The Geneva Papers on Risk and Insurance - Issues and Practice (2022) 47:698–736
https://doi.org/10.1057/s41288-022-00266-6

Cyber risk and cybersecurity: a systematic review of data availability

Frank Cremer[1] · Barry Sheehan[1] · Michael Fortmann[2] · Arash N. Kia[1] · Martin Mullins[1] · Finbarr Murphy[1] · Stefan Materne[2]

Received: 15 June 2021 / Accepted: 20 January 2022 / Published online: 17 February 2022

Abstract
Cybercrime is estimated to have cost the global economy just under USD 1 trillion in 2020, indicating an increase of more than 50% since 2018. With the average cyber insurance claim rising from USD 145,000 in 2019 to USD 359,000 in 2020, there is a growing necessity for better cyber information sources, standardised databases, mandatory reporting and public awareness. This research analyses the extant academic and industry literature on cybersecurity and cyber risk management with a particular focus on data availability. From a preliminary search resulting in 5219 cyber peer-reviewed studies, the application of the systematic methodology resulted in 79 unique datasets. We posit that the lack of available data on cyber risk poses a serious problem for stakeholders seeking to tackle this issue. In particular, we identify a lacuna in open databases that undermine collective endeavours to better manage this set of risks. The resulting data evaluation and categorisation will support cybersecurity researchers and the insurance industry in their efforts to comprehend, metricise and manage cyber risks.

Keywords Cyber insurance · Cyber risk · Open data · Systematic review · Cybersecurity

Introduction

Globalisation, digitalisation and smart technologies have escalated the propensity and severity of cybercrime. Whilst it is an emerging field of research and industry, the importance of robust cybersecurity defence systems has been highlighted at the corporate, national and supranational levels. The impacts of inadequate

✉ Barry Sheehan
barry.sheehan@ul.ie

[1] University of Limerick, Limerick, Ireland

[2] TH Köln University of Applied Sciences, Cologne, Germany

cybersecurity are estimated to have cost the global economy USD 945 billion in 2020 (Maleks Smith et al. 2020). Cyber vulnerabilities pose significant corporate risks, including business interruption, breach of privacy and financial losses (Sheehan et al. 2019). Despite the increasing relevance for the international economy, the availability of data on cyber risks remains limited. The reasons for this are many. Firstly, it is an emerging and evolving risk; therefore, historical data sources are limited (Biener et al. 2015). It could also be due to the fact that, in general, institutions that have been hacked do not publish the incidents (Eling and Schnell 2016). The lack of data poses challenges for many areas, such as research, risk management and cybersecurity (Falco et al. 2019). The importance of this topic is demonstrated by the announcement of the European Council in April 2021 that a centre of excellence for cybersecurity will be established to pool investments in research, technology and industrial development. The goal of this centre is to increase the security of the internet and other critical network and information systems (European Council 2021).

This research takes a risk management perspective, focusing on cyber risk and considering the role of cybersecurity and cyber insurance in risk mitigation and risk transfer. The study reviews the existing literature and open data sources related to cybersecurity and cyber risk. This is the first systematic review of data availability in the general context of cyber risk and cybersecurity. By identifying and critically analysing the available datasets, this paper supports the research community by aggregating, summarising and categorising all available open datasets. In addition, further information on datasets is attached to provide deeper insights and support stakeholders engaged in cyber risk control and cybersecurity. Finally, this research paper highlights the need for open access to cyber-specific data, without price or permission barriers.

The identified open data can support cyber insurers in their efforts on sustainable product development. To date, traditional risk assessment methods have been untenable for insurance companies due to the absence of historical claims data (Sheehan et al. 2021). These high levels of uncertainty mean that cyber insurers are more inclined to overprice cyber risk cover (Kshetri 2018). Combining external data with insurance portfolio data therefore seems to be essential to improve the evaluation of the risk and thus lead to risk-adjusted pricing (Bessy-Roland et al. 2021). This argument is also supported by the fact that some re/insurers reported that they are working to improve their cyber pricing models (e.g. by creating or purchasing databases from external providers) (EIOPA 2018). Figure 1 provides an overview of pricing tools and factors considered in the estimation of cyber insurance based on the findings of EIOPA (2018) and the research of Romanosky et al. (2019). The term cyber risk refers to all cyber risks and their potential impact.

Besides the advantage of risk-adjusted pricing, the availability of open datasets helps companies benchmark their internal cyber posture and cybersecurity measures. The research can also help to improve risk awareness and corporate behaviour. Many companies still underestimate their cyber risk (Leong and Chen 2020). For policymakers, this research offers starting points for a comprehensive recording of cyber risks. Although in many countries, companies are obliged to report data breaches to the respective supervisory authority, this information is usually not

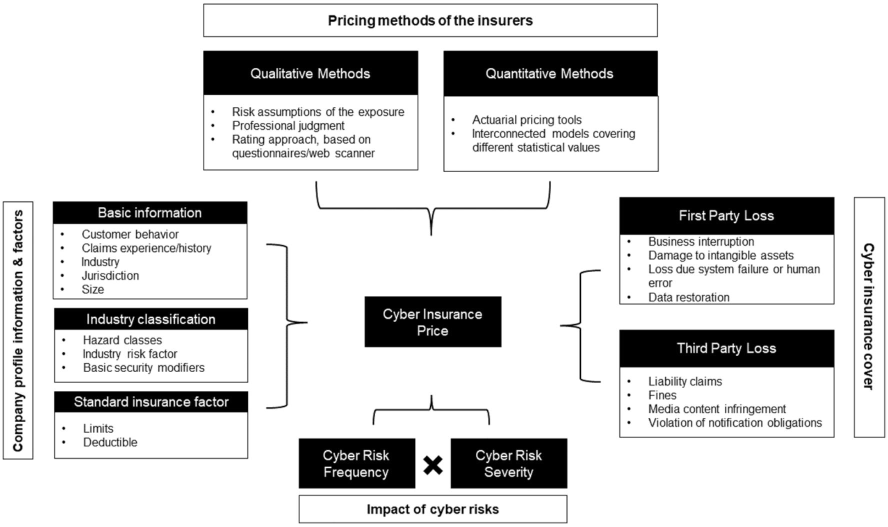

Fig. 1 An overview of the current cyber insurance informational and methodological landscape, adapted from EIOPA (2018) and Romanosky et al. (2019)

accessible to the research community. Furthermore, the economic impact of these breaches is usually unclear.

As well as the cyber risk management community, this research also supports cybersecurity stakeholders. Researchers are provided with an up-to-date, peer-reviewed literature of available datasets showing where these datasets have been used. For example, this includes datasets that have been used to evaluate the effectiveness of countermeasures in simulated cyberattacks or to test intrusion detection systems. This reduces a time-consuming search for suitable datasets and ensures a comprehensive review of those available. Through the dataset descriptions, researchers and industry stakeholders can compare and select the most suitable datasets for their purposes. In addition, it is possible to combine the datasets from one source in the context of cybersecurity or cyber risk. This supports efficient and timely progress in cyber risk research and is beneficial given the dynamic nature of cyber risks.

Cyber risks are defined as “operational risks to information and technology assets that have consequences affecting the confidentiality, availability, and/or integrity of information or information systems” (Cebula et al. 2014). Prominent cyber risk events include data breaches and cyberattacks (Agrafiotis et al. 2018). The increasing exposure and potential impact of cyber risk have been highlighted in recent industry reports (e.g. Allianz 2021; World Economic Forum 2020). Cyberattacks on critical infrastructures are ranked 5th in the World Economic Forum’s Global Risk Report. Ransomware, malware and distributed denial-of-service (DDoS) are examples of the evolving modes of a cyberattack. One example is the ransomware attack on the Colonial Pipeline, which shut down the 5500 mile pipeline system that delivers 2.5 million barrels of fuel per day and critical liquid fuel infrastructure from oil refineries to states along the U.S. East Coast (Brower and McCormick 2021). These and other cyber incidents have led the U.S. to strengthen its cybersecurity and introduce, among other things, a public body to analyse major cyber incidents and make recommendations to prevent a recurrence (Murphey 2021a). Another example of the scope of cyberattacks is the ransomware NotPetya in 2017. The damage amounted to USD 10 billion, as the ransomware exploited a vulnerability in the windows system, allowing it to spread independently worldwide in the network (GAO 2021). In the same year, the ransomware WannaCry was launched by cybercriminals. The cyberattack on Windows software took user data hostage in exchange for Bitcoin cryptocurrency (Smart 2018). The victims included the National Health Service in Great Britain. As a result, ambulances were redirected to other hospitals because of information technology (IT) systems failing, leaving people in need of urgent assistance waiting. It has been estimated that 19,000 cancelled treatment appointments resulted from losses of GBP 92 million (Field 2018). Throughout the COVID-19 pandemic, ransomware attacks increased significantly, as working from home arrangements increased vulnerability (Murphey 2021b).

Besides cyberattacks, data breaches can also cause high costs. Under the General Data Protection Regulation (GDPR), companies are obliged to protect personal data and safeguard the data protection rights of all individuals in the EU area. The GDPR allows data protection authorities in each country to impose sanctions and fines on organisations they find in breach. “For data breaches, the maximum fine can be €20 million or 4% of global turnover, whichever is higher” (GDPR.EU 2021). Data breaches often involve a large amount of sensitive data that has been accessed, unauthorised, by external parties, and are therefore considered important for information security due to their far-reaching impact (Goode et al. 2017). A data breach is defined as a “security incident in which sensitive, protected, or confidential data are copied, transmitted, viewed, stolen, or used by an unauthorized individual” (Freeha et al. 2021). Depending on the amount of data, the extent of the damage caused by a data breach can be significant, with the average cost being USD 392 million[1] (IBM Security 2020).

This research paper reviews the existing literature and open data sources related to cybersecurity and cyber risk, focusing on the datasets used to improve academic understanding and advance the current state-of-the-art in cybersecurity. Furthermore, important information about the available datasets is presented (e.g. use cases), and a plea is made for open data and the standardisation of cyber risk data for academic comparability and replication. The remainder of the paper is structured as follows. The next section describes the related work regarding cybersecurity and cyber risks. The third section outlines the review method used in this work and the process. The fourth section details the results of the identified literature. Further discussion is presented in the penultimate section and the final section concludes.

[1] Average cost of a breach of more than 50 million records.

Related work

Due to the significance of cyber risks, several literature reviews have been conducted in this field. Eling (2020) reviewed the existing academic literature on the topic of cyber risk and cyber insurance from an economic perspective. A total of 217 papers with the term 'cyber risk' were identified and classified in different categories. As a result, open research questions are identified, showing that research on cyber risks is still in its infancy because of their dynamic and emerging nature. Furthermore, the author highlights that particular focus should be placed on the exchange of information between public and private actors. An improved information flow could help to measure the risk more accurately and thus make cyber risks more insurable and help risk managers to determine the right level of cyber risk for their company. In the context of cyber insurance data, Romanosky et al. (2019) analysed the underwriting process for cyber insurance and revealed how cyber insurers understand and assess cyber risks. For this research, they examined 235 American cyber insurance policies that were publicly available and looked at three components (coverage, application questionnaires and pricing). The authors state in their findings that many of the insurers used very simple, flat-rate pricing (based on a single calculation of expected loss), while others used more parameters such as the asset value of the company (or company revenue) or standard insurance metrics (e.g. deductible, limits), and the industry in the calculation. This is in keeping with Eling (2020), who states that an increased amount of data could help to make cyber risk more accurately measured and thus more insurable. Similar research on cyber insurance and data was conducted by Nurse et al. (2020). The authors examined cyber insurance practitioners' perceptions and the challenges they face in collecting and using data. In addition, gaps were identified during the research where further data is needed. The authors concluded that cyber insurance is still in its infancy, and there are still several unanswered questions (for example, cyber valuation, risk calculation and recovery). They also pointed out that a better understanding of data collection and use in cyber insurance would be invaluable for future research and practice. Bessy-Roland et al. (2021) come to a similar conclusion. They proposed a multivariate Hawkes framework to model and predict the frequency of cyberattacks. They used a public dataset with characteristics of data breaches affecting the U.S. industry. In the conclusion, the authors make the argument that an insurer has a better knowledge of cyber losses, but that it is based on a small dataset and therefore combination with external data sources seems essential to improve the assessment of cyber risks.

Several systematic reviews have been published in the area of cybersecurity (Kruse et al. 2017; Lee et al. 2020; Loukas et al. 2013; Ulven and Wangen 2021). In these papers, the authors concentrated on a specific area or sector in the context of cybersecurity. This paper adds to this extant literature by focusing on data availability and its importance to risk management and insurance stakeholders. With a priority on healthcare and cybersecurity, Kruse et al. (2017) conducted a systematic literature review. The authors identified 472 articles with the keywords 'cybersecurity and healthcare' or 'ransomware' in the databases Cumulative Index of Nursing and Allied Health Literature, PubMed and Proquest. Articles were eligible for this review if they satisfied three criteria: (1) they were published between 2006 and 2016, (2) the full-text version of the article was available, and (3) the publication is a peer-reviewed or scholarly journal. The authors found that technological development and federal policies (in the U.S.) are the main factors exposing the health sector to cyber risks. Loukas et al. (2013) conducted a review with a focus on cyber risks and cybersecurity in emergency management. The authors provided an overview of cyber risks in communication, sensor, information management and vehicle technologies used in emergency management and showed areas for which there is still no solution in the literature. Similarly, Ulven and Wangen (2021) reviewed the literature on cybersecurity risks in higher education institutions. For the literature review, the authors used the keywords 'cyber', 'information threats' or 'vulnerability' in connection with the terms 'higher education, 'university' or 'academia'. A similar literature review with a focus on Internet of Things (IoT) cybersecurity was conducted by Lee et al. (2020). The review revealed that qualitative approaches focus on high-level frameworks, and quantitative approaches to cybersecurity risk management focus on risk assessment and quantification of cyberattacks and impacts. In addition, the findings presented a four-step IoT cyber risk management framework that identifies, quantifies and prioritises cyber risks.

Datasets are an essential part of cybersecurity research, underlined by the following works. Ilhan Firat et al. (2021) examined various cybersecurity datasets in detail. The study was motivated by the fact that with the proliferation of the internet and smart technologies, the mode of cyberattacks is also evolving. However, in order to prevent such attacks, they must first be detected; the dissemination and further development of cybersecurity datasets is therefore critical. In their work, the authors observed studies of datasets used in intrusion detection systems. Khraisat et al. (2019) also identified a need for new datasets in the context of cybersecurity. The researchers presented a taxonomy of current intrusion detection systems, a comprehensive review of notable recent work, and an overview of the datasets commonly used for assessment purposes. In their conclusion, the authors noted that new datasets are needed because most machine-learning techniques are trained and evaluated on the knowledge of old datasets. These datasets do not contain new and comprehensive information and are partly derived from datasets from 1999. The authors noted that the core of this issue is the availability of new public datasets as well as their quality. The availability of data, how it is used, created and shared was also investigated by Zheng et al. (2018). The researchers analysed 965 cybersecurity research papers published between 2012 and 2016. They created a taxonomy of the types of data that are created and shared and then analysed the data collected via datasets. The researchers concluded that while datasets are recognised as valuable for cybersecurity research, the proportion of publicly available datasets is limited.

