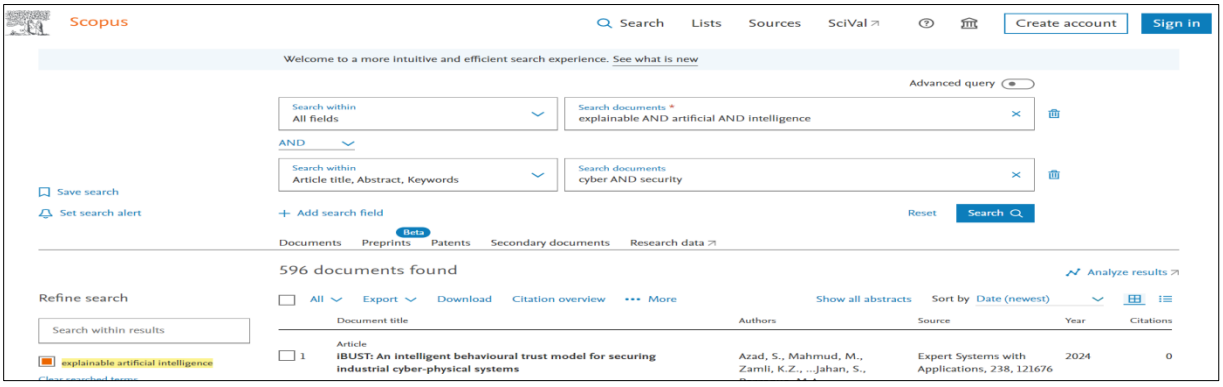


Figure 6. Finding 2: Total number of documents

When the search criteria were set, and a scan was performed on Scopus, the total number of documents on Explainable Artificial Intelligence and cyber security topics was 596 (Figure 6).

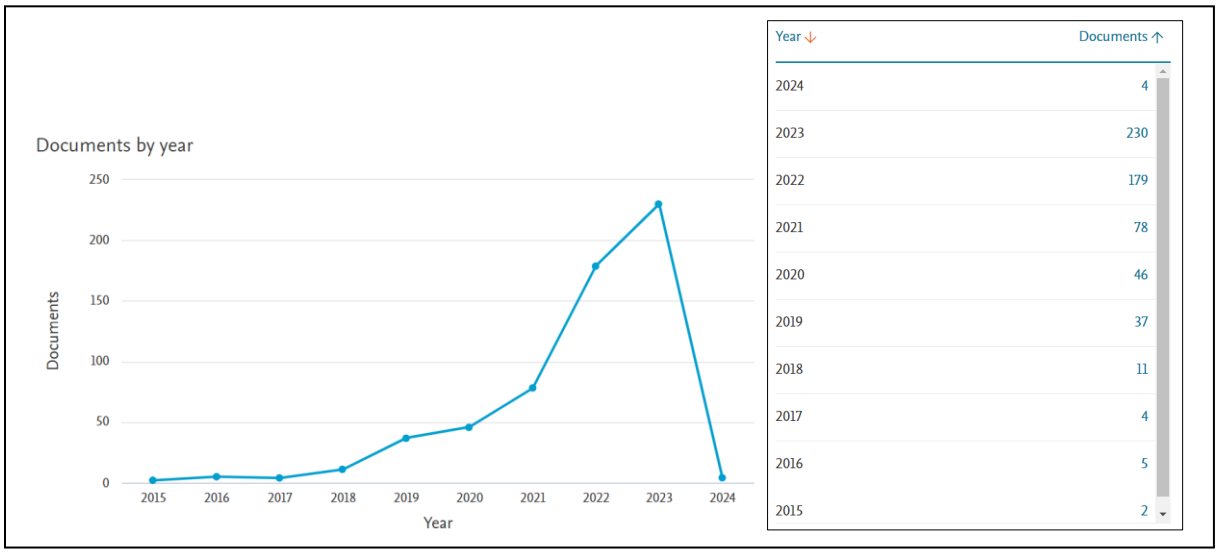


Figure 7. Finding 2: Distribution of documents by year

The distribution of documents by year is illustrated in Figure 7. According to this, the number of documents has been increased 2018, reaching the highest number (230) in 2023. Documents for the year 2024 have recently started. Similar to Finding 1, academic studies in this field have been conducted for six years.

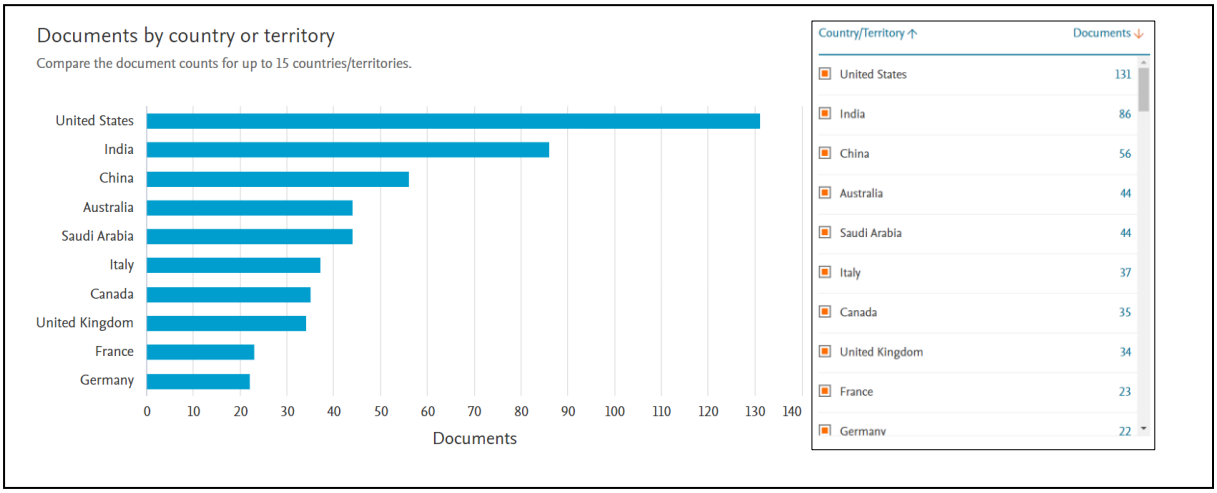


Figure 8. Finding 2: Distribution of documents by countries

The distribution of documents according to the countries in this section is presented in Figure 8. Unlike Finding 1, Australia and Saudi Arabia have entered the ranking. Upon examination of Figure 8, it is observed that the United States and India have more studies than other countries. The number of documents from Turkey is meager, totaling 18.

