Analyzing Global Cybersecurity Breach Data: A Secondary Data Study of Trends and Patterns

Rabia Nazir
Department of Computer Science
The University of Lahore
Lahore, Punjab, Pakistan

ABSTRACT
The cybersecurity situation in the world is becoming more dangerous, and the number of data breaches grows and grows. The paper is a secondary data analysis research of publicly available breach archives, such as Privacy Rights Clearinghouse, HaveIBeenPwned, and Kaggle datasets. The study was performed using a quantitative, descriptive research design with the help of the R programming language and tidyverse packages to perform data wrangling, statistical analysis, and visualization. Results indicate that breaches reported have doubled during this time, with 2023 showing the highest number of 59 million breached records due to a mega-breach in the technology sector. The most targeted sectors included healthcare and government sectors, with 22.2% of the incidents each, as this data is valuable. Hacking was the most common attack (44.4%), with ransomware and phishing (22.2% each). This study highlights the trends and patterns of global cybersecurity activities between 2021 and 2024. In this paper, the Healthcare and Government Services sectors are overrepresented. It is typical of the high value of the data that they comprise the sensitivity of their operations to interference. The results prove the ongoing incomparability of phishing as an initial access technique related to attack vectors. Moreover, it founded the drastic and recent escalation of ransomware as a major danger.

Keywords
Cybersecurity, cyber-attacks, ransomware, phishing, hacking, insider threats, Information security

1. INTRODUCTION
At a time when data has become one of the main assets of organizations and a constituent element of personal identity, unauthorized exposure or theft has become a universal phenomenon in the world [1]. Cybercrime is now a very structured and lucrative business and its expenses might be USD 10.5 trillion annually by 2025 and this amount is projected to go up to almost USD 24 trillion by 2027 [2]. It is also worrying how often these evil practices occur. In 2021, global cyber-attacks increased by 125 percent per year and the number of data breaches shot up by 200 percent between 2021 and 2024, and indicate that the rate of risk acceleration is not slowing down [3]. In 2024, the average total data breach cost increased to USD 4.88 million, the highest ever recorded between 2021 and 2024 since 2024 by a 10 percent surge [4]. This number includes direct financial losses related to ransomware and emergency IT services and regulatory penalties that have reached a strong increase. Nonetheless, indirect costs are more harmful and long-lasting in many cases. Those will be critical disruption of operation and system failure that will paralyze productivity and cause revenue loss [5].

Motivation for the Study
Despite rapid growth in cybersecurity spending and global awareness, cyber breaches continue to rise revealing a gap between security investment and real protection outcomes. Many studies discuss cybersecurity theoretically, but fewer provide evidence-based insights using large-scale breach datasets. Understanding which sectors are most targeted, which attack vectors dominate, and how breach frequency has evolved can help organizations strengthen prevention strategies. This motivated the present study to analyze global cyber breach records and uncover emerging patterns and trends using secondary breach repositories.

1.1 Research Problem
It is astonishing that, despite the unprecedented expenditure on cybersecurity protection, organizations of small and medium size, regardless of industries, are still unable to stop the unstoppable and complex flow of cyber-attacks. This gap is further compounded by the absence of a concrete, unified, and empirically based knowledge of the patterns of breaches based on large scale, real-world incidence data. Defensive postures can be built historically on the basis of threats or compliance models but do not necessarily depict the present day tactics, techniques and procedures (TTPs) of contemporary foes. The threat actors keep up with the pace of innovation, using emerging technologies such as artificial intelligence (AI) to automate and improve their attacks and exploit systemic vulnerabilities in the supply chains to increase their magnitude. Another resource that would be useful, though not perfect, in carrying out this large-scale analysis would be publicly available data breach repositories, which assemble notifications posted by government agencies.