The main contributions of this review and what differentiates it from previous studies can be summarised as follows. First, as far as we can tell, it is the first work to summarise all available datasets on cyber risk and cybersecurity in the context of a systematic review and present them to the scientific community and cyber insurance and cybersecurity stakeholders. Second, we investigated, analysed, and made available the datasets to support efficient and timely progress in cyber risk research.

And third, we enable comparability of datasets so that the appropriate dataset can be selected depending on the research area.

Methodology

Process and eligibility criteria

The structure of this systematic review is inspired by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Page et al. 2021), and the search was conducted from 3 to 10 May 2021. Due to the continuous development of cyber risks and their countermeasures, only articles published in the last 10 years were considered. In addition, only articles published in peer-reviewed journals written in English were included. As a final criterion, only articles that make use of one or more cybersecurity or cyber risk datasets met the inclusion criteria. Specifically, these studies presented new or existing datasets, used them for methods, or used them to verify new results, as well as analysed them in an economic context and pointed out their effects. The criterion was fulfilled if it was clearly stated in the abstract that one or more datasets were used. A detailed explanation of this selection criterion can be found in the 'Study selection' section.

Information sources

In order to cover a complete spectrum of literature, various databases were queried to collect relevant literature on the topic of cybersecurity and cyber risks. Due to the spread of related articles across multiple databases, the literature search was limited to the following four databases for simplicity: IEEE Xplore, Scopus, SpringerLink and Web of Science. This is similar to other literature reviews addressing cyber risks or cybersecurity, including Sardi et al. (2021), Franke and Brynielsson (2014), Lagerström (2019), Eling and Schnell (2016) and Eling (2020). In this paper, all databases used in the aforementioned works were considered. However, only two studies also used all the databases listed. The IEEE Xplore database contains electrical engineering, computer science, and electronics work from over 200 journals and three million conference papers (IEEE 2021). Scopus includes 23,400 peer-reviewed journals from more than 5000 international publishers in the areas of science, engineering, medicine, social sciences and humanities (Scopus 2021). SpringerLink contains 3742 journals and indexes over 10 million scientific documents (SpringerLink 2021). Finally, Web of Science indexes over 9200 journals in different scientific disciplines (Science 2021).

Search

A search string was created and applied to all databases. To make the search efficient and reproducible, the following search string with Boolean operator was used in all databases: cybersecurity OR cyber risk AND dataset OR database. To ensure uniformity of the search across all databases, some adjustments had to be made for the respective search engines. In Scopus, for example, the Advanced Search was used, and the field code 'Title-ABS-KEY' was integrated into the search string. For IEEE Xplore, the search was carried out with the Search String in the Command Search and 'All Metadata'. In the Web of Science database, the Advanced Search was used. The special feature of this search was that it had to be carried out in individual steps. The first search was carried out with the terms cybersecurity OR cyber risk with the field tag Topic (T.S. =) and the second search with dataset OR database. Subsequently, these searches were combined, which then delivered the searched articles for review. For SpringerLink, the search string was used in the Advanced Search under the category 'Find the resources with all of the words'. After conducting this search string, 5219 studies could be found. According to the eligibility criteria (period, language and only scientific journals), 1581 studies were identified in the databases:

- IEEE: 364
- Scopus: 135
- Springer Link: 548
- Web of Science: 534

An overview of the process is given in Fig. 2. Combined with the results from the four databases, 854 articles without duplicates were identified.

Study selection

In the final step of the selection process, the articles were screened for relevance. Due to a large number of results, the abstracts were analysed in the first step of the process. The aim was to determine whether the article was relevant for the systematic review. An article fulfilled the criterion if it was recognisable in the abstract that it had made a contribution to datasets or databases with regard to cyber risks or cybersecurity. Specifically, the criterion was considered to be met if the abstract used datasets that address the causes or impacts of cyber risks, and measures in the area of cybersecurity. In this process, the number of articles was reduced to 288. The articles were then read in their entirety, and an expert panel of six people decided whether they should be used. This led to a final number of 255 articles. The years in which the articles were published and the exact number can be seen in Fig. 3.

Data collection process and synthesis of the results

For the data collection process, various data were extracted from the studies, including the names of the respective creators, the name of the dataset or database and the corresponding reference. It was also determined where the data came from. In the context of accessibility, it was determined whether access is free, controlled, available for purchase or not available. It was also determined when the datasets were

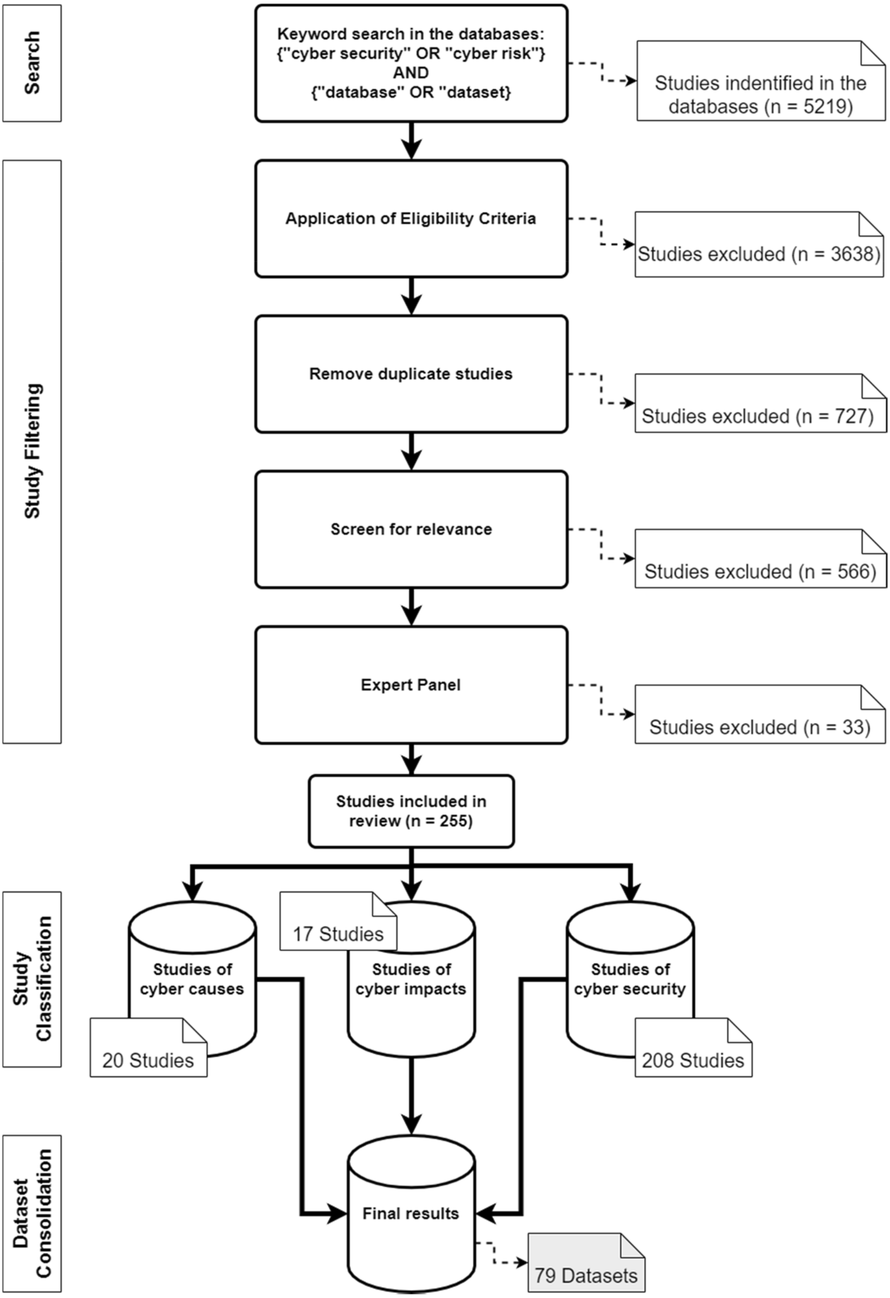
Search
Study Filtering
Study Classification
Dataset Consolidation
Keyword search in the databases: {"cyber security" OR "cyber risk"} AND {"database" OR "dataset}
Studies indentified in the databases (n = 5219)
Application of Eligibility Criteria
Studies excluded (n = 3638)
Remove duplicate studies
Studies excluded (n = 727)
Screen for relevance
Studies excluded (n = 566)
Expert Panel
Studies excluded (n = 33)
Studies included in review (n = 255)
Studies of cyber causes
20 Studies
17 Studies
Studies of cyber impacts
Studies of cyber security
208 Studies
Final results
79 Datasets

Fig. 2 Literature search process and categorisation of the studies

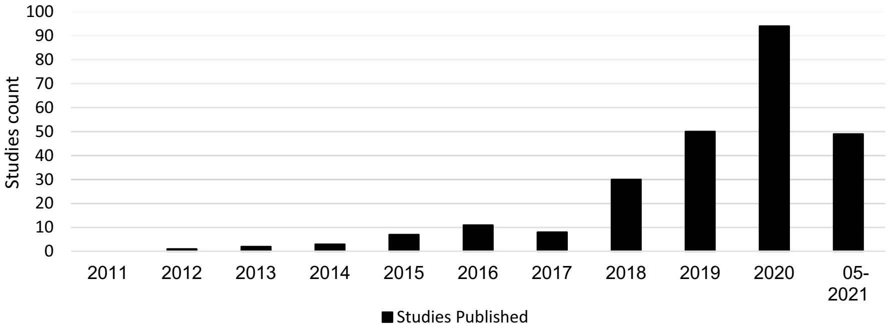

Fig. 3 Distribution of studies

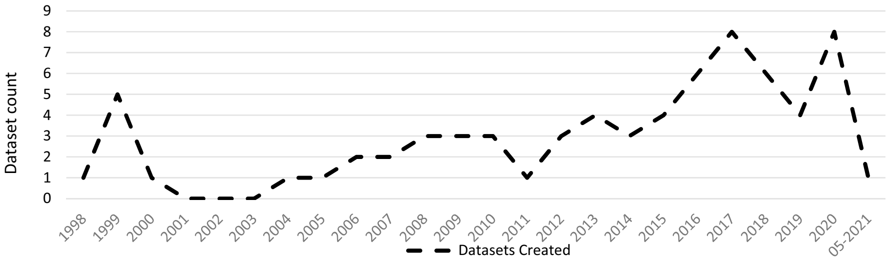

Fig. 4 Distribution of dataset results

created and the time period referenced. The application type and domain characteristics of the datasets were identified.

Results

This section analyses the results of the systematic literature review. The previously identified studies are divided into three categories: datasets on the causes of cyber risks, datasets on the effects of cyber risks and datasets on cybersecurity. The classification is based on the intended use of the studies. This system of classification makes it easier for stakeholders to find the appropriate datasets. The categories are evaluated individually. Although complete information is available for a large proportion of datasets, this is not true for all of them. Accordingly, the abbreviation N/A has been inserted in the respective characters to indicate that this information could not be determined by the time of submission. The term 'use cases in the literature' in the following and supplementary tables refers to the application areas in which the corresponding datasets were used in the literature. The areas listed there refer to the topic area on which the researchers conducted their research. Since some datasets were used interdisciplinarily, the listed use cases in the literature are correspondingly longer. Before discussing each category in the next sections, Fig. 4 provides an overview of the number of datasets found and their year of creation.

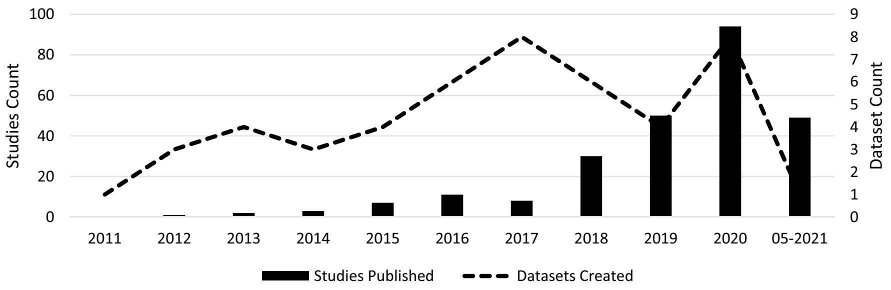

Fig. 5 Correlation between the studies and the datasets

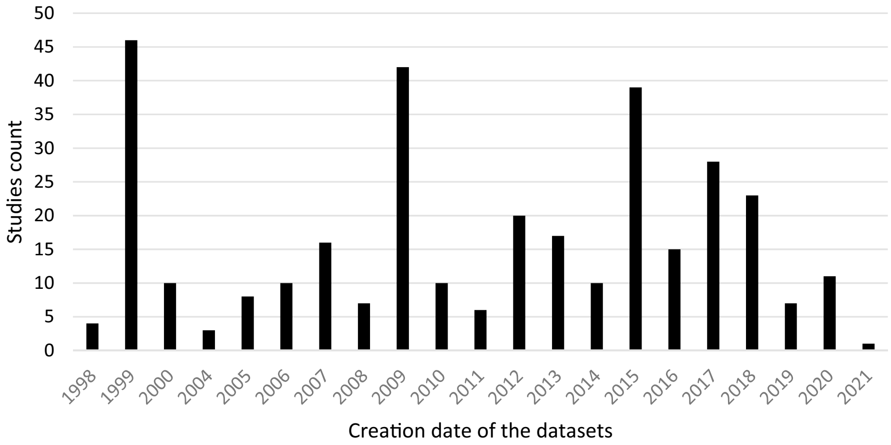

Fig. 6 Distribution of studies and their use of datasets

Figure 5 then shows the relationship between studies and datasets in the period under consideration. Figure 6 shows the distribution of studies, their use of datasets and their creation date. The number of datasets used is higher than the number of studies because the studies often used several datasets (Table 1).

Most of the datasets are generated in the U.S. (up to 58.2%). Canada and Australia rank next, with 11.3% and 5% of all the reviewed datasets, respectively.

Additionally, to create value for the datasets for the cyber insurance industry, an assessment of the applicability of each dataset has been provided for cyber insurers. This ‘Use Case Assessment’ includes the use of the data in the context of different analyses, calculation of cyber insurance premiums, and use of the information for the design of cyber insurance contracts or for additional customer services. To reasonably account for the transition of direct hyperlinks in the future, references were directed to the main websites for longevity (nearest resource point). In addition, the links to the main pages contain further information on the datasets and different versions related to the operating systems. The references were chosen in such a way that practitioners get the best overview of the respective datasets.