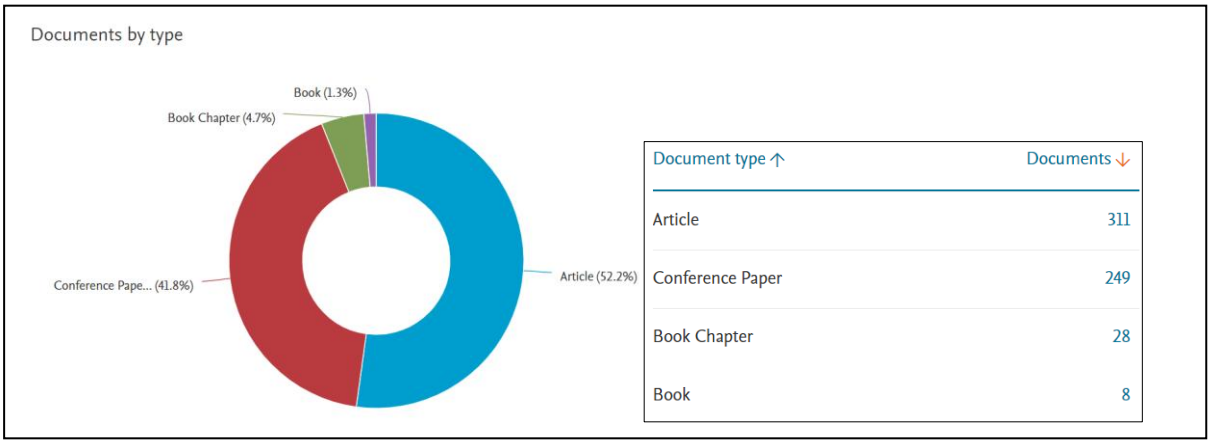


Figure 9. Finding 2: Distribution of documents by type

When examining the desired publication types based on the search criteria in Figure 9, it is observed that the 'Article' type of study is once again the most prevalent. Subsequently, conference papers, book chapters, and book publications follow. The least prevalent type of document remains the same: book publications.

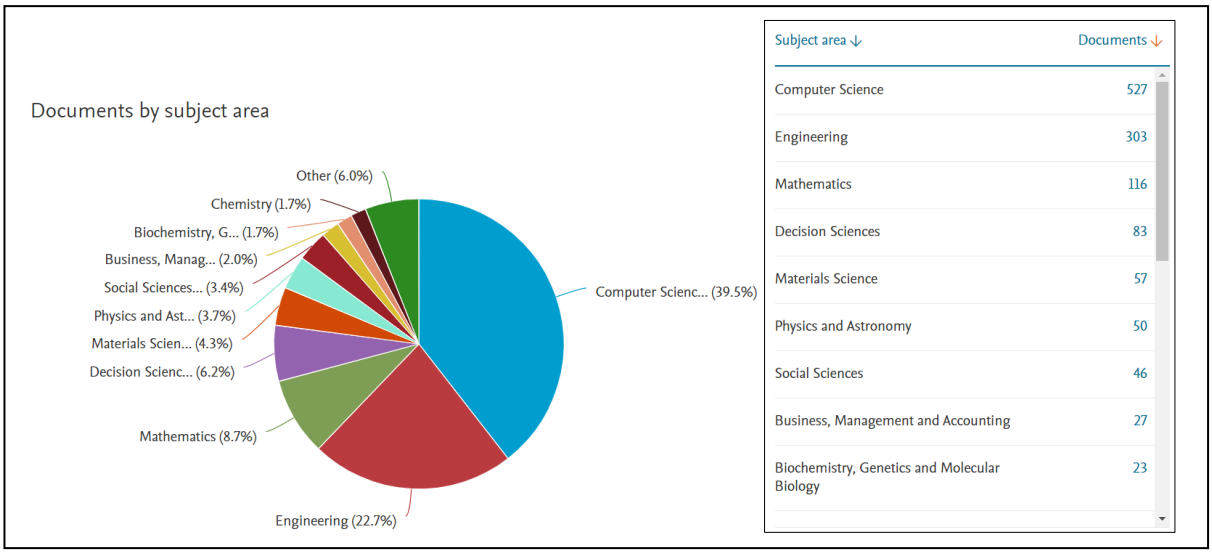


Figure 10. Finding 2: Distribution of documents by subject area

Based on the determination of the fields or topics in which the documents are located, Figure 10 has been obtained. The top 3 areas with the highest number of documents are Computer Science, Engineering, and Mathematics. This situation aligns with Finding 1.

4.3.Finding 3: Scanning based on “explainable artificial intelligence,” and “cyber security,“ and “cyber attack” keywords

A search was conducted based on the keywords 'explainable artificial intelligence,' 'cyber security,' and 'cyber attack,' and the findings (Finding 3) obtained from this search are presented here.

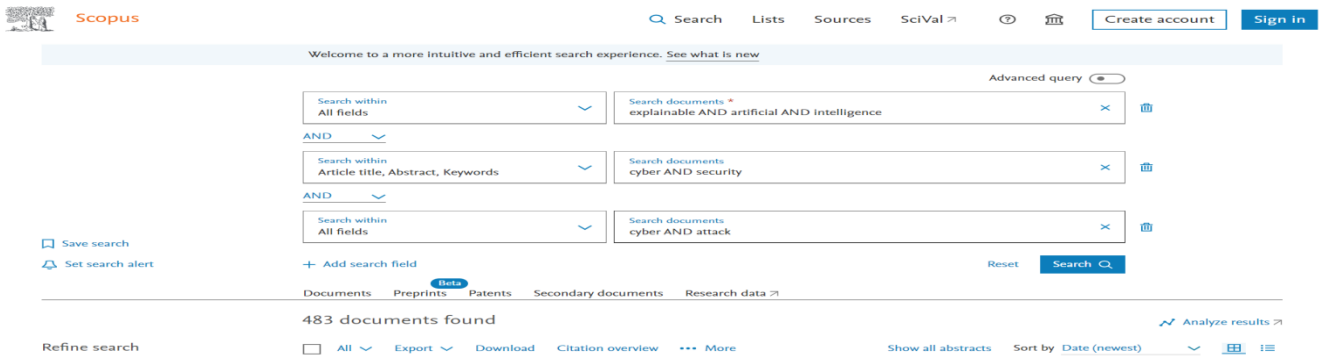


Figure 11. Finding 3: Total number of documents

When the search criteria were set and a search was conducted on Scopus, the total number of documents on the topics of ‘explainable artificial intelligence’, ‘cyber security’, and ‘cyber attack' was found to be 483 (Figure 11).