1.2 Research Questions
1. **RQ1:** How have the frequency, and scale (number of records compromised) of cybersecurity breaches evolved between 2021 and 2024?
2. **RQ2:** Which industries are most frequently targeted by cyber-attacks, and are there discernible patterns in the types of attacks they experience?
3. **RQ3:** What are the most prevalent types of cyber-attacks (e.g., ransomware, phishing, insider threats), and how has their prevalence changed between 2021 and 2024?
4. **RQ4:** What is the geographic distribution of reported data breaches, and which regions are most significantly affected?

1.3 Significance of the Study

The practical and strategic implications of the study findings are immense to a wide range of stakeholders. This study has the potential to present information based on data that can be applied by executive management and security officers to make decisions intended to control risks and allocate resources more effectively. To ensure the maximization of security investments, companies can align the risk profile of their organization with the trends in the sector, leaving the compliance-driven posture and adopting the intelligence-driven one [6] [7]. In addition, the data regarding the geographic location of violations and the transnational nature of threat actors can be utilized to design the policy of creating closer cooperation and information sharing between states and combating the crimes in cyberspace [8] [9].

2. LITERATURE REVIEW

2.1 A Taxonomy of Modern Cyber Attacks

2.1.1 Ransomware: A dual-extortion strategy of encrypting the data and exfiltration it has become a frequent practice of threat actors who threaten to release the sensitive data onto the Internet publicly unless the ransom is paid [10]. This introduces an element of blackmail which puts a lot of pressure on victim organizations. Also being more mature is the business model of Ransomware-as-a-Service (RaaS), in which the developers of the malware sell it to affiliates who then share a portion of the revenue with the author without requiring high-skill levels [11]. The ransomware effect is significant, and the case of the February 2024 Change Healthcare medical payment processor attack shows that. The attack shook healthcare payments throughout the United States over weeks, and it is an example of the possible impact of ransomware crippling critical infrastructure [12]. In 2023 and 2024, other large organizations, including Starbucks, AT&T, and Amazon Web Services (AWS) users, became the victims of the threat, proving the widespread and nonselective character of this threat [13] [14].

2.1.2 Phishing and Social Engineering: Phishing is a form of social engineering that uses deceptive emails, text messages, or websites to trick individuals into revealing sensitive information, such as login credentials [15]. It remains one of the most common and effective initial access vectors for cybercriminals, responsible for a significant portion of all data breaches [16]. Phishing attacks vary in sophistication. Bulk phishing campaigns cast a wide net, often impersonating well-known brands to lure a small percentage of a large number of recipients [17]. More targeted forms include spear phishing, which uses personalized information about the target to make the lure more convincing, and *whaling*, which specifically targets high-level executives [18].

2.1.3 Hacking and Malware: An eminent case is the massive exploitation of the zero-day vulnerability in the MOVEit file transfer system that affected thousands of businesses worldwide and was a major cause of the 180% growth in the vulnerability exploitation observed in 2023 [19]. It consists of different types, including viruses which are attached to legitimate software; worms which spread themselves through networks; Trojans which simulate harmless software but run dangerous background tasks; and spywares that are used to steal information on a computer belonging to a user [20].

2.1.4 Insider Threats: Insider threats refer to security threats of existing employees or former employees, contractors or business partners who have authorized access to the systems and data of an organization [21]. These threats are classified into two major types. Inspirations malicious insiders are intentional, and driven by revenge, or corporate espionage [22]. A good illustration of ill intent is the case of the 2022 incident in Yahoo where a senior research scientist stole 570,000 pages of trade secrets after a rival company offered him a position [23]. The second and more widespread type is the non-malicious insider. Such events are related to human mistakes, carelessness, or the deceit of a worker by some social engineering attack by an outsider [24]. The percentage of non-malicious insiders is an astounding 75 percent of all insider-related incidents, which underscores the essence of understanding security and the need to have strong internal controls [25]. The magnitude of the issue is high as 83 percent of organizations said that they had at least one insider attack in 2024 [26].