Table 1 Percentage contribution of datasets for each place of origin

Rank	Place of origin	Percentage of datasets
1	U.S.	58.2
2	Canada	11.3
3	Australia	5
4	Germany	3.7
5	U.K.	3.7
6	France	2.5
7	Italy	2.5
8	Spain	2.5
9	China	1.2
10	Czech Republic	1.2
11	Greece	1.2
12	Japan	1.2
13	Lithuania	1.2
14	Luxembourg	1.2
15	Netherlands	1.2
16	Republic of Korea	1.2
17	Turkey	1.2

Case datasets

This section presents selected articles that use the datasets to analyse the causes of cyber risks. The datasets help identify emerging trends and allow pattern discovery in cyber risks. This information gives cybersecurity experts and cyber insurers the data to make better predictions and take appropriate action. For example, if certain vulnerabilities are not adequately protected, cyber insurers will demand a risk surcharge leading to an improvement in the risk-adjusted premium. Due to the capricious nature of cyber risks, existing data must be supplemented with new data sources (for example, new events, new methods or security vulnerabilities) to determine prevailing cyber exposure. The datasets of cyber risk causes could be combined with existing portfolio data from cyber insurers and integrated into existing pricing tools and factors to improve the valuation of cyber risks.

A portion of these datasets consists of several taxonomies and classifications of cyber risks. Aassal et al. (2020) propose a new taxonomy of phishing characteristics based on the interpretation and purpose of each characteristic. In comparison, Hindy et al. (2020) presented a taxonomy of network threats and the impact of current datasets on intrusion detection systems. A similar taxonomy was suggested by Kiwia et al. (2018). The authors presented a cyber kill chain-based taxonomy of banking Trojans features. The taxonomy built on a real-world dataset of 127 banking Trojans collected from December 2014 to January 2016 by a major U.K.-based financial organisation.

In the context of classification, Aamir et al. (2021) showed the benefits of machine learning for classifying port scans and DDoS attacks in a mixture of normal and attack traffic. Guo et al. (2020) presented a new method to improve malware classification based on entropy sequence features. The evaluation of this new method was conducted on different malware datasets.

To reconstruct attack scenarios and draw conclusions based on the evidence in the alert stream, Barzegar and Shajari (2018) use the DARPA2000 and MACCDC 2012 dataset for their research. Giudici and Raffinetti (2020) proposed a rank-based statistical model aimed at predicting the severity levels of cyber risk. The model used cyber risk data from the University of Milan. In contrast to the previous datasets, Skrjanc et al. (2018) used the older dataset KDD99 to monitor large-scale cyberattacks using a cauchy clustering method.

Amin et al. (2021) used a cyberattack dataset from the Canadian Institute for Cybersecurity to identify spatial clusters of countries with high rates of cyberattacks. In the context of cybercrime, Junger et al. (2020) examined crime scripts, key characteristics of the target company and the relationship between criminal effort and financial benefit. For their study, the authors analysed 300 cases of fraudulent activities against Dutch companies. With a similar focus on cybercrime, Mireles et al. (2019) proposed a metric framework to measure the effectiveness of the dynamic evolution of cyberattacks and defensive measures. To validate its usefulness, they used the DEFCON dataset.

Due to the rapidly changing nature of cyber risks, it is often impossible to obtain all information on them. Kim and Kim (2019) proposed an automated dataset generation system called CTIMiner that collects threat data from publicly available security reports and malware repositories. They released a dataset to the public containing about 640,000 records from 612 security reports published between January 2008 and 2019. A similar approach is proposed by Kim et al. (2020), using a named entity recognition system to extract core information from cyber threat reports automatically. They created a 498,000-tag dataset during their research (Ulven and Wangen 2021).

Within the framework of vulnerabilities and cybersecurity issues, Ulven and Wangen (2021) proposed an overview of mission-critical assets and everyday threat events, suggested a generic threat model, and summarised common cybersecurity vulnerabilities. With a focus on hospitality, Chen and Fiscus (2018) proposed several issues related to cybersecurity in this sector. They analysed 76 security incidents from the Privacy Rights Clearinghouse database. Supplementary Table 1 lists all findings that belong to the cyber causes dataset.

Impact datasets

This section outlines selected findings of the cyber impact dataset. For cyber insurers, these datasets can form an important basis for information, as they can be used to calculate cyber insurance premiums, evaluate specific cyber risks, formulate inclusions and exclusions in cyber wordings, and re-evaluate as well as supplement the data collected so far on cyber risks. For example, information on financial losses can help to better assess the loss potential of cyber risks. Furthermore, the datasets can provide insight into the frequency of occurrence of these cyber risks. The new datasets can be used to close any data gaps that were previously based on very approximate estimates or to find new results.

Eight studies addressed the costs of data breaches. For instance, Eling and Jung (2018) reviewed 3327 data breach events from 2005 to 2016 and identified an asymmetric dependence of monthly losses by breach type and industry. The authors used datasets from the Privacy Rights Clearinghouse for analysis. The Privacy Rights Clearinghouse datasets and the Breach level index database were also used by De Giovanni et al. (2020) to describe relationships between data breaches and bitcoin-related variables using the cointegration methodology. The data were obtained from the Department of Health and Human Services of healthcare facilities reporting data breaches and a national database of technical and organisational infrastructure information. Also in the context of data breaches, Algarni et al. (2021) developed a comprehensive, formal model that estimates the two components of security risks: breach cost and the likelihood of a data breach within 12 months. For their survey, the authors used two industrial reports from the Ponemon institute and VERIZON. To illustrate the scope of data breaches, Neto et al. (2021) identified 430 major data breach incidents among more than 10,000 incidents. The database created is available and covers the period 2018 to 2019.

With a direct focus on insurance, Biener et al. (2015) analysed 994 cyber loss cases from an operational risk database and investigated the insurability of cyber risks based on predefined criteria. For their study, they used data from the company SAS OpRisk Global Data. Similarly, Eling and Wirfs (2019) looked at a wide range of cyber risk events and actual cost data using the same database. They identified cyber losses and analysed them using methods from statistics and actuarial science. Using a similar reference, Farkas et al. (2021) proposed a method for analysing cyber claims based on regression trees to identify criteria for classifying and evaluating claims. Similar to Chen and Fiscus (2018), the dataset used was the Privacy Rights Clearinghouse database. Within the framework of reinsurance, Moro (2020) analysed cyber index-based information technology activity to see if index-parametric reinsurance coverage could suggest its cedant using data from a Symantec dataset.

Paté-Cornell et al. (2018) presented a general probabilistic risk analysis framework for cybersecurity in an organisation to be specified. The results are distributions of losses to cyberattacks, with and without considered countermeasures in support of risk management decisions based both on past data and anticipated incidents. The data used were from The Common Vulnerability and Exposures database and via confidential access to a database of cyberattacks on a large, U.S.-based organisation. A different conceptual framework for cyber risk classification and assessment was proposed by Sheehan et al. (2021). This framework showed the importance of proactive and reactive barriers in reducing companies' exposure to cyber risk and quantifying the risk. Another approach to cyber risk assessment and mitigation was proposed by Mukhopadhyay et al. (2019). They estimated the probability of an attack using generalised linear models, predicted the security technology required to reduce the probability of cyberattacks, and used gamma and exponential distributions to best approximate the average loss data for each malicious attack. They also calculated the expected loss due to cyberattacks, calculated the net premium that would need to be charged by a cyber insurer, and suggested cyber insurance as a strategy to minimise losses. They used the CSI-FBI survey (1997–2010) to conduct their research.

In order to highlight the lack of data on cyber risks, Eling (2020) conducted a literature review in the areas of cyber risk and cyber insurance. Available information on the frequency, severity, and dependency structure of cyber risks was filtered out. In addition, open questions for future cyber risk research were set up. Another example of data collection on the impact of cyberattacks is provided by Sornette et al. (2013), who use a database of newspaper articles, press reports and other media to provide a predictive method to identify triggering events and potential accident scenarios and estimate their severity and frequency. A similar approach to data collection was used by Arcuri et al. (2020) to gather an original sample of global cyberattacks from newspaper reports sourced from the LexisNexis database. This collection is also used and applied to the fields of dynamic communication and cyber risk perception by Fang et al. (2021). To create a dataset of cyber incidents and disputes, Valeriano and Maness (2014) collected information on cyber interactions between rival states.

To assess trends and the scale of economic cybercrime, Levi (2017) examined datasets from different countries and their impact on crime policy. Pooser et al. (2018) investigated the trend in cyber risk identification from 2006 to 2015 and company characteristics related to cyber risk perception. The authors used a dataset of various reports from cyber insurers for their study. Walker-Roberts et al. (2020) investigated the spectrum of risk of a cybersecurity incident taking place in the cyber-physical-enabled world using the VERIS Community Database. The datasets of impacts identified are presented below. Due to overlap, some may also appear in the causes dataset (Supplementary Table 2).

Cybersecurity datasets

General intrusion detection

General intrusion detection systems account for the largest share of countermeasure datasets. For companies or researchers focused on cybersecurity, the datasets can be used to test their own countermeasures or obtain information about potential vulnerabilities. For example, Al-Omari et al. (2021) proposed an intelligent intrusion detection model for predicting and detecting attacks in cyberspace, which was applied to dataset UNSW-NB 15. A similar approach was taken by Choras and Kozik (2015), who used machine learning to detect cyberattacks on web applications. To evaluate their method, they used the HTTP dataset CSIC 2010. For the identification of unknown attacks on web servers, Kamarudin et al. (2017) proposed an anomaly-based intrusion detection system using an ensemble classification approach. Ganeshan and Rodrigues (2020) showed an intrusion detection system approach, which clusters the database into several groups and detects the presence of intrusion in the clusters. In comparison, AlKadi et al. (2019) used a localisation-based model to discover abnormal patterns in network traffic. Hybrid models have been recommended by Bhattacharya et al. (2020) and Agrawal et al. (2019); the former is a machine-learning model based on principal component analysis for the classification of intrusion detection system datasets, while the latter is a hybrid ensemble intrusion detection system for anomaly detection using different datasets to detect patterns in network traffic that deviate from normal behaviour.

Agarwal et al. (2021) used three different machine learning algorithms in their research to find the most suitable for efficiently identifying patterns of suspicious network activity. The UNSW-NB15 dataset was used for this purpose. Kasongo and Sun (2020), Feed-Forward Deep Neural Network (FFDNN), Keshk et al. (2021), the privacy-preserving anomaly detection framework, and others also use the UNSW-NB 15 dataset as part of intrusion detection systems. The same dataset and others were used by Binbusayyis and Vaiyapuri (2019) to identify and compare key features for cyber intrusion detection. Atefinia and Ahmadi (2021) proposed a deep neural network model to reduce the false positive rate of an anomaly-based intrusion detection system. Fossaceca et al. (2015) focused in their research on the development of a framework that combined the outputs of multiple learners in order to improve the efficacy of network intrusion, and Gauthama Raman et al. (2020) presented a search algorithm based on Support Vector machine to improve the performance of the detection and false alarm rate to improve intrusion detection techniques. Ahmad and Alsemmeari (2020) targeted extreme learning machine techniques due to their good capabilities in classification problems and handling huge data. They used the NSL-KDD dataset as a benchmark.

With reference to prediction, Bakdash et al. (2018) used datasets from the U.S. Department of Defence to predict cyberattacks by malware. This dataset consists of weekly counts of cyber events over approximately seven years. Another prediction method was presented by Fan et al. (2018), which showed an improved integrated cybersecurity prediction method based on spatial-time analysis. Also, with reference to prediction, Ashtiani and Azgomi (2014) proposed a framework for the distributed simulation of cyberattacks based on high-level architecture. Kirubavathi and Anitha (2016) recommended an approach to detect botnets, irrespective of their structures, based on network traffic flow behaviour analysis and machine-learning techniques. Dwivedi et al. (2021) introduced a multi-parallel adaptive technique to utilise an adaption mechanism in the group of swarms for network intrusion detection. AlEroud and Karabatis (2018) presented an approach that used contextual information to automatically identify and query possible semantic links between different types of suspicious activities extracted from network flows.

Intrusion detection systems with a focus on IoT

In addition to general intrusion detection systems, a proportion of studies focused on IoT. Habib et al. (2020) presented an approach for converting traditional intrusion detection systems into smart intrusion detection systems for IoT networks. To enhance the process of diagnostic detection of possible vulnerabilities with an IoT system, Georgescu et al. (2019) introduced a method that uses a named entity recognition-based solution. With regard to IoT in the smart home sector, Heartfield et al.

(2021) presented a detection system that is able to autonomously adjust the decision function of its underlying anomaly classification models to a smart home's changing condition. Another intrusion detection system was suggested by Keserwani et al. (2021), which combined Grey Wolf Optimization and Particle Swam Optimization to identify various attacks for IoT networks. They used the KDD Cup 99, NSL-KDD and CICIDS-2017 to evaluate their model. Abu Al-Haija and Zein-Sabatto (2020) provide a comprehensive development of a new intelligent and autonomous deep-learning-based detection and classification system for cyberattacks in IoT communication networks that leverage the power of convolutional neural networks, abbreviated as IoT-IDCS-CNN (IoT-based Intrusion Detection and Classification System using Convolutional Neural Network). To evaluate the development, the authors used the NSL-KDD dataset. Biswas and Roy (2021) recommended a model that identifies malicious botnet traffic using novel deep-learning approaches like artificial neural networks gutted recurrent units and long- or short-term memory models. They tested their model with the Bot-IoT dataset.

With a more forensic background, Koroniotis et al. (2020) submitted a network forensic framework, which described the digital investigation phases for identifying and tracing attack behaviours in IoT networks. The suggested work was evaluated with the Bot-IoT and UINSW-NB15 datasets. With a focus on big data and IoT, Chhabra et al. (2020) presented a cyber forensic framework for big data analytics in an IoT environment using machine learning. Furthermore, the authors mentioned different publicly available datasets for machine-learning models.

A stronger focus on a mobile phones was exhibited by Alazab et al. (2020), which presented a classification model that combined permission requests and application programme interface calls. The model was tested with a malware dataset containing 27,891 Android apps. A similar approach was taken by Li et al. (2019a, b), who proposed a reliable classifier for Android malware detection based on factorisation machine architecture and extraction of Android app features from manifest files and source code.

Literature reviews

In addition to the different methods and models for intrusion detection systems, various literature reviews on the methods and datasets were also found. Liu and Lang (2019) proposed a taxonomy of intrusion detection systems that uses data objects as the main dimension to classify and summarise machine learning and deep learning-based intrusion detection literature. They also presented four different benchmark datasets for machine-learning detection systems. Ahmed et al. (2016) presented an in-depth analysis of four major categories of anomaly detection techniques, which include classification, statistical, information theory and clustering. Hajj et al. (2021) gave a comprehensive overview of anomaly-based intrusion detection systems. Their article gives an overview of the requirements, methods, measurements and datasets that are used in an intrusion detection system.

Within the framework of machine learning, Chattopadhyay et al. (2018) conducted a comprehensive review and meta-analysis on the application of machine-learning techniques in intrusion detection systems. They also compared different machine learning techniques in different datasets and summarised the performance. Vidros et al. (2017) presented an overview of characteristics and methods in automatic detection of online recruitment fraud. They also published an available dataset of 17,880 annotated job ads, retrieved from the use of a real-life system. An empirical study of different unsupervised learning algorithms used in the detection of unknown attacks was presented by Meira et al. (2020).