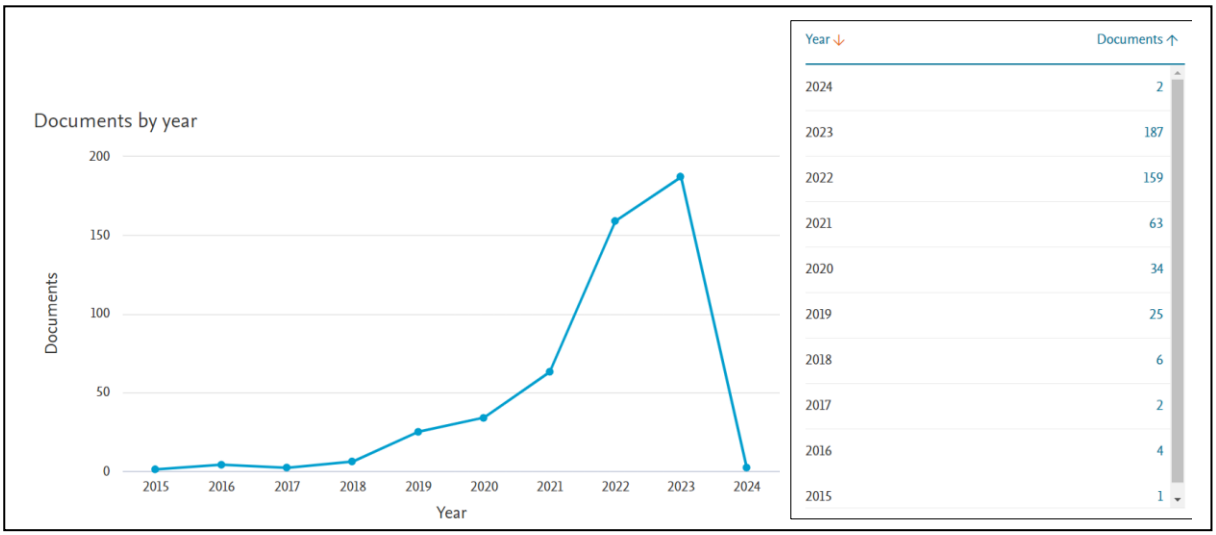


Figure 12. Finding 3: Distribution of documents by year

The distribution of documents based on the years they were published is illustrated in Figure 12. Accordingly, similar to other findings, the number of documents has increased from 2018 to the present, reaching the highest number in 2023 (187). There are two publications for the year 2024. Academic studies have been conducted over the last six years, consistent with other findings.

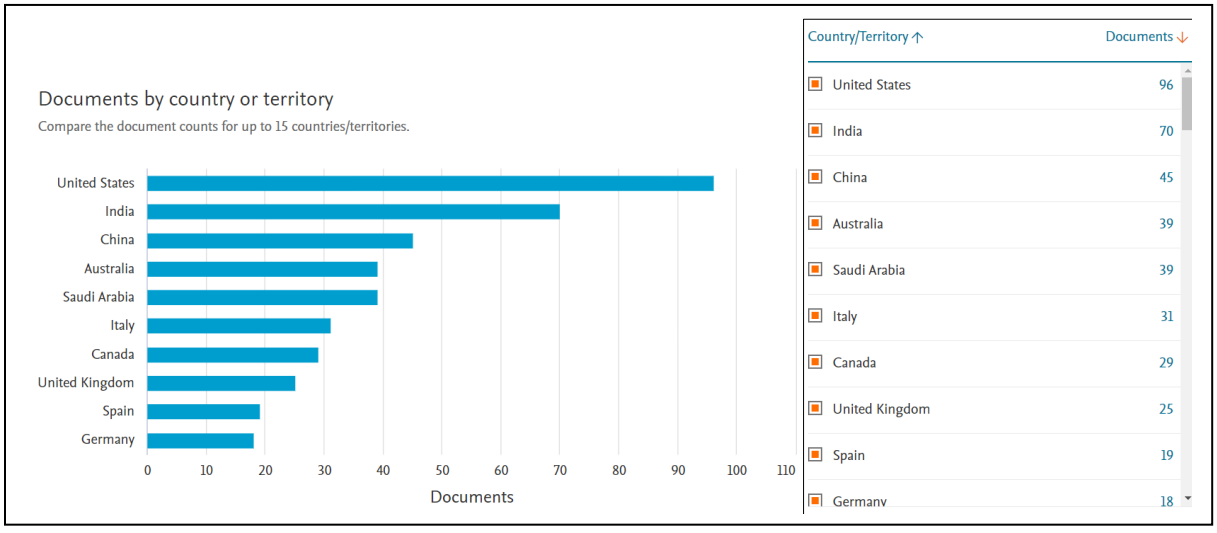


Figure 13. Finding 3: Distribution of documents by countries

The distribution of documents based on the countries where they were conducted is presented in Figure 13. Similar findings to Finding 2 are reported. When examining Figure 13, it is observed that the United States,

India, and China have more publications than other countries. The number of publications from Turkey needs to be improved, (totaling 16).