2.2 Synthesis of Major Industry and Academic Research

2.2.1 Verizon Data Breach Investigations Report (DBIR): The Verizon DBIR is among the oldest and most exhaustive studies on real-life security breaches. An analysis of more than 30,000 incidents and 10,000 proven breaches conducted in 2024 found several of the most important trends. One of them is the importance of the human factor, which occurred in 68% of breaches, including mistakes, social engineering, and credential misuse [27]. The number of used vulnerabilities as the first step to breach almost tripled, with a 180% growth compared to the previous year, and most of this is due to the weaponization of zero-day vulnerabilities, such as the one that affected MOVEit. Moreover, in the DBIR 2024, it was noted that the risk of extortion is on the rise, and ransomware or other extortion methods were used in an estimated one-third of all breaches. The report further quantified the escalating risk of interrelated digital ecosystems and indicated that attacks on a third party or a supply chain element had risen by 68% to become 15% of all attacks [28].

2.2.2 IBM Cost of a Data Breach Report: In the 2024 report, which is founded on the interview with organizations who had a breach, the world average cost was at a new high of USD 4.88 million [29]. The IBM reports also give valuable insights on factors which would lessen these costs. One of the most important facts is the direct proportion between the time to respond the breach detection and containment within less than 200 days saved organizations more than a million dollars as compared to longer breaches [30]. Likewise, mature security operations were proven to save companies USD 1.9 million on average in comparison to companies with no security AI and automation, and this aspect is quite relevant to the high utilization of these tools [31].

2.2.3 Previous Academic Studies: The findings of these industry reports are supported with academic research and then expanded. Research on the analysis of databases of publicly reported breaches has reported a steady rise in the rate of breaches of data between 2021 and 2024 with hacking being the most prevalent contributor [32]. Studies have also identified as the most significant motive,

especially when the business-related organizations are attacked as compared to non-profits. This coincides with the industry reports that demonstrate that the overwhelming numbers of attacks are effectively driven. It has also been confirmed in academic analyses that the healthcare industry is among the most targeted industries, and it can be found in numerous pieces of data [33]. One of the common motifs of the academic literature is that it may not always be an easy task to conduct such type of research because of the difficulty of gathering full and objective data because organizations are often unwilling to reveal all details of security incidents as they are afraid of reputational and financial losses [34]. The ongoing focus on the human aspect of the report such as the DBIR indicates the effectiveness and scalability of social engineering attacks that appeal to the employees. Meanwhile, the dramatic increase in exploitation of vulnerabilities also evidences that advanced participants are becoming swifter and more effective dealing with the software and infrastructure-only technical imperfections [35].

2.3 R as a Tool for Cybersecurity Data Analysis

The statistical analysis and visualization platform that was required in this study was robust and flexible due to the quantitative nature of this study. The choice of R programming language was because of its power, extensibility, and its extensive use among the data science and statistics fraternity [36]. Being a free and open-source environment, R encourages reproducible research, and it is possible to create scripts that are easily shared, validated, and reused. Its huge environment of user-submitted packages offers specialized software at each phase of the information examination procedur. To carry out the first steps of data preparation, this paper used the tidy verse, an integrated set of R packages that are useful in data science [37]. Data manipulation tools like dplyr and tidyr provide a standardized grammar to perform tasks in data manipulation like row and column selection, new variable creation and data reshaping between wide and long formats. An important characteristic of the tidy verse is the pipeline operator, enabling the sequencing of multiple operations in a single operation into one readable and linear line of script, which significantly enhances the scalability of the analysis script and its ability to be [38].

2.4 Gaps in the Existing Research

Although the industry reports and the studies conducted by scholars are an invaluable input to the knowledge base on cybersecurity attacks, some gaps are still present. Large industry reports, including those of Verizon and IBM, are background but are usually created using proprietary data based on their own incident response activities, customer surveys, or partner submissions. This has the risk of selection bias, as their datasets may not represent the world picture as a whole, especially with organizations that are not their clients. However, academic research often recognizes the large difficulty in obtaining detailed and authoritative data, commonly depending on one of the public sources or smaller data sets. Moreover, the problem of underreporting is a thoroughly established drawback of cybersecurity studies. The number of incidents never reported publicly is substantial, which implies that databases available to the public are just a tip of the iceberg. According to some academic estimates, the number of cyber incidents reported by the public can be only 3% of the total number of cyber incidents [39].