New datasets

Kilincer et al. (2021) reviewed different intrusion detection system datasets in detail. They had a closer look at the UNS-NB15, ISCX-2012, NSL-KDD and CIDDS-001 datasets. Stojanovic et al. (2020) also provided a review on datasets and their creation for use in advanced persistent threat detection in the literature. Another review of datasets was provided by Sarker et al. (2020), who focused on cybersecurity data science as part of their research and provided an overview from a machine-learning perspective. Avila et al. (2021) conducted a systematic literature review on the use of security logs for data leak detection. They recommended a new classification of information leak, which uses the GDPR principles, identified the most widely publicly available dataset for threat detection, described the attack types in the datasets and the algorithms used for data leak detection. Tuncer et al. (2020) presented a bytecode-based detection method consisting of feature extraction using local neighbourhood binary patterns. They chose a byte-based malware dataset to investigate the performance of the proposed local neighbourhood binary pattern-based detection method. With a different focus, Mauro et al. (2020) gave an experimental overview of neural-based techniques relevant to intrusion detection. They assessed the value of neural networks using the Bot-IoT and UNSW-DB15 datasets.

Another category of results in the context of countermeasure datasets is those that were presented as new. Moreno et al. (2018) developed a database of 300 security-related accidents from European and American sources. The database contained cybersecurity-related events in the chemical and process industry. Damasevicius et al. (2020) proposed a new dataset (LITNET-2020) for network intrusion detection. The dataset is a new annotated network benchmark dataset obtained from the real-world academic network. It presents real-world examples of normal and under-attack network traffic. With a focus on IoT intrusion detection systems, Alsaedi et al. (2020) proposed a new benchmark IoT/IIot datasets for assessing intrusion detection system-enabled IoT systems. Also in the context of IoT, Vaccari et al. (2020) proposed a dataset focusing on message queue telemetry transport protocols, which can be used to train machine-learning models. To evaluate the performance of machine-learning classifiers, Mahfouz et al. (2020) created a dataset called Game Theory and Cybersecurity (GTCS). A dataset containing 22,000 malware and benign samples was constructed by Martin et al. (2019). The dataset can be used as a benchmark to test the algorithm for Android malware classification and clustering techniques. In addition, Laso et al. (2017) presented a dataset created to investigate how data and information quality estimates enable the detection of anomalies and malicious acts in cyber-physical systems. The dataset contained various cyberattacks and is publicly available.

Other

In addition to the results described above, several other studies were found that fit into the category of countermeasures. Johnson et al. (2016) examined the time between vulnerability disclosures. Using another vulnerabilities database, Common Vulnerabilities and Exposures (CVE), Subroto and Apriyana (2019) presented an algorithm model that uses big data analysis of social media and statistical machine learning to predict cyber risks. A similar databank but with a different focus, Common Vulnerability Scoring System, was used by Chatterjee and Thekdi (2020) to present an iterative data-driven learning approach to vulnerability assessment and management for complex systems. Using the CICIDS2017 dataset to evaluate the performance, Malik et al. (2020) proposed a control plane-based orchestration for varied, sophisticated threats and attacks. The same dataset was used in another study by Lee et al. (2019), who developed an artificial security information event management system based on a combination of event profiling for data processing and different artificial network methods. To exploit the interdependence between multiple series, Fang et al. (2021) proposed a statistical framework. In order to validate the framework, the authors applied it to a dataset of enterprise-level security breaches from the Privacy Rights Clearinghouse and Identity Theft Center database. Another framework with a defensive aspect was recommended by Li et al. (2021) to increase the robustness of deep neural networks against adversarial malware evasion attacks. Sarabi et al. (2016) investigated whether and to what extent business details can help assess an organisation's risk of data breaches and the distribution of risk across different types of incidents to create policies for protection, detection and recovery from different forms of security incidents. They used data from the VERIS Community Database.

Datasets that have been classified into the cybersecurity category are detailed in Supplementary Table 3. Due to overlap, records from the previous tables may also be included.

Discussion

This paper presented a systematic literature review of studies on cyber risk and cybersecurity that used datasets. Within this framework, 255 studies were fully reviewed and then classified into three different categories. Then, 79 datasets were consolidated from these studies. These datasets were subsequently analysed, and important information was selected through a process of filtering out. This information was recorded in a table and enhanced with further information as part of the literature analysis. This made it possible to create a comprehensive overview of the datasets. For example, each dataset contains a description of where the data came from and how the data has been used to date. This allows different datasets to be compared and the appropriate dataset for the use case to be selected. This research certainly has limitations, so our selection of datasets cannot necessarily be taken as a representation of all available datasets related to cyber risks and cybersecurity. For example, literature searches were conducted in four academic databases and only found datasets that were used in the literature. Many research projects also used old datasets that may no longer consider current developments. In addition, the data are often focused on only one observation and are limited in scope. For example, the datasets can only be applied to specific contexts and are also subject to further limitations (e.g. region, industry, operating system). In the context of the applicability of the datasets, it is unfortunately not possible to make a clear statement on the extent to which they can be integrated into academic or practical areas of application or how great this effort is. Finally, it remains to be pointed out that this is an overview of currently available datasets, which are subject to constant change.

Due to the lack of datasets on cyber risks in the academic literature, additional datasets on cyber risks were integrated as part of a further search. The search was conducted on the Google Dataset search portal. The search term used was ‘cyber risk datasets’. Over 100 results were found. However, due to the low significance and verifiability, only 20 selected datasets were included. These can be found in Table 2 in the “Appendix”.

The results of the literature review and datasets also showed that there continues to be a lack of available, open cyber datasets. This lack of data is reflected in cyber insurance, for example, as it is difficult to find a risk-based premium without a sufficient database (Nurse et al. 2020). The global cyber insurance market was estimated at USD 5.5 billion in 2020 (Dyson 2020). When compared to the USD 1 trillion global losses from cybercrime (Maleks Smith et al. 2020), it is clear that there exists a significant cyber risk awareness challenge for both the insurance industry and international commerce. Without comprehensive and qualitative data on cyber losses, it can be difficult to estimate potential losses from cyberattacks and price cyber insurance accordingly (GAO 2021). For instance, the average cyber insurance loss increased from USD 145,000 in 2019 to USD 359,000 in 2020 (FitchRatings 2021). Cyber insurance is an important risk management tool to mitigate the financial impact of cybercrime. This is particularly evident in the impact of different industries. In the Energy & Commodities financial markets, a ransomware attack on the Colonial Pipeline led to a substantial impact on the U.S. economy. As a result of the attack, about 45% of the U.S. East Coast was temporarily unable to obtain supplies of diesel, petrol and jet fuel. This caused the average price in the U.S. to rise 7 cents to USD 3.04 per gallon, the highest in seven years (Garber 2021). In addition, Colonial Pipeline confirmed that it paid a USD 4.4 million ransom to a hacker gang after the attack. Another ransomware attack occurred in the healthcare and government sector. The victim of this attack was the Irish Health Service Executive (HSE). A ransom payment of USD 20 million was demanded from the Irish government to restore services after the hack (Tidy 2021). In the car manufacturing sector, Miller and Valasek (2015) initiated a cyberattack that resulted in the recall of 1.4 million vehicles and cost manufacturers EUR 761 million. The risk that arises in the context of these events is the potential for the accumulation of cyber losses, which is why cyber insurers are not expanding their capacity. An example of this accumulation of cyber risks is the NotPetya malware attack, which originated in Russia, struck in Ukraine, and rapidly spread around the world, causing at least USD 10 billion in damage (GAO 2021). These events highlight the importance of proper cyber risk management.

This research provides cyber insurance stakeholders with an overview of cyber datasets. Cyber insurers can use the open datasets to improve their understanding and assessment of cyber risks. For example, the impact datasets can be used to better measure financial impacts and their frequencies. These data could be combined with existing portfolio data from cyber insurers and integrated with existing pricing tools and factors to better assess cyber risk valuation. Although most cyber insurers have sparse historical cyber policy and claims data, they remain too small at present for accurate prediction (Bessy-Roland et al. 2021). A combination of portfolio data and external datasets would support risk-adjusted pricing for cyber insurance, which would also benefit policyholders. In addition, cyber insurance stakeholders can use the datasets to identify patterns and make better predictions, which would benefit sustainable cyber insurance coverage. In terms of cyber risk cause datasets, cyber insurers can use the data to review their insurance products. For example, the data could provide information on which cyber risks have not been sufficiently considered in product design or where improvements are needed. A combination of cyber cause and cybersecurity datasets can help establish uniform definitions to provide greater transparency and clarity. Consistent terminology could lead to a more sustainable cyber market, where cyber insurers make informed decisions about the level of coverage and policyholders understand their coverage (The Geneva Association 2020).

In addition to the cyber insurance community, this research also supports cybersecurity stakeholders. The reviewed literature can be used to provide a contemporary, contextual and categorised summary of available datasets. This supports efficient and timely progress in cyber risk research and is beneficial given the dynamic nature of cyber risks. With the help of the described cybersecurity datasets and the identified information, a comparison of different datasets is possible. The datasets can be used to evaluate the effectiveness of countermeasures in simulated cyberattacks or to test intrusion detection systems.

Conclusion

In this paper, we conducted a systematic review of studies on cyber risk and cybersecurity databases. We found that most of the datasets are in the field of intrusion detection and machine learning and are used for technical cybersecurity aspects. The available datasets on cyber risks were relatively less represented. Due to the dynamic nature and lack of historical data, assessing and understanding cyber risk is a major challenge for cyber insurance stakeholders. To address this challenge, a greater density of cyber data is needed to support cyber insurers in risk management and researchers with cyber risk-related topics. With reference to ‘Open Science’ FAIR data (Jacobsen et al. 2020), mandatory reporting of cyber incidents could help improve cyber understanding, awareness and loss prevention among companies and insurers. Through greater availability of data, cyber risks can be better understood, enabling researchers to conduct more in-depth research into these risks. Companies could incorporate this new knowledge into their corporate culture to reduce cyber risks. For insurance companies, this would have the advantage that all insurers would have the same understanding of cyber risks, which would support sustainable risk-based pricing. In addition, common definitions of cyber risks could be derived from new data.

The cybersecurity databases summarised and categorised in this research could provide a different perspective on cyber risks that would enable the formulation of common definitions in cyber policies. The datasets can help companies addressing cybersecurity and cyber risk as part of risk management assess their internal cyber posture and cybersecurity measures. The paper can also help improve risk awareness and corporate behaviour, and provides the research community with a comprehensive overview of peer-reviewed datasets and other available datasets in the area of cyber risk and cybersecurity. This approach is intended to support the free availability of data for research. The complete tabulated review of the literature is included in the Supplementary Material.

This work provides directions for several paths of future work. First, there are currently few publicly available datasets for cyber risk and cybersecurity. The older datasets that are still widely used no longer reflect today's technical environment. Moreover, they can often only be used in one context, and the scope of the samples is very limited. It would be of great value if more datasets were publicly available that reflect current environmental conditions. This could help intrusion detection systems to consider current events and thus lead to a higher success rate. It could also compensate for the disadvantages of older datasets by collecting larger quantities of samples and making this contextualisation more widespread. Another area of research may be the integratability and adaptability of cybersecurity and cyber risk datasets. For example, it is often unclear to what extent datasets can be integrated or adapted to existing data. For cyber risks and cybersecurity, it would be helpful to know what requirements need to be met or what is needed to use the datasets appropriately. In addition, it would certainly be helpful to know whether datasets can be modified to be used for cyber risks or cybersecurity. Finally, the ability for stakeholders to identify machine-readable cybersecurity datasets would be useful because it would allow for even clearer delineations or comparisons between datasets. Due to the lack of publicly available datasets, concrete benchmarks often cannot be applied.

Appendix

Table 2 Summary of Google datasets

No	Dataset creator	Name of the dataset	Data availability	Year of creation/start year	Description
1	ActionFraud	Cyber Crime Dashboard	Public	2020	Shows cybercrime and fraud reported in the U.K..
2	Carlos E. Jimenez-Gomez	Data Breaches 2004–2017	Public	2018	Includes 270 records and 11 variables of data breaches. The data breaches happened between 2004–2017. Only data breaches with over 30,000 records are considered.
3	Chubb	Chubb Cyber Index	Public	2007	Shows cyber claims for more than two decades. In this dashboard, there is the possibility to get information about different areas regarding claims cost. Furthermore, it is possible to get an overview of claims of different years.
4	CMS	DGDPR Enforcement Tracker	Public	2018	An overview of fines and penalties, which data protection authorities within the EU have imposed under the EU GDPR.
5	DSGVO Portal	DSGVO—Portal	Public	2014	Fines for violations of the GDPR and other data protection laws.

Table 2 (continued)

No	Dataset creator	Name of the dataset	Data availability	Year of creation/start year	Description
6	Federal Bureau of Investigation	Internet Crime Report 2020	Public	2021	Includes the cyber risk impact situation in the U.S..
7	Government of Canada	No name	Public	2017	Percentage of enterprises impacted by specific types of cybersecurity incidents by the North American Industry Classification System (NAICS) and size of the enterprise.
8	Hiscox	Hisco Cyber Readiness Report 2020	Public	2020	The average cost of all cyberattacks to firms from Europe and the U.S. in 2020, by size, in USD.
9	IBM Security	Cost of a Data Breach Report 2020	Public	2020	Includes the cost of data breaches from 2014 to 2020.
10	Information is beautiful	World's Biggest Data Breaches & Hacks	Public	2004	Selected events over 30,000 records.
11	Ipsos Mori	Cyber Security Breaches Survey	Public	2020	Displays the share of businesses that have had certain outcomes after experiencing a cybersecurity breach or attack in the last 12 months in the U.K. in 2020
12	Kaspersky	Damage Control: The Cost of Security Breaches	Public	2020	Analyses the different data of Kaspersky

Table 2 (continued)

No	Dataset creator	Name of the dataset	Data availability	Year of creation/start year	Description
13	Marsch—Mircosoft—Global Cyber Risk Perception Survey	Marsch—Mircosoft—Global Cyber Risk Perception Survey	Public	2018	Presents the greatest potential imp.acts to an organisation due to cyber loss scenarios, according to senior executives
14	Mendeley Data	California Data Breach Notification Data	Public	2019	An empirical study of security breach notifications filed in California during 2012–2016.
15	Norton	2019 Cyber Safety Insights Report	Public	2020	A survey of internet users who have experienced an internet crime.
16	Paolo Passeri	Hackmageddon	Access controlled	2011	Overview of collected timelines with a focus on cyberattacks.
17	Pierangelo and Theo	Data Breach Dataset	Public	2020	Consists of 506 data breaches and associated characteristics that affected U.S.-listed companies over a 10-year period from April 2005 to March 2015. The dataset was gathered from the Privacy Rights Clearinghouse (PRC) and then augmented with manual data collection.