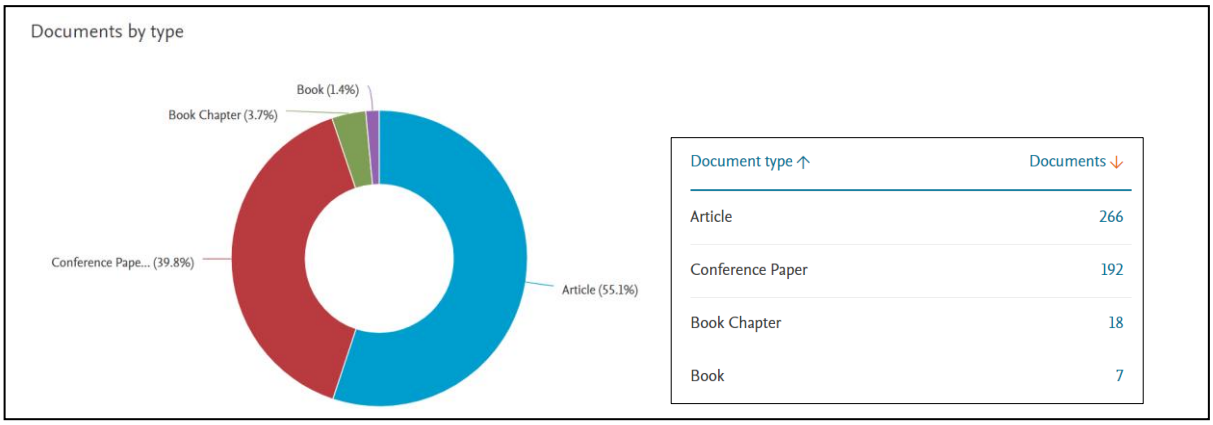


Figure 14. Finding 3: Distribution of documents by type

When examining Figure 14 based on the desired document types for scanning, it is again evident that the most common type of study is the 'Article.' Following this, conference papers, book chapters, and book studies are observed. The least common type of study remains book publications. The results of the first two findings align with each other.

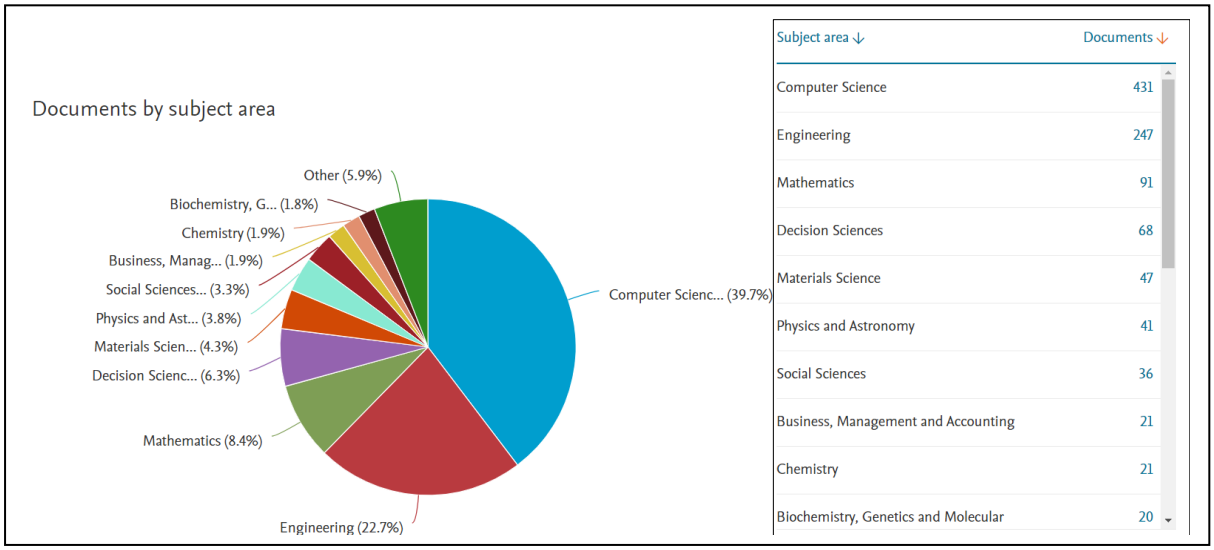


Figure 15. Finding 3: Distribution of documents by subject area

Based on determining in which fields or topics the documents in the final item are located, Figure 15 has been obtained. The top three fields with the highest number of documents are Computer Science, Engineering, and Mathematics.

4.4. Finding 4: Scanning based on "explainable artificial intelligence", and "cyber security", and "attack detection system" keywords

Scanning was conducted based on the keywords 'explainable artificial intelligence', 'cyber security', and 'attack detection system', and the findings (Finding 4) obtained from this search are presented here.

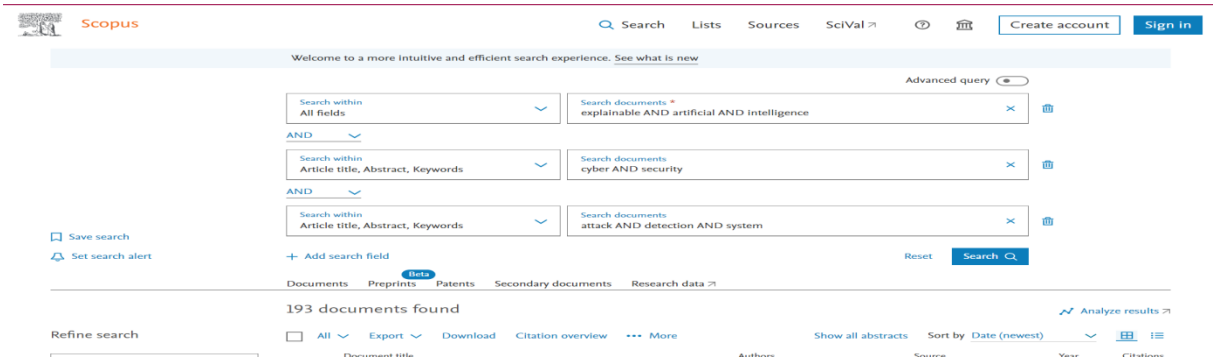


Figure 16. Finding 4: Total number of documents

When the search criteria were set, and a scan was conducted on Scopus, the total number of documents on the topics of 'explainable artificial intelligence', 'cyber security', and 'attack detection system' was found to be 193 (Figure 16).