3. METHODOLOGY

The current research design was a quantitative, descriptive research design to study the trends and patterns in the data of cybersecurity breaches across the world. The secondary analysis of publicly available datasets was the methodological focus. The selected non-experimental design fits the research questions since the main objective was not to test a hypothesis on causal relationships but to observe, describe, and summarize the nature of breach incidents as reported from 2021 to 2024.

3.1 Data Sources

The PRC database gathers and systematizes breach notices, which have been publicly disclosed by state and federal government agencies in the U.S., including different Attorneys General and the U.S. Department of Health and Human Services [40]. Its advantages are that it has a unified method of categorizing types of organizations (e.g., MED in the case of healthcare), and types of breaches (e.g., HACK in case of hacking, INSD in case of insider threat) that are directly related to the research questions in this paper [41]. HIBP gathers information through a massive range of sources, such as the publication of the breached datasets on the hacking and dark web forums [42].

Although it does not contain any industry or organization-specific classifications, its core data i.e. BreachDate and PwnCount (the number of compromised accounts) are very useful in following the chronology and extent of significant international events [43]. Its advantage lies in the fact that it is able to reach the whole world and address the breaches of email credentials.

3.1.1 Dataset Preparation and Selection

To confirm reproducibility, cybersecurity breach data were collected from three publicly available repositories. The first one is Privacy Rights Clearinghouse (PRC),second one is Have I Been Pwned (HIBP) database, and third one the Kaggle Data Breaches dataset. These datasets were accessed and downloaded between March and April 2025. Information from these three sources was reviewed to identify publicly reported breach incidents occurring between 2021 and 2024. The collected data were then cleaned, and duplicate incidents were removed. Only records meeting the study criteria were obtained for analysis. The final and exact analytical dataset consisted of 18 cybersecurity breach incidents.

3.1.2 Inclusion Criteria

The following principles were used to select breach incidents for examine:

- The breach occurred between 2021 and 2024 and was publicly reported.
- The record contained enough and important information on the breach date, organization, and attack type.
- It included information on the number of compromised records whenever available.
- Duplicate incidents across multiple repositories were removed.

3.1.3 Exclusion Criteria

Records were accepted from the analysis if they:

- Occurred outside the study period (before 2021 or after 2024).
- Were duplicate/double entries across different repositories.
- Lacked necessary information such as breach date or organization name.
- Could not be verified using the original data source.

Table 1: Comparison of selected Public Data Breach Repositories

Feature	Privacy Rights Clearinghouse (PRC)	HaveIBeenPwned (HIBP)	Kaggle Aggregated Datasets
Data Collection Method	Compiles and standardizes notifications from U.S. government agencies [44].	Processes breached datasets found online, often from hacking forums or direct submissions [45].	Aggregates information from public sources like press reports and news articles [46].
Key Variables	Date, Organization Type, Breach Type, Records Affected, Location (U.S. State) [44].	Breach Date, PwnCount (records), Domain, Data Classes Compromised [45].	Year, Entity, Records, Organization Type, Method of Breach [46].
Geographic Scope	Primarily United States [44].	Global [45].	Global, with potential regional biases based on source material [46].
Strengths	Structured, standardized data; consistent classification system; based on official reports [44].	Extensive coverage of global breaches, especially those involving credentials; large scale [45].	Broad scope; includes incidents not officially reported to government agencies; easy access [46].
Limitations	Limited to U.S. reports; data completeness depends on state reporting laws [44].	Lacks industry classification; data may come from unverified sources; focuses on email breaches [45].	Inconsistent data quality; relies on media reporting which can be incomplete or inaccurate; potential for duplicates [46].