Table 2 (continued)

No	Dataset creator	Name of the dataset	Data availability	Year of creation/start year	Description
18	PwC	2015 Information Security Breaches Survey	Public	2015	Illustrates the ranking of what made a particular security breach incident the worst of the year in the U.K. in 2015.
19	Spy Cloud	Spy Cloud	Private	-	-
20	Willis Towers Watson	Cyber claims analysis report	Public	2020	Uses analysed claims data of Willis Towers Watson to provide specific insight.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1057/s41288-022-00266-6.

Funding Open Access funding provided by the IReL Consortium.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aamir, M., S.S.H. Rizvi, M.A. Hashmani, M. Zubair, and J. Ahmad. 2021. Machine learning classification of port scanning and DDoS attacks: A comparative analysis. *Mehran University Research Journal of Engineering and Technology* 40 (1): 215–229. https://doi.org/10.22581/muet1982.2101.19.

Aamir, M., and S.M.A. Zaidi. 2019. DDoS attack detection with feature engineering and machine learning: The framework and performance evaluation. *International Journal of Information Security* 18 (6): 761–785. https://doi.org/10.1007/s10207-019-00434-1.

Aassal, A. El, S. Baki, A. Das, and R.M. Verma. 2020. 2020. An in-depth benchmarking and evaluation of phishing detection research for security needs. *IEEE Access* 8: 22170–22192. https://doi.org/10.1109/ACCESS.2020.2969780.

Abu Al-Haija, Q., and S. Zein-Sabatto. 2020. An efficient deep-learning-based detection and classification system for cyber-attacks in IoT communication networks. *Electronics* 9 (12): 26. https://doi.org/10.3390/electronics9122152.

Adhikari, U., T.H. Morris, and S.Y. Pan. 2018. Applying Hoeffding adaptive trees for real-time cyber-power event and intrusion classification. *IEEE Transactions on Smart Grid* 9 (5): 4049–4060. https://doi.org/10.1109/tsg.2017.2647778.

Agarwal, A., P. Sharma, M. Alshehri, A.A. Mohamed, and O. Alfarraj. 2021. Classification model for accuracy and intrusion detection using machine learning approach. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.437.

Agrafiotis, I., J.R.C.. Nurse, M. Goldsmith, S. Creese, and D. Upton. 2018. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity* 4: tyy006.

Agrawal, A., S. Mohammed, and J. Fiaidhi. 2019. Ensemble technique for intruder detection in network traffic. *International Journal of Security and Its Applications* 13 (3): 1–8. https://doi.org/10.33832/ijsia.2019.13.3.01.

Ahmad, I., and R.A. Alsemmeari. 2020. Towards improving the intrusion detection through ELM (extreme learning machine). *CMC Computers Materials & Continua* 65 (2): 1097–1111. https://doi.org/10.32604/cmc.2020.011732.

Ahmed, M., A.N. Mahmood, and J.K. Hu. 2016. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* 60: 19–31. https://doi.org/10.1016/j.jnca.2015.11.016.

Al-Jarrah, O.Y., O. Alhussein, P.D. Yoo, S. Muhaidat, K. Taha, and K. Kim. 2016. Data randomization and cluster-based partitioning for Botnet intrusion detection. *IEEE Transactions on Cybernetics* 46 (8): 1796–1806. https://doi.org/10.1109/TCYB.2015.2490802.

Al-Mhiqani, M.N., R. Ahmad, Z.Z. Abidin, W. Yassin, A. Hassan, K.H. Abdulkareem, N.S. Ali, and Z. Yunos. 2020. A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations. *Applied Sciences—Basel* 10 (15): 41. https://doi.org/10.3390/app10155208.

Al-Omari, M., M. Rawashdeh, F. Qutaishat, M. Alshira'H, and N. Ababneh. 2021. An intelligent tree-based intrusion detection model for cyber security. *Journal of Network and Systems Management* 29 (2): 18. https://doi.org/10.1007/s10922-021-09591-y.

Alabdallah, A., and M. Awad. 2018. Using weighted Support Vector Machine to address the imbalanced classes problem of Intrusion Detection System. *KSII Transactions on Internet and Information Systems* 12 (10): 5143–5158. https://doi.org/10.3837/tiis.2018.10.027.

Alazab, M., M. Alazab, A. Shalaginov, A. Mesleh, and A. Awajan. 2020. Intelligent mobile malware detection using permission requests and API calls. *Future Generation Computer Systems—the International Journal of eScience* 107: 509–521. https://doi.org/10.1016/j.future.2020.02.002.

Albahar, M.A., R.A. Al-Falluji, and M. Binsawad. 2020. An empirical comparison on malicious activity detection using different neural network-based models. *IEEE Access* 8: 61549–61564. https://doi.org/10.1109/ACCESS.2020.2984157.

AlEroud, A.F., and G. Karabatis. 2018. Queryable semantics to detect cyber-attacks: A flow-based detection approach. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48 (2): 207–223. https://doi.org/10.1109/TSMC.2016.2600405.

Algarni, A.M., V. Thayananthan, and Y.K. Malaiya. 2021. Quantitative assessment of cybersecurity risks for mitigating data breaches in business systems. *Applied Sciences (switzerland)*. https://doi.org/10.3390/app11083678.

Alhowaide, A., I. Alsmadi, and J. Tang. 2021. Towards the design of real-time autonomous IoT NIDS. *Cluster Computing—the Journal of Networks Software Tools and Applications*. https://doi.org/10.1007/s10586-021-03231-5.

Ali, S., and Y. Li. 2019. Learning multilevel auto-encoders for DDoS attack detection in smart grid network. *IEEE Access* 7: 108647–108659. https://doi.org/10.1109/ACCESS.2019.2933304.

AlKadi, O., N. Moustafa, B. Turnbull, and K.K.R. Choo. 2019. Mixture localization-based outliers models for securing data migration in cloud centers. *IEEE Access* 7: 114607–114618. https://doi.org/10.1109/ACCESS.2019.2935142.

Allianz. 2021. Allianz Risk Barometer. https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/Allianz-Risk-Barometer-2021.pdf. Accessed 15 May 2021.

Almiani, M., A. AbuGhazleh, A. Al-Rahayfeh, S. Atiewi, and Razaque, A. 2020. Deep recurrent neural network for IoT intrusion detection system. *Simulation Modelling Practice and Theory* 101: 102031. https://doi.org/10.1016/j.simpat.2019.102031

Alsaedi, A., N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar. 2020. TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access* 8: 165130–165150. https://doi.org/10.1109/access.2020.3022862.

Alsamiri, J., and K. Alsubhi. 2019. Internet of Things cyber attacks detection using machine learning. *International Journal of Advanced Computer Science and Applications* 10 (12): 627–634.

Alsharafat, W. 2013. Applying artificial neural network and eXtended classifier system for network intrusion detection. *International Arab Journal of Information Technology* 10 (3): 230–238.

Amin, R.W., H.E. Sevil, S. Kocak, G. Francia III., and P. Hoover. 2021. The spatial analysis of the malicious uniform resource locators (URLs): 2016 dataset case study. *Information (switzerland)* 12 (1): 1–18. https://doi.org/10.3390/info12010002.

Arcuri, M.C., L.Z. Gai, F. Ielasi, and E. Ventisette. 2020. Cyber attacks on hospitality sector: Stock market reaction. *Journal of Hospitality and Tourism Technology* 11 (2): 277–290. https://doi.org/10.1108/jhtt-05-2019-0080.

Arp, D., M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C.E.R.T. Siemens. 2014. Drebin: Effective and explainable detection of android malware in your pocket. *In Ndss* 14: 23–26.

Ashtiani, M., and M.A. Azgomi. 2014. A distributed simulation framework for modeling cyber attacks and the evaluation of security measures. *Simulation* 90 (9): 1071–1102. https://doi.org/10.1177/0037549714540221.

Atefinia, R., and M. Ahmadi. 2021. Network intrusion detection using multi-architectural modular deep neural network. *Journal of Supercomputing* 77 (4): 3571–3593. https://doi.org/10.1007/s11227-020-03410-y.

Avila, R., R. Khoury, R. Khoury, and F. Petrillo. 2021. Use of security logs for data leak detection: A systematic literature review. *Security and Communication Networks* 2021: 29. https://doi.org/10.1155/2021/6615899.

Azeez, N.A., T.J. Ayemobola, S. Misra, R. Maskeliunas, and R. Damasevicius. 2019. Network Intrusion Detection with a Hashing Based Apriori Algorithm Using Hadoop MapReduce. *Computers* 8 (4): 15. https://doi.org/10.3390/computers8040086.

Bakdash, J.Z., S. Hutchinson, E.G. Zaroukian, L.R. Marusich, S. Thirumuruganathan, C. Sample, B. Hoffman, and G. Das. 2018. Malware in the future forecasting of analyst detection of cyber events. *Journal of Cybersecurity*. https://doi.org/10.1093/cybsec/tyy007.

Barletta, V.S., D. Caivano, A. Nannavecchia, and M. Scalera. 2020. Intrusion detection for in-vehicle communication networks: An unsupervised Kohonen SOM approach. *Future Internet*. https://doi.org/10.3390/FI12070119.

Barzegar, M., and M. Shajari. 2018. Attack scenario reconstruction using intrusion semantics. *Expert Systems with Applications* 108: 119–133. https://doi.org/10.1016/j.eswa.2018.04.030.

Bessy-Roland, Y., A. Boumezoued, and C. Hillairet. 2021. Multivariate Hawkes process for cyber insurance. *Annals of Actuarial Science* 15 (1): 14–39.

Bhardwaj, A., V. Mangat, and R. Vig. 2020. Hyperband tuned deep neural network with well posed stacked sparse AutoEncoder for detection of DDoS attacks in cloud. *IEEE Access* 8: 181916–181929. https://doi.org/10.1109/ACCESS.2020.3028690.

Bhati, B.S., C.S. Rai, B. Balamurugan, and F. Al-Turjman. 2020. An intrusion detection scheme based on the ensemble of discriminant classifiers. *Computers & Electrical Engineering* 86: 9. https://doi.org/10.1016/j.compeleceng.2020.106742.

Bhattacharya, S., S.S.R. Krishnan, P.K.R. Maddikunta, R. Kaluri, S. Singh, T.R. Gadekallu, M. Alazab, and U. Tariq. 2020. A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU. *Electronics* 9 (2): 16. https://doi.org/10.3390/electronics9020219.

Bibi, I., A. Akhunzada, J. Malik, J. Iqbal, A. Musaddiq, and S. Kim. 2020. A dynamic DL-driven architecture to combat sophisticated android malware. *IEEE Access* 8: 129600–129612. https://doi.org/10.1109/ACCESS.2020.3009819.

Biener, C., M. Eling, and J.H. Wirfs. 2015. Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance—Issues and Practice* 40 (1): 131–158. https://doi.org/10.1057/gpp.2014.19.

Binbusayyis, A., and T. Vaiyapuri. 2019. Identifying and benchmarking key features for cyber intrusion detection: An ensemble approach. *IEEE Access* 7: 106495–106513. https://doi.org/10.1109/ACCESS.2019.2929487.

Biswas, R., and S. Roy. 2021. Botnet traffic identification using neural networks. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-021-10765-8.

Bouyeddou, B., F. Harrou, B. Kadri, and Y. Sun. 2021. Detecting network cyber-attacks using an integrated statistical approach. *Cluster Computing—the Journal of Networks Software Tools and Applications* 24 (2): 1435–1453. https://doi.org/10.1007/s10586-020-03203-1.

Bozkir, A.S., and M. Aydos. 2020. LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition. *Computers & Security* 95: 18. https://doi.org/10.1016/j.cose.2020.101855.

Brower, D., and M. McCormick. 2021. Colonial pipeline resumes operations following ransomware attack. *Financial Times*.

Cai, H., F. Zhang, and A. Levi. 2019. An unsupervised method for detecting shilling attacks in recommender systems by mining item relationship and identifying target items. *The Computer Journal* 62 (4): 579–597. https://doi.org/10.1093/comjnl/bxy124.

Cebula, J.J., M.E. Popeck, and L.R. Young. 2014. *A Taxonomy of Operational Cyber Security Risks Version 2*.

Chadza, T., K.G. Kyriakopoulos, and S. Lambotharan. 2020. Learning to learn sequential network attacks using hidden Markov models. *IEEE Access* 8: 134480–134497. https://doi.org/10.1109/ACCESS.2020.3011293.

Chatterjee, S., and S. Thekdi. 2020. An iterative learning and inference approach to managing dynamic cyber vulnerabilities of complex systems. *Reliability Engineering and System Safety*. https://doi.org/10.1016/j.ress.2019.106664.

Chattopadhyay, M., R. Sen, and S. Gupta. 2018. A comprehensive review and meta-analysis on applications of machine learning techniques in intrusion detection. *Australasian Journal of Information Systems* 22: 27.

Chen, H.S., and J. Fiscus. 2018. The inhospitable vulnerability: A need for cybersecurity risk assessment in the hospitality industry. *Journal of Hospitality and Tourism Technology* 9 (2): 223–234. https://doi.org/10.1108/JHTT-07-2017-0044.

Chhabra, G.S., V.P. Singh, and M. Singh. 2020. Cyber forensics framework for big data analytics in IoT environment using machine learning. *Multimedia Tools and Applications* 79 (23–24): 15881–15900. https://doi.org/10.1007/s11042-018-6338-1.

Chiba, Z., N. Abghour, K. Moussaid, A. Elomri, and M. Rida. 2019. Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. *Computers and Security* 86: 291–317. https://doi.org/10.1016/j.cose.2019.06.013.

Choras, M., and R. Kozik. 2015. Machine learning techniques applied to detect cyber attacks on web applications. *Logic Journal of the IGPL* 23 (1): 45–56. https://doi.org/10.1093/jigpal/jzu038.

Chowdhury, S., M. Khanzadeh, R. Akula, F. Zhang, S. Zhang, H. Medal, M. Marufuzzaman, and L. Bian. 2017. Botnet detection using graph-based feature clustering. *Journal of Big Data* 4 (1): 14. https://doi.org/10.1186/s40537-017-0074-7.

Cost Of A Cyber Incident: Systematic Review And Cross-Validation, *Cybersecurity & Infrastructure Agency*, 1, https://www.cisa.gov/sites/default/files/publications/CISA-OCE_Cost_of_Cyber_Incidents_Study-FINAL_508.pdf (2020).

D'Hooge, L., T. Wauters, B. Volckaert, and F. De Turck. 2019. Classification hardness for supervised learners on 20 years of intrusion detection data. *IEEE Access* 7: 167455–167469. https://doi.org/10.1109/access.2019.2953451.

Damasevicius, R., A. Venckauskas, S. Grigaliunas, J. Toldinas, N. Morkevicius, T. Aleliunas, and P. Smuikys. 2020. LITNET-2020: An annotated real-world network flow dataset for network intrusion detection. *Electronics* 9 (5): 23. https://doi.org/10.3390/electronics9050800.