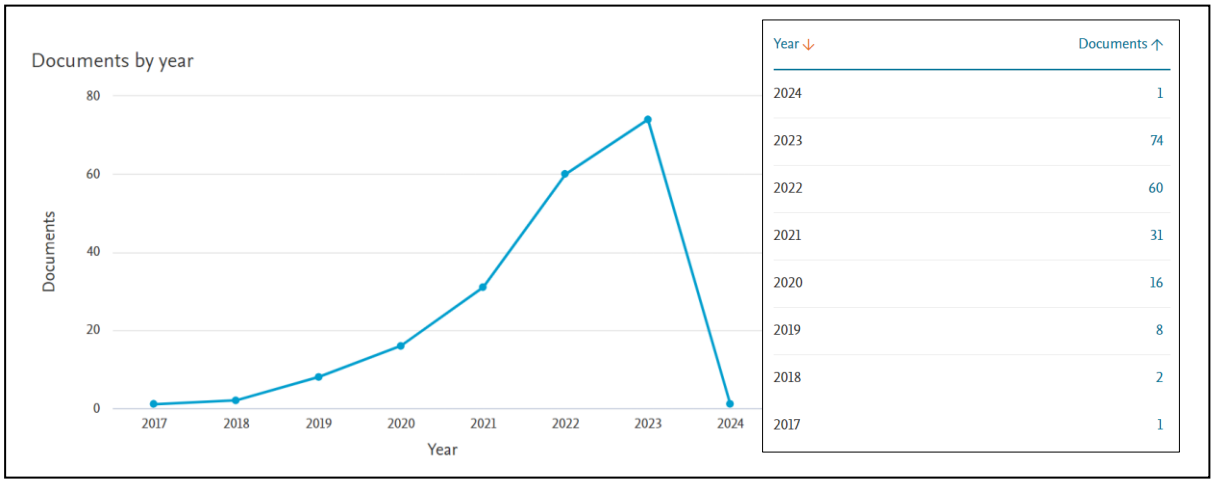


Figure 17. Finding 4: Distribution of documents by year

The distribution of documents over the years is illustrated in Figure 17. Accordingly, similar to other findings, the number of documents has been increasing from 2018 to the present, reaching its peak in 2023 with 74 publications. The first publication for the year 2024 has been released. Academic studies on this topic have consistently been conducted for the past six years.

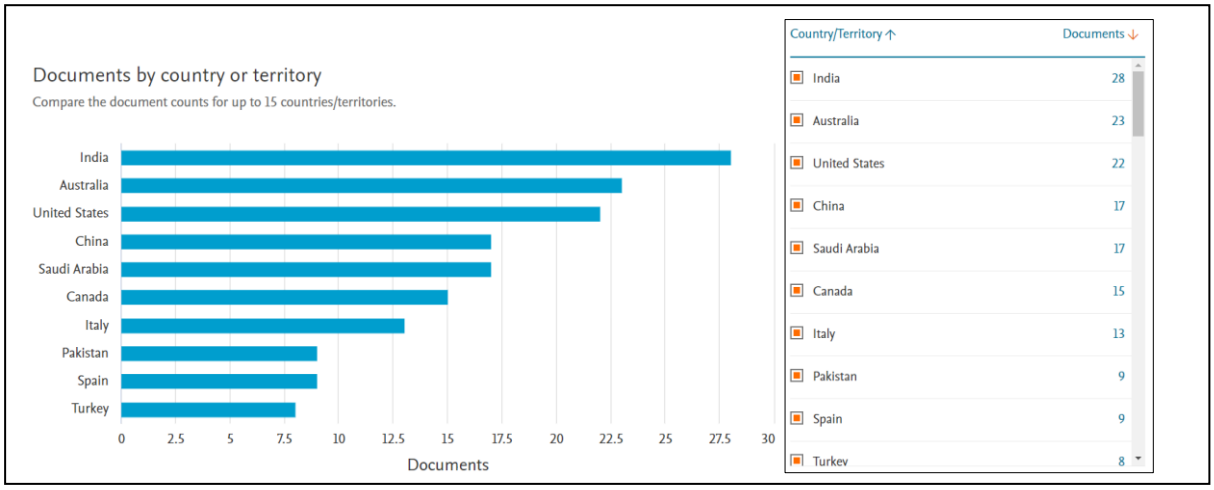


Figure 18. Finding 4: Distribution of documents by countries

The distribution of documents based on the countries where they were conducted is presented in Figure 18. Similar findings to Findings 2 and 3 are observed. Upon examination of Figure 18, it is evident that the United States, India, and Australia have more publications than other countries. The number of publications from Turkey is deficient, (totaling 8).

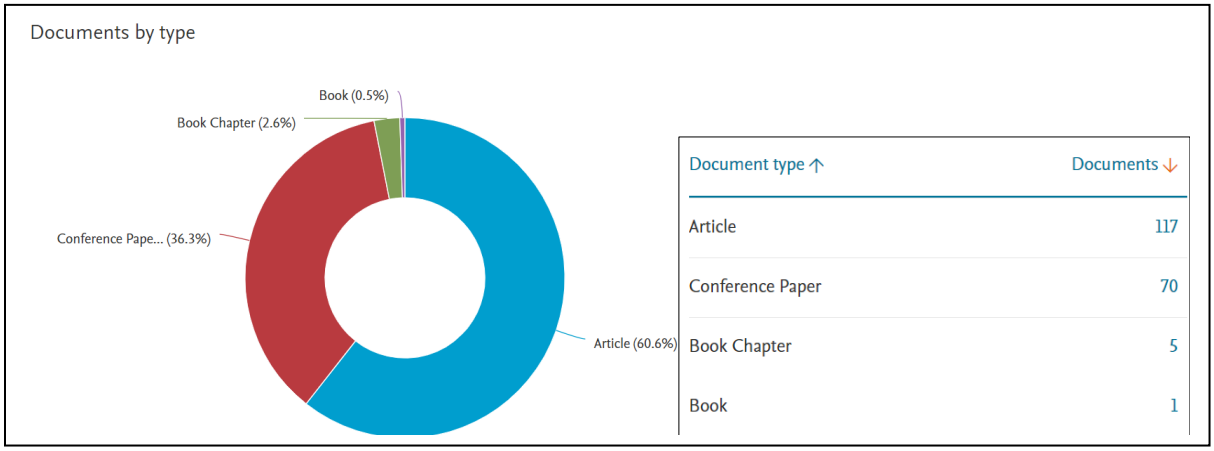


Figure 19. Finding 4:

Distribution of documents by type

When examining the desired document types in Figure 19, it is observed that the Article type is the most frequently conducted study. Subsequently, conference papers, book chapters, and book publications follow. The least common type of document is, again, book publications. The results for the top 3 findings are consistent.