De Giovanni, A.L.D., and M. Pirra. 2020. On the determinants of data breaches: A cointegration analysis. *Decisions in Economics and Finance*. https://doi.org/10.1007/s10203-020-00301-y.

Deng, L., D. Li, X. Yao, and H. Wang. 2019. Retracted Article: Mobile network intrusion detection for IoT system based on transfer learning algorithm. *Cluster Computing* 22 (4): 9889–9904. https://doi.org/10.1007/s10586-018-1847-2.

Donkal, G., and G.K. Verma. 2018. A multimodal fusion based framework to reinforce IDS for securing Big Data environment using Spark. *Journal of Information Security and Applications* 43: 1–11. https://doi.org/10.1016/j.jisa.2018.10.001.

Dunn, C., N. Moustafa, and B. Turnbull. 2020. Robustness evaluations of sustainable machine learning models against data Poisoning attacks in the Internet of Things. *Sustainability* 12 (16): 17. https://doi.org/10.3390/su12166434.

Dwivedi, S., M. Vardhan, and S. Tripathi. 2021. Multi-parallel adaptive grasshopper optimization technique for detecting anonymous attacks in wireless networks. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-021-08368-5.

Dyson, B. 2020. COVID-19 crisis could be 'watershed' for cyber insurance, says Swiss Re exec. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/covid-19-crisis-could-be-watershed-for-cyber-insurance-says-swiss-re-exec-59197154. Accessed 7 May 2020.

EIOPA. 2018. Understanding cyber insurance—a structured dialogue with insurance companies. https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa_understanding_cyber_insurance.pdf. Accessed 28 May 2018

Elijah, A.V., A. Abdullah, N.Z. JhanJhi, M. Supramaniam, and O.B. Abdullateef. 2019. Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: An empirical study. *International Journal of Advanced Computer Science and Applications* 10 (9): 520–528.

Eling, M., and K. Jung. 2018. Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance Mathematics & Economics* 82: 167–180. https://doi.org/10.1016/j.insmatheco.2018.07.003.

Eling, M., and W. Schnell. 2016. What do we know about cyber risk and cyber risk insurance? *Journal of Risk Finance* 17 (5): 474–491. https://doi.org/10.1108/jrf-09-2016-0122.

Eling, M., and J. Wirfs. 2019. What are the actual costs of cyber risk events? *European Journal of Operational Research* 272 (3): 1109–1119. https://doi.org/10.1016/j.ejor.2018.07.021.

Eling, M. 2020. Cyber risk research in business and actuarial science. *European Actuarial Journal* 10 (2): 303–333.

Elmasry, W., A. Akbulut, and A.H. Zaim. 2019. Empirical study on multiclass classification-based network intrusion detection. *Computational Intelligence* 35 (4): 919–954. https://doi.org/10.1111/coin.12220.

Elsaid, S.A., and N.S. Albatati. 2020. An optimized collaborative intrusion detection system for wireless sensor networks. *Soft Computing* 24 (16): 12553–12567. https://doi.org/10.1007/s00500-020-04695-0.

Estepa, R., J.E. Díaz-Verdejo, A. Estepa, and G. Madinabeitia. 2020. How much training data is enough? A case study for HTTP anomaly-based intrusion detection. *IEEE Access* 8: 44410–44425. https://doi.org/10.1109/ACCESS.2020.2977591.

European Council. 2021. Cybersecurity: how the EU tackles cyber threats. https://www.consilium.europa.eu/en/policies/cybersecurity/. Accessed 10 May 2021

Falco, G. et al. 2019. Cyber risk research impeded by disciplinary barriers. *Science (American Association for the Advancement of Science)* 366 (6469): 1066–1069.

Fan, Z.J., Z.P. Tan, C.X. Tan, and X. Li. 2018. An improved integrated prediction method of cyber security situation based on spatial-time analysis. *Journal of Internet Technology* 19 (6): 1789–1800. https://doi.org/10.3966/160792642018111906015.

Fang, Z.J., M.C. Xu, S.H. Xu, and T.Z. Hu. 2021. A framework for predicting data breach risk: Leveraging dependence to cope with sparsity. *IEEE Transactions on Information Forensics and Security* 16: 2186–2201. https://doi.org/10.1109/tifs.2021.3051804.

Farkas, S., O. Lopez, and M. Thomas. 2021. Cyber claim analysis using Generalized Pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics* 98: 92–105. https://doi.org/10.1016/j.insmatheco.2021.02.009.

Farsi, H., A. Fanian, and Z. Taghiyarrenani. 2019. A novel online state-based anomaly detection system for process control networks. *International Journal of Critical Infrastructure Protection* 27: 11. https://doi.org/10.1016/j.ijcip.2019.100323.

Ferrag, M.A., L. Maglaras, S. Moschoyiannis, and H. Janicke. 2020. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications* 50: 19. https://doi.org/10.1016/j.jisa.2019.102419.

Field, M. 2018. WannaCry cyber attack cost the NHS £92m as 19,000 appointments cancelled. https://www.telegraph.co.uk/technology/2018/10/11/wannacry-cyber-attack-cost-nhs-92m-19000-appointments-cancelled/. Accessed 9 May 2018.

FitchRatings. 2021. U.S. Cyber Insurance Market Update (Spike in Claims Leads to Decline in 2020 Underwriting Performance). https://www.fitchratings.com/research/insurance/us-cyber-insurance-market-update-spike-in-claims-leads-to-decline-in-2020-underwriting-performance-26-05-2021.

Fossaceca, J.M., T.A. Mazzuchi, and S. Sarkani. 2015. MARK-ELM: Application of a novel Multiple Kernel Learning framework for improving the robustness of network intrusion detection. *Expert Systems with Applications* 42 (8): 4062–4080. https://doi.org/10.1016/j.eswa.2014.12.040.

Franke, U., and J. Brynielsson. 2014. Cyber situational awareness–a systematic review of the literature. *Computers & security* 46: 18–31.

Freeha, K., K.J. Hwan, M. Lars, and M. Robin. 2021. Data breach management: An integrated risk model. *Information & Management* 58 (1): 103392. https://doi.org/10.1016/j.im.2020.103392.

Ganeshan, R., and P. Rodrigues. 2020. Crow-AFL: Crow based adaptive fractional lion optimization approach for the intrusion detection. *Wireless Personal Communications* 111 (4): 2065–2089. https://doi.org/10.1007/s11277-019-06972-0.

GAO. 2021. CYBER INSURANCE—Insurers and policyholders face challenges in an evolving market. https://www.gao.gov/assets/gao-21-477.pdf. Accessed 16 May 2021.

Garber, J. 2021. Colonial Pipeline fiasco foreshadows impact of Biden energy policy. https://www.foxbusiness.com/markets/colonial-pipeline-fiasco-foreshadows-impact-of-biden-energy-policy. Accessed 4 May 2021.

Gauthama Raman, M.R., N. Somu, S. Jagarapu, T. Manghnani, T. Selvam, K. Krithivasan, and V.S. Shankar Sriram. 2020. An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. *Artificial Intelligence Review* 53 (5): 3255–3286. https://doi.org/10.1007/s10462-019-09762-z.

Gavel, S., A.S. Raghuvanshi, and S. Tiwari. 2021. Distributed intrusion detection scheme using dual-axis dimensionality reduction for Internet of things (IoT). *Journal of Supercomputing*. https://doi.org/10.1007/s11227-021-03697-5.

GDPR.EU. 2021. FAQ. https://gdpr.eu/faq/. Accessed 10 May 2021.

Georgescu, T.M., B. Iancu, and M. Zurini. 2019. Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks. *Sensors (switzerland)*. https://doi.org/10.3390/s19153380.

Giudici, P., and E. Raffinetti. 2020. Cyber risk ordering with rank-based statistical models. *AStA Advances in Statistical Analysis*. https://doi.org/10.1007/s10182-020-00387-0.

Goh, J., S. Adepu, K.N. Junejo, and A. Mathur. 2016. A dataset to support research in the design of secure water treatment systems. In CRITIS.

Gong, X.Y., J.L. Lu, Y.F. Zhou, H. Qiu, and R. He. 2021. Model uncertainty based annotation error fixing for web attack detection. *Journal of Signal Processing Systems for Signal Image and Video Technology* 93 (2–3): 187–199. https://doi.org/10.1007/s11265-019-01494-1.

Goode, S., H. Hoehle, V. Venkatesh, and S.A. Brown. 2017. USER compensation as a data breach recovery action: An investigation of the sony playstation network breach. *MIS Quarterly* 41 (3): 703–727.

Guo, H., S. Huang, C. Huang, Z. Pan, M. Zhang, and F. Shi. 2020. File entropy signal analysis combined with wavelet decomposition for malware classification. *IEEE Access* 8: 158961–158971. https://doi.org/10.1109/ACCESS.2020.3020330.

Habib, M., I. Aljarah, and H. Faris. 2020. A Modified multi-objective particle swarm optimizer-based Lévy flight: An approach toward intrusion detection in Internet of Things. *Arabian Journal for Science and Engineering* 45 (8): 6081–6108. https://doi.org/10.1007/s13369-020-04476-9.

Hajj, S., R. El Sibai, J.B. Abdo, J. Demerjian, A. Makhoul, and C. Guyeux. 2021. Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets. *Transactions on Emerging Telecommunications Technologies* 32 (4): 36. https://doi.org/10.1002/ett.4240.

Heartfield, R., G. Loukas, A. Bezemskij, and E. Panaousis. 2021. Self-configurable cyber-physical intrusion detection for smart homes using reinforcement learning. *IEEE Transactions on Information Forensics and Security* 16: 1720–1735. https://doi.org/10.1109/tifs.2020.3042049.

Hemo, B., T. Gafni, K. Cohen, and Q. Zhao. 2020. Searching for anomalies over composite hypotheses. *IEEE Transactions on Signal Processing* 68: 1181–1196. https://doi.org/10.1109/TSP.2020.2971438

Hindy, H., D. Brosset, E. Bayne, A.K. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens. 2020. A taxonomy of network threats and the effect of current datasets on intrusion detection systems. *IEEE Access* 8: 104650–104675. https://doi.org/10.1109/ACCESS.2020.3000179.

Hong, W., D. Huang, C. Chen, and J. Lee. 2020. Towards accurate and efficient classification of power system contingencies and cyber-attacks using recurrent neural networks. *IEEE Access* 8: 123297–123309. https://doi.org/10.1109/ACCESS.2020.3007609.

Husák, M., M. Zádník, V. Bartos, and P. Sokol. 2020. Dataset of intrusion detection alerts from a sharing platform. *Data in Brief* 33: 106530.

IBM Security. 2020. Cost of a Data breach Report. https://www.capita.com/sites/g/files/nginej291/files/2020-08/Ponemon-Global-Cost-of-Data-Breach-Study-2020.pdf. Accessed 19 May 2021.

IEEE. 2021. IEEE Quick Facts. https://www.ieee.org/about/at-a-glance.html. Accessed 11 May 2021.

Kilincer, I.F., F. Ertam, and S. Abdulkadir. 2021. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks* 188: 107840. https://doi.org/10.1016/j.comnet.2021.107840.

Jaber, A.N., and S. Ul Rehman. 2020. FCM-SVM based intrusion detection system for cloud computing environment. *Cluster Computing—the Journal of Networks Software Tools and Applications* 23 (4): 3221–3231. https://doi.org/10.1007/s10586-020-03082-6.

Jacobs, J., S. Romanosky, B. Edwards, M. Roytman, and I. Adjerid. 2019. Exploit prediction scoring system (epss). arXiv:1908.04856

Jacobsen, A. et al. 2020. FAIR principles: Interpretations and implementation considerations. *Data Intelligence* 2 (1–2): 10–29. https://doi.org/10.1162/dint_r_00024.

Jahromi, A.N., S. Hashemi, A. Dehghantanha, R.M. Parizi, and K.K.R. Choo. 2020. An enhanced stacked LSTM method with no random initialization for malware threat hunting in safety and time-critical systems. *IEEE Transactions on Emerging Topics in Computational Intelligence* 4 (5): 630–640. https://doi.org/10.1109/TETCI.2019.2910243.

Jang, S., S. Li, and Y. Sung. 2020. FastText-based local feature visualization algorithm for merged image-based malware classification framework for cyber security and cyber defense. *Mathematics* 8 (3): 13. https://doi.org/10.3390/math8030460.

Javeed, D., T.H. Gao, and M.T. Khan. 2021. SDN-enabled hybrid DL-driven framework for the detection of emerging cyber threats in IoT. *Electronics* 10 (8): 16. https://doi.org/10.3390/electronics10080918.

Johnson, P., D. Gorton, R. Lagerstrom, and M. Ekstedt. 2016. Time between vulnerability disclosures: A measure of software product vulnerability. *Computers & Security* 62: 278–295. https://doi.org/10.1016/j.cose.2016.08.004.

Johnson, P., R. Lagerström, M. Ekstedt, and U. Franke. 2018. Can the common vulnerability scoring system be trusted? A Bayesian analysis. *IEEE Transactions on Dependable and Secure Computing* 15 (6): 1002–1015. https://doi.org/10.1109/TDSC.2016.2644614.

Junger, M., V. Wang, and M. Schlömer. 2020. Fraud against businesses both online and offline: Crime scripts, business characteristics, efforts, and benefits. *Crime Science* 9 (1): 13. https://doi.org/10.1186/s40163-020-00119-4.

Kalutarage, H.K., H.N. Nguyen, and S.A. Shaikh. 2017. Towards a threat assessment framework for apps collusion. *Telecommunication Systems* 66 (3): 417–430. https://doi.org/10.1007/s11235-017-0296-1.

Kamarudin, M.H., C. Maple, T. Watson, and N.S. Safa. 2017. A LogitBoost-based algorithm for detecting known and unknown web attacks. *IEEE Access* 5: 26190–26200. https://doi.org/10.1109/ACCESS.2017.2766844.

Kasongo, S.M., and Y.X. Sun. 2020. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. *Computers & Security* 92: 15. https://doi.org/10.1016/j.cose.2020.101752.

Keserwani, P.K., M.C. Govil, E.S. Pilli, and P. Govil. 2021. A smart anomaly-based intrusion detection system for the Internet of Things (IoT) network using GWO–PSO–RF model. *Journal of Reliable Intelligent Environments* 7 (1): 3–21. https://doi.org/10.1007/s40860-020-00126-x.

Keshk, M., E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil. 2021. An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems. *IEEE Transactions on Sustainable Computing* 6 (1): 66–79. https://doi.org/10.1109/TSUSC.2019.2906657.

Khan, I.A., D.C. Pi, A.K. Bhatia, N. Khan, W. Haider, and A. Wahab. 2020. Generating realistic IoT-based IDS dataset centred on fuzzy qualitative modelling for cyber-physical systems. *Electronics Letters* 56 (9): 441–443. https://doi.org/10.1049/el.2019.4158.

Khraisat, A., I. Gondal, P. Vamplew, J. Kamruzzaman, and A. Alazab. 2020. Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vector machine. *Electronics* 9 (1): 18. https://doi.org/10.3390/electronics9010173.

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

Kilincer, I.F., F. Ertam, and A. Sengur. 2021. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks* 188: 16. https://doi.org/10.1016/j.comnet.2021.107840.

Kim, D., and H.K. Kim. 2019. Automated dataset generation system for collaborative research of cyber threat analysis. *Security and Communication Networks* 2019: 10. https://doi.org/10.1155/2019/6268476.

Kim, G., C. Lee, J. Jo, and H. Lim. 2020. Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. *International Journal of Machine Learning and Cybernetics* 11 (10): 2341–2355. https://doi.org/10.1007/s13042-020-01122-6.

Kirubavathi, G., and R. Anitha. 2016. Botnet detection via mining of traffic flow characteristics. *Computers & Electrical Engineering* 50: 91–101. https://doi.org/10.1016/j.compeleceng.2016.01.012.

Kiwia, D., A. Dehghantanha, K.K.R. Choo, and J. Slaughter. 2018. A cyber kill chain based taxonomy of banking Trojans for evolutionary computational intelligence. *Journal of Computational Science* 27: 394–409. https://doi.org/10.1016/j.jocs.2017.10.020.

Koroniotis, N., N. Moustafa, and E. Sitnikova. 2020. A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Generation Computer Systems* 110: 91–106. https://doi.org/10.1016/j.future.2020.03.042.

Kruse, C.S., B. Frederick, T. Jacobson, and D. Kyle Monticone. 2017. Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care* 25 (1): 1–10.

Kshetri, N. 2018. The economics of cyber-insurance. *IT Professional* 20 (6): 9–14. https://doi.org/10.1109/MITP.2018.2874210.

Kumar, R., P. Kumar, R. Tripathi, G.P. Gupta, T.R. Gadekallu, and G. Srivastava. 2021. SP2F: A secured privacy-preserving framework for smart agricultural Unmanned Aerial Vehicles. *Computer Networks*. https://doi.org/10.1016/j.comnet.2021.107819.

Kumar, R., and R. Tripathi. 2021. DBTP2SF: A deep blockchain-based trustworthy privacy-preserving secured framework in industrial internet of things systems. *Transactions on Emerging Telecommunications Technologies* 32 (4): 27. https://doi.org/10.1002/ett.4222.

Laso, P.M., D. Brosset, and J. Puentes. 2017. Dataset of anomalies and malicious acts in a cyber-physical subsystem. *Data in Brief* 14: 186–191. https://doi.org/10.1016/j.dib.2017.07.038.

Lee, J., J. Kim, I. Kim, and K. Han. 2019. Cyber threat detection based on artificial neural networks using event profiles. *IEEE Access* 7: 165607–165626. https://doi.org/10.1109/ACCESS.2019.2953095.

Lee, S.J., P.D. Yoo, A.T. Asyhari, Y. Jhi, L. Chermak, C.Y. Yeun, and K. Taha. 2020. IMPACT: Impersonation attack detection via edge computing using deep Autoencoder and feature abstraction. *IEEE Access* 8: 65520–65529. https://doi.org/10.1109/ACCESS.2020.2985089.

Leong, Y.-Y., and Y.-C. Chen. 2020. Cyber risk cost and management in IoT devices-linked health insurance. *The Geneva Papers on Risk and Insurance—Issues and Practice* 45 (4): 737–759. https://doi.org/10.1057/s41288-020-00169-4.

Levi, M. 2017. Assessing the trends, scale and nature of economic cybercrimes: overview and Issues: In Cybercrimes, cybercriminals and their policing, in crime, law and social change. *Crime, Law and Social Change* 67 (1): 3–20. https://doi.org/10.1007/s10611-016-9645-3.

Li, C., K. Mills, D. Niu, R. Zhu, H. Zhang, and H. Kinawi. 2019a. Android malware detection based on factorization machine. *IEEE Access* 7: 184008–184019. https://doi.org/10.1109/ACCESS.2019.2958927.

Li, D.Q., and Q.M. Li. 2020. Adversarial deep ensemble: evasion attacks and defenses for malware detection. *IEEE Transactions on Information Forensics and Security* 15: 3886–3900. https://doi.org/10.1109/tifs.2020.3003571.

Li, D.Q., Q.M. Li, Y.F. Ye, and S.H. Xu. 2021. A framework for enhancing deep neural networks against adversarial malware. *IEEE Transactions on Network Science and Engineering* 8 (1): 736–750. https://doi.org/10.1109/tnse.2021.3051354.

Li, R.H., C. Zhang, C. Feng, X. Zhang, and C.J. Tang. 2019b. Locating vulnerability in binaries using deep neural networks. *IEEE Access* 7: 134660–134676. https://doi.org/10.1109/access.2019.2942043.

Li, X., M. Xu, P. Vijayakumar, N. Kumar, and X. Liu. 2020. Detection of low-frequency and multi-stage attacks in industrial Internet of Things. *IEEE Transactions on Vehicular Technology* 69 (8): 8820–8831. https://doi.org/10.1109/TVT.2020.2995133.

Liu, H.Y., and B. Lang. 2019. Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences—Basel* 9 (20): 28. https://doi.org/10.3390/app9204396.

Lopez-Martin, M., B. Carro, and A. Sanchez-Esguevillas. 2020. Application of deep reinforcement learning to intrusion detection for supervised problems. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2019.112963.

Loukas, G., D. Gan, and Tuan Vuong. 2013. A review of cyber threats and defence approaches in emergency management. *Future Internet* 5: 205–236.

Luo, C.C., S. Su, Y.B. Sun, Q.J. Tan, M. Han, and Z.H. Tian. 2020. A convolution-based system for malicious URLs detection. *CMC—Computers Materials Continua* 62 (1): 399–411.

Mahbooba, B., M. Timilsina, R. Sahal, and M. Serrano. 2021. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity* 2021: 11. https://doi.org/10.1155/2021/6634811.

Mahdavifar, S., and A.A. Ghorbani. 2020. DeNNeS: Deep embedded neural network expert system for detecting cyber attacks. *Neural Computing & Applications* 32 (18): 14753–14780. https://doi.org/10.1007/s00521-020-04830-w.

Mahfouz, A., A. Abuhussein, D. Venugopal, and S. Shiva. 2020. Ensemble classifiers for network intrusion detection using a novel network attack dataset. *Future Internet* 12 (11): 1–19. https://doi.org/10.3390/fi12110180.

Maleks Smith, Z., E. Lostri, and J.A. Lewis. 2020. The hidden costs of cybercrime. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf. Accessed 16 May 2021.

Malik, J., A. Akhunzada, I. Bibi, M. Imran, A. Musaddiq, and S.W. Kim. 2020. Hybrid deep learning: An efficient reconnaissance and surveillance detection mechanism in SDN. *IEEE Access* 8: 134695–134706. https://doi.org/10.1109/ACCESS.2020.3009849.

Manimurugan, S. 2020. IoT-Fog-Cloud model for anomaly detection using improved Naive Bayes and principal component analysis. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-02723-3.

Martin, A., R. Lara-Cabrera, and D. Camacho. 2019. Android malware detection through hybrid features fusion and ensemble classifiers: The AndroPyTool framework and the OmniDroid dataset. *Information Fusion* 52: 128–142. https://doi.org/10.1016/j.inffus.2018.12.006.

Mauro, M.D., G. Galatro, and A. Liotta. 2020. Experimental review of neural-based approaches for network intrusion management. *IEEE Transactions on Network and Service Management* 17 (4): 2480–2495. https://doi.org/10.1109/TNSM.2020.3024225.

McLeod, A., and D. Dolezel. 2018. Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decision Support Systems* 108: 57–68. https://doi.org/10.1016/j.dss.2018.02.007.

Meira, J., R. Andrade, I. Praca, J. Carneiro, V. Bolon-Canedo, A. Alonso-Betanzos, and G. Marreiros. 2020. Performance evaluation of unsupervised techniques in cyber-attack anomaly detection. *Journal of Ambient Intelligence and Humanized Computing* 11 (11): 4477–4489. https://doi.org/10.1007/s12652-019-01417-9.

Miao, Y., J. Ma, X. Liu, J. Weng, H. Li, and H. Li. 2019. Lightweight fine-grained search over encrypted data in Fog computing. *IEEE Transactions on Services Computing* 12 (5): 772–785. https://doi.org/10.1109/TSC.2018.2823309.

Miller, C., and C. Valasek. 2015. Remote exploitation of an unaltered passenger vehicle. *Black Hat USA* 2015 (S 91).

Mireles, J.D., E. Ficke, J.H. Cho, P. Hurley, and S.H. Xu. 2019. Metrics towards measuring cyber agility. *IEEE Transactions on Information Forensics and Security* 14 (12): 3217–3232. https://doi.org/10.1109/tifs.2019.2912551.

Mishra, N., and S. Pandya. 2021. Internet of Things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3073408.

Monshizadeh, M., V. Khatri, B.G. Atli, R. Kantola, and Z. Yan. 2019. Performance evaluation of a combined anomaly detection platform. *IEEE Access* 7: 100964–100978. https://doi.org/10.1109/ACCESS.2019.2930832.

Moreno, V.C., G. Reniers, E. Salzano, and V. Cozzani. 2018. Analysis of physical and cyber security-related events in the chemical and process industry. *Process Safety and Environmental Protection* 116: 621–631. https://doi.org/10.1016/j.psep.2018.03.026.

Moro, E.D. 2020. Towards an economic cyber loss index for parametric cover based on IT security indicator: A preliminary analysis. *Risks*. https://doi.org/10.3390/risks8020045.

Moustafa, N., E. Adi, B. Turnbull, and J. Hu. 2018. A new threat intelligence scheme for safeguarding industry 4.0 systems. *IEEE Access* 6: 32910–32924. https://doi.org/10.1109/ACCESS.2018.2844794.

Moustakidis, S., and P. Karlsson. 2020. A novel feature extraction methodology using Siamese convolutional neural networks for intrusion detection. *Cybersecurity*. https://doi.org/10.1186/s42400-020-00056-4.

Mukhopadhyay, A., S. Chatterjee, K.K. Bagchi, P.J. Kirs, and G.K. Shukla. 2019. Cyber Risk Assessment and Mitigation (CRAM) framework using Logit and Probit models for cyber insurance. *Information Systems Frontiers* 21 (5): 997–1018. https://doi.org/10.1007/s10796-017-9808-5.

Murphey, H. 2021a. Biden signs executive order to strengthen US cyber security. https://www.ft.com/content/4d808359-b504-4014-85f6-68e7a2851bf1?accessToken=zwAAAXl0_ifgkc9NgINZtQRAFNOF9mjnooUb8Q.MEYCIQDw46SFWsMn1iyuz3kvgAmn6mxc0rIVfw10Lg1ovJSfJwIhAK2X2URzfSqHwIS7ddRCvSt2nGC2DcdoiDTG49-4TeEt&sharetype=gift?token=fbcd6323-1ecf-4fc3-b136-b5b0dd6a8756. Accessed 7 May 2021.

Murphey, H. 2021b. Millions of connected devices have security flaws, study shows. https://www.ft.com/content/0bf92003-926d-4dee-87d7-b01f7c3e9621?accessToken=zwAAAXnA7f2Ikc8L-SADkm1N7tOH17AffD6WIQ.MEQCIDjBuROvhmYV0Mx3iB0cEV7m5oND1uaCICxJu0mzxM0PAiBam98q9zfHiTB6hKGr1gGl0Azt85yazdpX9K5sI8se3Q&sharetype=gift?token=2538218d-77d9-4dd3-9649-3cb556a34e51. Accessed 6 May 2021.

Murugesan, V., M. Shalinie, and M.H. Yang. 2018. Design and analysis of hybrid single packet IP traceback scheme. *IET Networks* 7 (3): 141–151. https://doi.org/10.1049/iet-net.2017.0115.

Mwitondi, K.S., and S.A. Zargari. 2018. An iterative multiple sampling method for intrusion detection. *Information Security Journal* 27 (4): 230–239. https://doi.org/10.1080/19393555.2018.1539790.

Neto, N.N., S. Madnick, A.M.G. De Paula, and N.M. Borges. 2021. Developing a global data breach database and the challenges encountered. *ACM Journal of Data and Information Quality* 13 (1): 33. https://doi.org/10.1145/3439873.

Nurse, J.R.C., L. Axon, A. Erola, I. Agrafiotis, M. Goldsmith, and S. Creese. 2020. The data that drives cyber insurance: A study into the underwriting and claims processes. In 2020 International conference on cyber situational awareness, data analytics and assessment (CyberSA), 15–19 June 2020.

Oliveira, N., I. Praca, E. Maia, and O. Sousa. 2021. Intelligent cyber attack detection and classification for network-based intrusion detection systems. *Applied Sciences—Basel* 11 (4): 21. https://doi.org/10.3390/app11041674.

Page, M.J. et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews* 10 (1): 89. https://doi.org/10.1186/s13643-021-01626-4.

Pajouh, H.H., R. Javidan, R. Khayami, A. Dehghantanha, and K.R. Choo. 2019. A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. *IEEE Transactions on Emerging Topics in Computing* 7 (2): 314–323. https://doi.org/10.1109/TETC.2016.2633228.

Parra, G.D., P. Rad, K.K.R. Choo, and N. Beebe. 2020. Detecting Internet of Things attacks using distributed deep learning. *Journal of Network and Computer Applications* 163: 13. https://doi.org/10.1016/j.jnca.2020.102662.

Paté-Cornell, M.E., M. Kuypers, M. Smith, and P. Keller. 2018. Cyber risk management for critical infrastructure: A risk analysis model and three case studies. *Risk Analysis* 38 (2): 226–241. https://doi.org/10.1111/risa.12844.

Pooser, D.M., M.J. Browne, and O. Arkhangelska. 2018. Growth in the perception of cyber risk: evidence from U.S. P&C Insurers. *The Geneva Papers on Risk and Insurance—Issues and Practice* 43 (2): 208–223. https://doi.org/10.1057/s41288-017-0077-9.

Pu, G., L. Wang, J. Shen, and F. Dong. 2021. A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology* 26 (2): 146–153. https://doi.org/10.26599/TST.2019.9010051.

Qiu, J., W. Luo, L. Pan, Y. Tai, J. Zhang, and Y. Xiang. 2019. Predicting the impact of android malicious samples via machine learning. *IEEE Access* 7: 66304–66316. https://doi.org/10.1109/ACCESS.2019.2914311.

Qu, X., L. Yang, K. Guo, M. Sun, L. Ma, T. Feng, S. Ren, K. Li, and X. Ma. 2020. Direct batch growth hierarchical self-organizing mapping based on statistics for efficient network intrusion detection. *IEEE Access* 8: 42251–42260. https://doi.org/10.1109/ACCESS.2020.2976810.

Rahman, Md.S., S. Halder, Md. Ashraf Uddin, and U.K. Acharjee. 2021. An efficient hybrid system for anomaly detection in social networks. *Cybersecurity* 4 (1): 10. https://doi.org/10.1186/s42400-021-00074-w.