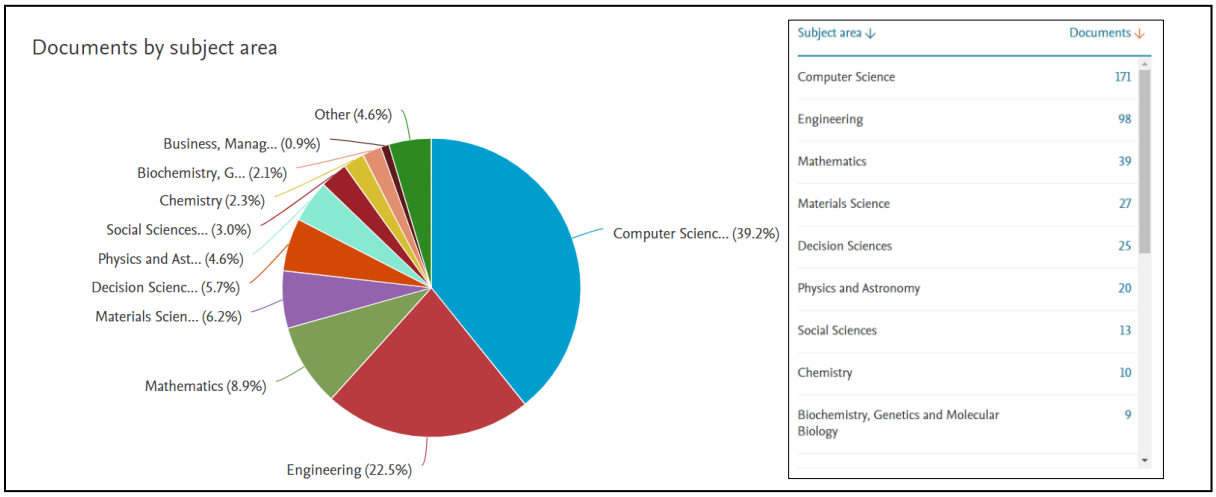


Figure 20. Finding 4: Distribution of documents by subject area

Figure 20 has been obtained based on the determination of the fields or topics in which the documents, the subject of the last finding, are located. Similar to the top 3 findings, the majority of the publications are in the fields of Computer Science, Engineering, and Mathematics.

4.5.Finding 5: Scanning based on “explainable artificial intelligence”, and “cyber security“, and “intrusion detection system” keywords

The search was conducted based on the keywords 'explainable artificial intelligence,' 'cyber security,' and 'intrusion detection system,' and the findings (Finding 5) obtained from this search are presented here. In the search for Finding 4, the term 'attack' was replaced with the term 'intrusion.' These two terms are used interchangeably in studies.

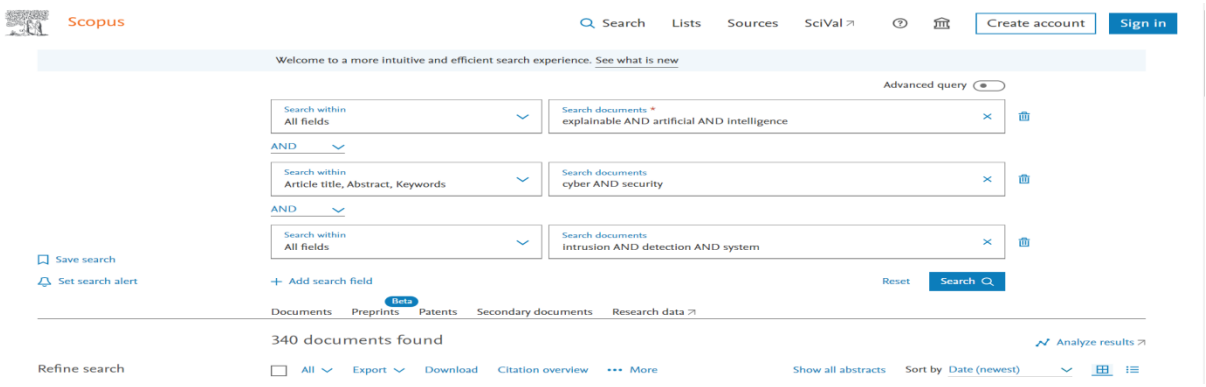


Figure 21. Finding 5: Total number of documents

When the search criteria were set and a search was conducted on Scopus, the total number of publications on the topics of 'explainable artificial intelligence,' 'cyber security,' and 'intrusion detection system' was found to be 340 (Figure 21).

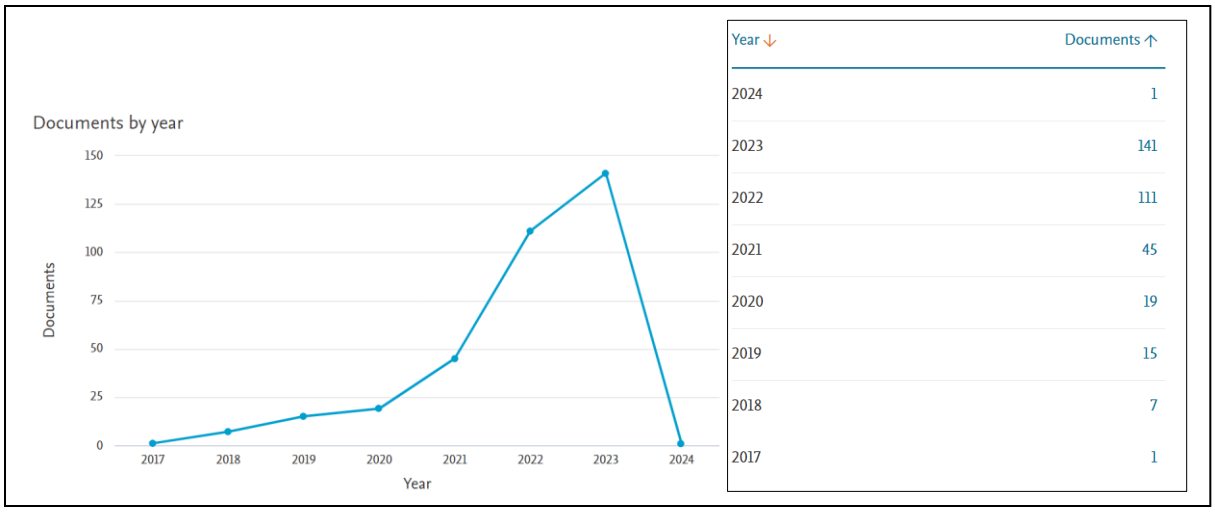


Figure 22. Finding 5: Distribution of documents by year

The distribution of documents over the years is presented in Figure 22. Accordingly, the number of documents has increased from 2018 to the present, reaching the highest number of publications in 2023 (141). The first publication for the year 2024 has already been released. Consistent with other findings, academic studies in this field have been conducted over the past six years.

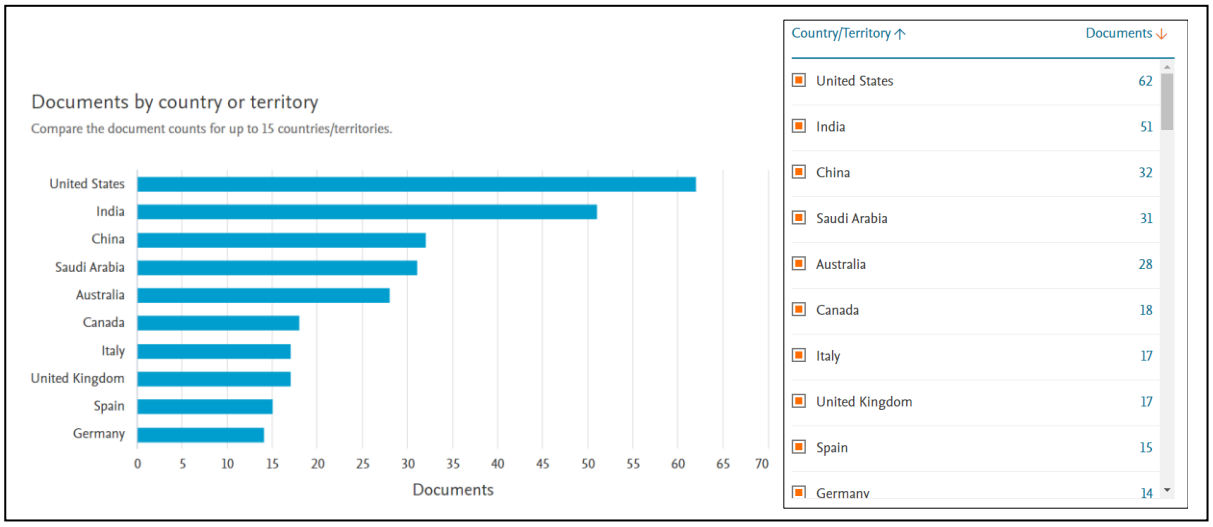


Figure 23. Finding 5: Distribution of documents by countries

The distribution of documents according to the countries in which they were conducted is presented in Figure 23. Similar findings to Findings 2, 3, and 4 are observed. Upon examination of Figure 23, it is evident that the United States, India, and China have more studies compared than other countries. Turkey's publication count is notably low, with 11 publications.

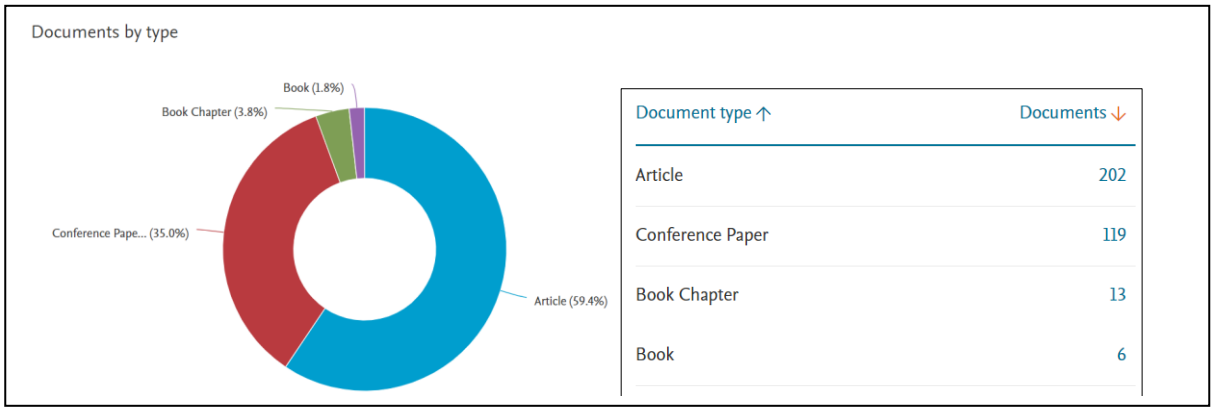


Figure 24. Finding 5: Distribution of documents by type

Upon examination of Figure 24 based on the desired document types, it is observed that the Article type is the most frequently conducted study. Subsequently, conference papers, book chapters, and book publications

frequently follow. The least conducted type of study remains book publications, maintaining consistency with the first four findings.