Ramaiah, M., V. Chandrasekaran, V. Ravi, and N. Kumar. 2021. An intrusion detection system using optimized deep neural network architecture. *Transactions on Emerging Telecommunications Technologies* 32 (4): 17. https://doi.org/10.1002/ett.4221.

Raman, M.R.G., K. Kannan, S.K. Pal, and V.S.S. Sriram. 2016. Rough set-hypergraph-based feature selection approach for intrusion detection systems. *Defence Science Journal* 66 (6): 612–617. https://doi.org/10.14429/dsj.66.10802.

Rathore, S., J.H. Park. 2018. Semi-supervised learning based distributed attack detection framework for IoT. *Applied Soft Computing* 72: 79–89. https://doi.org/10.1016/j.asoc.2018.05.049.

Romanosky, S., L. Ablon, A. Kuehn, and T. Jones. 2019. Content analysis of cyber insurance policies: How do carriers price cyber risk? *Journal of Cybersecurity (oxford)* 5 (1): tyz002.

Sarabi, A., P. Naghizadeh, Y. Liu, and M. Liu. 2016. Risky business: Fine-grained data breach prediction using business profiles. *Journal of Cybersecurity* 2 (1): 15–28. https://doi.org/10.1093/cybsec/tyw004.

Sardi, Alberto, Alessandro Rizzi, Enrico Sorano, and Anna Guerrieri. 2021. Cyber risk in health facilities: A systematic literature review. *Sustainability* 12 (17): 7002.

Sarker, Iqbal H., A.S.M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. 2020. Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data* 7 (1): 41. https://doi.org/10.1186/s40537-020-00318-5.

Scopus. 2021. Factsheet. https://www.elsevier.com/__data/assets/pdf_file/0017/114533/Scopus_GlobalResearch_Factsheet2019_FINAL_WEB.pdf. Accessed 11 May 2021.

Sentuna, A., A. Alsadoon, P.W.C. Prasad, M. Saadeh, and O.H. Alsadoon. 2021. A novel Enhanced Naïve Bayes Posterior Probability (ENBPP) using machine learning: Cyber threat analysis. *Neural Processing Letters* 53 (1): 177–209. https://doi.org/10.1007/s11063-020-10381-x.

Shaukat, K., S.H. Luo, V. Varadharajan, I.A. Hameed, S. Chen, D.X. Liu, and J.M. Li. 2020. Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies* 13 (10): 27. https://doi.org/10.3390/en13102509.

Sheehan, B., F. Murphy, M. Mullins, and C. Ryan. 2019. Connected and autonomous vehicles: A cyber-risk classification framework. *Transportation Research Part a: Policy and Practice* 124: 523–536. https://doi.org/10.1016/j.tra.2018.06.033.

Sheehan, B., F. Murphy, A.N. Kia, and R. Kiely. 2021. A quantitative bow-tie cyber risk classification and assessment framework. *Journal of Risk Research* 24 (12): 1619–1638.

Shlomo, A., M. Kalech, and R. Moskovitch. 2021. Temporal pattern-based malicious activity detection in SCADA systems. *Computers & Security* 102: 17. https://doi.org/10.1016/j.cose.2020.102153.

Singh, K.J., and T. De. 2020. Efficient classification of DDoS attacks using an ensemble feature selection algorithm. *Journal of Intelligent Systems* 29 (1): 71–83. https://doi.org/10.1515/jisys-2017-0472.

Skrjanc, I., S. Ozawa, T. Ban, and D. Dovzan. 2018. Large-scale cyber attacks monitoring using Evolving Cauchy Possibilistic Clustering. *Applied Soft Computing* 62: 592–601. https://doi.org/10.1016/j.asoc.2017.11.008.

Smart, W. 2018. Lessons learned review of the WannaCry Ransomware Cyber Attack. https://www.england.nhs.uk/wp-content/uploads/2018/02/lessons-learned-review-wannacry-ransomware-cyber-attack-cio-review.pdf. Accessed 7 May 2021.

Sornette, D., T. Maillart, and W. Kröger. 2013. Exploring the limits of safety analysis in complex technological systems. *International Journal of Disaster Risk Reduction* 6: 59–66. https://doi.org/10.1016/j.ijdrr.2013.04.002.

Sovacool, B.K. 2008. The costs of failure: A preliminary assessment of major energy accidents, 1907–2007. *Energy Policy* 36 (5): 1802–1820. https://doi.org/10.1016/j.enpol.2008.01.040.

SpringerLink. 2021. Journal Search. https://rd.springer.com/search?facet-content-type=%22Journal%22. Accessed 11 May 2021.

Stojanovic, B., K. Hofer-Schmitz, and U. Kleb. 2020. APT datasets and attack modeling for automated detection methods: A review. *Computers & Security* 92: 19. https://doi.org/10.1016/j.cose.2020.101734.

Subroto, A., and A. Apriyana. 2019. Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0216-1.

Tan, Z., A. Jamdagni, X. He, P. Nanda, R.P. Liu, and J. Hu. 2015. Detection of denial-of-service attacks based on computer vision techniques. *IEEE Transactions on Computers* 64 (9): 2519–2533. https://doi.org/10.1109/TC.2014.2375218.

Tidy, J. 2021. Irish cyber-attack: Hackers bail out Irish health service for free. https://www.bbc.com/news/world-europe-57197688. Accessed 6 May 2021.

Tuncer, T., F. Ertam, and S. Dogan. 2020. Automated malware recognition method based on local neighborhood binary pattern. *Multimedia Tools and Applications* 79 (37–38): 27815–27832. https://doi.org/10.1007/s11042-020-09376-6.

Uhm, Y., and W. Pak. 2021. Service-aware two-level partitioning for machine learning-based network intrusion detection with high performance and high scalability. *IEEE Access* 9: 6608–6622. https://doi.org/10.1109/ACCESS.2020.3048900.

Ulven, J.B., and G. Wangen. 2021. A systematic review of cybersecurity risks in higher education. *Future Internet* 13 (2): 1–40. https://doi.org/10.3390/fi13020039.

Vaccari, I., G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso. 2020. MQTTset, a new dataset for machine learning techniques on MQTT. *Sensors* 20 (22): 17. https://doi.org/10.3390/s20226578.

Valeriano, B., and R.C. Maness. 2014. The dynamics of cyber conflict between rival antagonists, 2001–11. *Journal of Peace Research* 51 (3): 347–360. https://doi.org/10.1177/0022343313518940.

Varghese, J.E., and B. Muniyal. 2021. An Efficient IDS framework for DDoS attacks in SDN environment. *IEEE Access* 9: 69680–69699. https://doi.org/10.1109/ACCESS.2021.3078065.

Varsha, M. V., P. Vinod, K.A. Dhanya. 2017 Identification of malicious android app using manifest and opcode features. *Journal of Computer Virology and Hacking Techniques* 13 (2): 125–138. https://doi.org/10.1007/s11416-016-0277-z

Velliangiri, S., and H.M. Pandey. 2020. Fuzzy-Taylor-elephant herd optimization inspired Deep Belief Network for DDoS attack detection and comparison with state-of-the-arts algorithms. *Future Generation Computer Systems—the International Journal of Escience* 110: 80–90. https://doi.org/10.1016/j.future.2020.03.049.

Verma, A., and V. Ranga. 2020. Machine learning based intrusion detection systems for IoT applications. *Wireless Personal Communications* 111 (4): 2287–2310. https://doi.org/10.1007/s11277-019-06986-8.

Vidros, S., C. Kolias, G. Kambourakis, and L. Akoglu. 2017. Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet* 9 (1): 19. https://doi.org/10.3390/fi9010006.

Vinayakumar, R., M. Alazab, K.P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman. 2019. Deep learning approach for intelligent intrusion detection system. *IEEE Access* 7: 41525–41550. https://doi.org/10.1109/access.2019.2895334.

Walker-Roberts, S., M. Hammoudeh, O. Aldabbas, M. Aydin, and A. Dehghantanha. 2020. Threats on the horizon: Understanding security threats in the era of cyber-physical systems. *Journal of Supercomputing* 76 (4): 2643–2664. https://doi.org/10.1007/s11227-019-03028-9.

Web of Science. 2021. Web of Science: Science Citation Index Expanded. https://clarivate.com/webofsciencegroup/solutions/webofscience-scie/. Accessed 11 May 2021.

World Economic Forum. 2020. WEF Global Risk Report. http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf. Accessed 13 May 2020.

Xin, Y., L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang. 2018. Machine learning and deep learning methods for cybersecurity. *IEEE Access* 6: 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950.

Xu, C., J. Zhang, K. Chang, and C. Long. 2013. Uncovering collusive spammers in Chinese review websites. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management.

Yang, J., T. Li, G. Liang, W. He, and Y. Zhao. 2019. A Simple recurrent unit model based intrusion detection system with DCGAN. *IEEE Access* 7: 83286–83296. https://doi.org/10.1109/ACCESS.2019.2922692.

Yuan, B.G., J.F. Wang, D. Liu, W. Guo, P. Wu, and X.H. Bao. 2020. Byte-level malware classification based on Markov images and deep learning. *Computers & Security* 92: 12. https://doi.org/10.1016/j.cose.2020.101740.

Zhang, S., X.M. Ou, and D. Caragea. 2015. Predicting cyber risks through national vulnerability database. *Information Security Journal* 24 (4–6): 194–206. https://doi.org/10.1080/19393555.2015.1111961.

Zhang, Y., P. Li, and X. Wang. 2019. Intrusion detection for IoT based on improved genetic algorithm and deep belief network. *IEEE Access* 7: 31711–31722.

Zheng, Muwei, Hannah Robbins, Zimo Chai, Prakash Thapa, and Tyler Moore. 2018. Cybersecurity research datasets: taxonomy and empirical analysis. In 11th {USENIX} workshop on cyber security experimentation and test ({CSET} 18).

Zhou, X., W. Liang, S. Shimizu, J. Ma, and Q. Jin. 2021. Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics* 17 (8): 5790–5798. https://doi.org/10.1109/TII.2020.3047675.

Zhou, Y.Y., G. Cheng, S.Q. Jiang, and M. Dai. 2020. Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer Networks* 174: 17. https://doi.org/10.1016/j.comnet.2020.107247.

About the authors

Frank Cremer is a PhD student at the Kemmy Business School, University of Limerick, as part of the Emerging Risk Group (ERG). He is researching in joint cooperation with the Institute for Insurance Studies (ivwKöln), TH Köln, where he is working as a Research Assistant at the Cologne Research Centre for Reinsurance. His current research interests include cyber risks, cyber insurance and cybersecurity. Frank is a Fellow of the Chartered Insurance Institute (FCII) and a member of the German Association for Insurance Studies (DVfVW).

Barry Sheehan is a Lecturer in Risk and Finance at the Kemmy Business School at the University of Limerick. In his research, Dr Sheehan investigates novel risk metrication and machine learning methodologies in the context of insurance and finance, attentive to a changing private and public emerging risk environment. He is a researcher with significant insurance industry and academic experience. With a professional background in actuarial science, his research uses machine-learning techniques to estimate the changing risk profile produced by emerging technologies. He is a senior member of the Emerging Risk Group (ERG) at the University of Limerick, which has long-established expertise in insurance and risk management and has continued success within large research consortia including a number of SFI, FP7 and EU H2020 research projects. In particular, he contributed to the successful completion of three Horizon 2020 EU-funded projects, including PROTECT, Vision Inspired Driver Assistance Systems (VIDAS) and Cloud Large Scale Video Analysis (Cloud-LSVA).

Michael Fortmann is a Professor at the Institute of Insurance at the Technical University of Cologne. His activities include teaching and research in insurance law and liability insurance. His research focuses include D&O, corporate liability, fidelity and cyber insurance. In addition, he heads the Master's degree programme in insurance law and is the Academic Director of the Automotive Insurance Manager and Cyber Insurance Manager certificate programmes. He is also chairman of the examination board at the Institute of Insurance Studies.

Arash Negahdari Kia is a postdoctoral Marie Cuire scholar and Research Fellow at the Kemmy Business School (KBS), University of Limerick (UL), a member of the Lero Software Research Center and Emerging Risk Group (ERG). He researches the cybersecurity risks of autonomous vehicles using machine-learning algorithms in a team supervised by Dr Finbarr Murphy at KBS, UL. For his PhD, he developed two graph-based, semi-supervised algorithms for multivariate time series for global stock market indices prediction. For his Master's, he developed neural network models for Forex market prediction. Arash's other research interests include text mining, graph mining and bioinformatics.

Martin Mullins is a Professor in Risk and Insurance at the Kemmy Business School, University of Limerick. He worked on a number of insurance-related research projects, including four EU Commission-funded projects around emerging technologies and risk transfer. Prof. Mullins maintains strong links with the international insurance industry and works closely with Lloyd's of London and XL Catlin on emerging risk. His work also encompasses the area of applied ethics as it pertains to new technologies. In the field of applied ethics, Dr Mullins works closely with the insurance industry and lectures on cultural and technological breakthroughs of high societal relevance. In that respect, Dr Martin Mullins has been appointed to a European expert group to advise EIOPA on the development of digital responsibility principles in insurance.

Finbarr Murphy is Executive Dean Kemmy Business School. A computer engineering graduate, Finbarr worked for over 10 years in investment banking before returning to academia and completing his PhD in 2010. Finbarr has authored or co-authored over 70 refereed journal papers, edited books and book chapters. His research has been published in leading research journals in his discipline, such as Nature Nanotechnology, Small, Transportation Research A-F and the Review of Derivatives Research. A former Fulbright Scholar and Erasmus Mundus Exchange Scholar, Finbarr has delivered numerous guest lectures in America, mainland Europe, Israel, Russia, China and Vietnam. His research interests include quantitative finance and, more recently, emerging technological risk. Finbarr is currently engaged in several EU H2020 projects and with the Irish Science Foundation Ireland.

Stefan Materne (FCII) has held the Chair of Reinsurance at the Institute of Insurance of TH Köln since 1998, focusing on the efficiency of reinsurance, industrial insurance and alternative risk transfer (ART). He studied mathematics and computer science with a focus on artificial intelligence and researched from 1988 to 1991 at the Fraunhofer Institute for Autonomous Intelligent Systems (AiS) in Schloß Birlinghoven. From 1991 to 2004, Prof. Materne worked for Gen Re (formerly Cologne Re) in various management positions in Germany and abroad, and from 2001 to 2003, he served as General Manager of Cologne Re of Dublin in Ireland. In 2008, Prof. Materne founded the Cologne Reinsurance Research Centre, of which he is the Director. Current issues in reinsurance and related fields are analysed and discussed with practitioners, with valuable contacts through the 'Förderkreis Rückversicherung' and the organisation of the annual Cologne Reinsurance Symposium. Prof. Materne holds various international supervisory boards, board of directors and advisory board mandates at insurance and reinsurance companies, captives, InsurTechs, EIOPA, as well as at insurance-scientific institutions. He also acts as an arbitrator and party representative in arbitration proceedings.