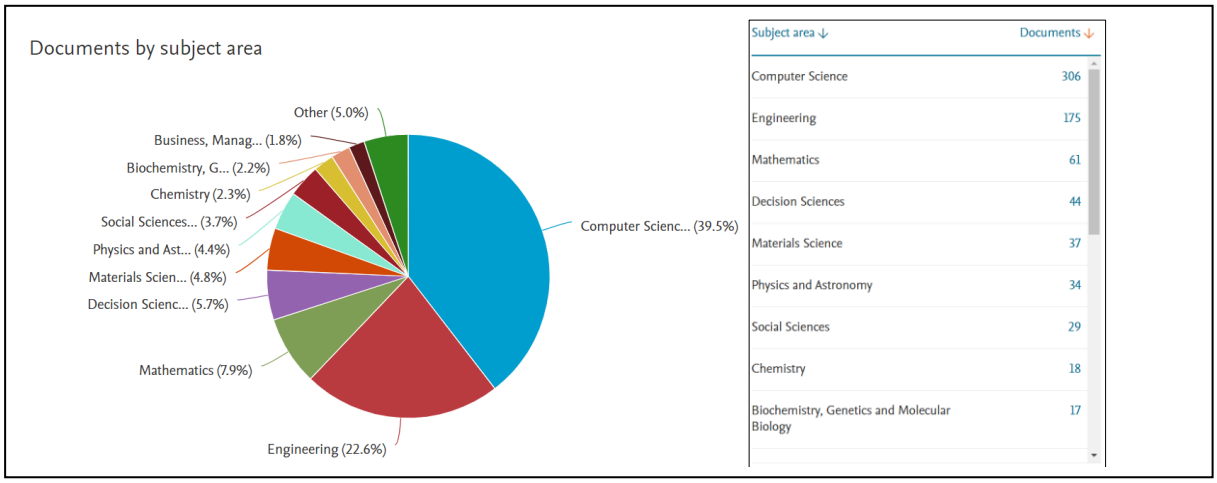


Figure 25. Finding 5: Distribution of documents by subject area

Figure 25 was obtained based on the determining the fields or topics in which the documents in the last finding are located. Like the first four findings, most publications are in Computer Science, Engineering, and Mathematics.

5. Conclusion and Comments

A systematic review method was employed based on the selected keywords to illustrate the relationship between Explainable Artificial Intelligence (XAI) and cyber security and the overall global status. The findings obtained were evaluated under five headings. Conclusions were drawn from all the findings.

Academic publications meeting the specified criteria have seen a notable increase since 2018. When examining the total number of documents from 2014 to the early months of 2024, it is observed that 35% of these were published in 2023. Moreover, 288 publications have already been prepared for 2024. The research areas that gained momentum in the last six years, experiencing a surge in 2023, are likely to continue growing in popularity in the coming years. The increasing trust in artificial intelligence and the shift towards Explainable Artificial Intelligence have contributed to the recent rise in studies. The growing number and types of cyber attacks have also accelerated the convergence of cyber security with Explainable Artificial Intelligence.

When examining the countries of affiliation for authors contributing to these documents, the top five are the United States, India, China, Saudi Arabia, and Australia. Following them are Germany and the United Kingdom. These countries emphasize cyber security and try to integrate Explainable Artificial Intelligence into this field. In contrast, Turkey represents approximately 1% according to Finding 1, 3% according to Findings 2, 3, and 5, and 4% according to Finding 4. Of the 100 studies conducted, roughly three are associated with our country. To enhance our country's prosperity in the fields of Explainable Artificial Intelligence and cyber security, it is imperative to undertake more extensive research and publications. Developing more effective and rapid solutions against cyber attacks is critical for our country and should be expedited accordingly.

In the types of documents conducted over the past decade, articles rank first in all findings, followed by conference papers in the second place, book chapters in the third place, and books in the last place. The occurrence of conference papers following articles indicates that Explainable Artificial Intelligence and cyber security are discussed, and knowledge exchange occurs at scientific conferences. The percentage of publications in the book category is approximately 1%. Articles constitute approximately 56% of the total. To enhance the scientific dissemination of knowledge, there is a need to increase the number of books, provide scholarly information in book format, and even elevate book publications to a level suitable for instructional purposes.

When examining the documents resulting from the search, it is observed that the top three fields are Computer Science, Engineering, and Mathematics. This pattern is consistent across all findings. Subsequently, publications such as Decision Sciences and Materials Science have also been identified. From this, it is understood that Explainable Artificial Intelligence and cyber security are contemporary topics relevant to various domains.

One keyword was altered in the search conducted for Findings 4 and 5. 'Attack' and 'Intrusion' are interchangeable terms in the cyber security domain. Using the keyword 'Attack' yielded 193 publications in Finding 4, while using the keyword 'Intrusion' in Finding 5 resulted in 340 publications. These highlights the significance of the keywords chosen during scientific and academic searches. Alternative keywords should be considered, and diverse searches should be performed. Additionally, when preparing publications, attention should be given to these considerations to enhance the visibility and utilization of publications in the areas of interest.

References

1. Söğüt, E., & Erdem, O. A. (2023). “A Multi-Model Proposal for Classification and Detection of DDoS Attacks on SCADA Systems”. *Applied Sciences*, 13(10), 5993. https://doi.org/10.3390/app13105993

2. Yaz, Ö., & Süzen, A. A. (2023). "Kurumsal Ağlara Yapılan Saldırılarda Açıklanabilir Yapay Zekânın Yeri". *The 5 th International Conference on Applied Engineering and Natural Sciences.* 528-534. Konya.
3. Erhan, B. A. Ş., & Süzen, A. A. (2023). "Web Uygulama Zafiyetlerinin Keşfinde Açıklanabilir Yapay Zekânın Yeri". *The 5 th International Conference on Applied Engineering and Natural Sciences.* 517-527. Konya.
4. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). "Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model", *Complexity*, vol. 2021, Article ID 6634811, 11 pages. https://doi.org/10.1155/2021/6634811
5. Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). "Explainable artificial intelligence in cybersecurity: A survey". *IEEE Access*, vol. 10, 93575-93600. https://doi.org/10.1109/ACCESS.2022.3204171
6. Şencan, Ö. A., Atacak, İ., & Doğru, İ. A. (2022). "Sosyal Ağlarda Topluluk ve Konu Tespiti: Bir Sistematik Literatür Taraması". *International Journal of Informatics Technologies*, 15(3). https://doi.org/10.17671/gazibtd.1061332
7. Kılınç, H. (2022). "Nesnelerin İnterneti Konusunda Gerçekleştirilen Akademik Çalışmaların Eğilimleri: Bir Sistematik Tarama". *Dijital Teknolojiler ve Eğitim Dergisi*, 1(1), 57-67. https://doi.org/10.5281/zenodo.6647667

Scopus web site, https://www.scopus.com/home.uri, Access Date: September, 2023.

View publication stats