3.2 Data Collection and Variables

- **Loading:** Reading the raw data files into RStudio.
- **Cleaning:** Before statistical analysis, the dataset was evaluated for missing, incomplete, and appropriate values in the variables Records_Compromised, Attack_Type, Industry, Location, and Year. Where information was incomplete or unidentified, the publicly available source records were reviewed to identify and verify the data whenever possible. Records lacking necessary information required for the analysis were not include during the data preparation stage. Only records containing enough and important information for the study objectives were retained in the final analytical dataset.
- **Harmonization:** Creating a unified classification scheme for key categorical variables. For instance, the Organization type variable from the Kaggle dataset and the organization type classification from the PRC were mapped to a common set of categories (e.g., Healthcare, Finance, Technology, Government, Retail). Similarly, the Method and Breach Type variables were harmonized into a single Attack_Type variable (e.g., Ransomware, Phishing, Hacking, and Insider Threat).
- **Merging:** Combining the cleaned and harmonized datasets into a single, master analytical data frame. Duplicate breach incidents looking in more than one repository were recognized using a combination of breach name, organization name, breach year, and number of compromised records. Where duplicate records discussed to the same incident, the most complete record was recollected. When corresponding information was available across datasets, variables were combined to create a single collective record.

The final analytical dataset contained the following core variables:

- **Date:** A numerical variable demonstrating the calendar year and month in which the cybersecurity breach followed or was publicly revealed.
- **Industry:** A categorical variable for the affected organization's sector.
- **Attack_Type:** A categorical variable for the method of the breach such as Ransomware, malware, phishing etc.
- **Records_Compromised:** A numerical variable indicating the number of individuals or records affected.
- **Location:** A categorical variable for the country and/or U.S. state and regional location of the breached entity.

3.3 Data Analysis Plan

3.3.1 Descriptive Statistics: The first step ushered in the computation of basic descriptive statistics of the whole data. This involved counts and percentages of frequency of categorical variables, such as Industry and AttackType, to get a sense of their total distribution. To provide a summary on the extent of the breaches, measures of central tendency (mean, median) and dispersion (standard deviation, range) were calculated using the numerical variable

RecordsCompromised. This first summary was a broad summary of the characteristics of the dataset.

3.3.2 Trend Analysis: A longitudinal study was conducted to answer RQ1 (from 2021 to 2024) and RQ3 (prevalence of types of attacks). The aggregation was done by year (and month, when the granularity allowed) to form time-series datasets. The dplyr functions groupby and summarise took into consideration the total number of breaches and records compromised totals per time period. It was possible to analyze how the frequency and scale of breaches have changed between the 2021 and 2024.

3.3.3 Comparative Analysis: To respond to RQ2 (patterns in the industry) and RQ4 (geographic distribution), the information was split and contrasted among various groups. The data were aggregated by industry to find the total number of breaches and records compromised in each sector to find the industries most targeted. Cross-tabulations (contingency tables) were created to investigate the correlation between Industry and Attack

4. RESULTS

4.1 Trends between 2021 and 2024 (RQ1)

Figure 1 and Table 2 shows the trends in reported cybersecurity breach incidents and the total number of compromised records between (2021–2024). The number of reported breach incidents increased rapidly from 3 in 2021 to 6 in 2024. It indicates the gradual rise in reported cyber incidents. Although the number of breaches increased over time, the number of compromised records varied gradually across the four years. The highest number of compromised records was observed in 2023 (71.7 million). It reflected the influence of several large-scale breaches, whereas 2022 recorded the lowest total (10.3 million). These findings advised that the severity of cybersecurity breaches is influenced not only by the frequency of incidents but also by the scale of individual breach events.

Table 2: Breach Incidents and Records Compromised by Year

Year	Breach Count	Total Records Compromised
2021	3	18,700,000
2022	4	10,300,000
2023	5	71,700,000
2024	6	31,650,000

Figure 1: Data Breach Trends Over Time

```
library(dplyr)
library(ggplot2)
library(plotly)
library(scales)

# Create Year variable from Date
breach_data$Year <-
as.numeric(format(as.Date(breach_data$Date),
"%Y"))

# Prepare yearly summary
trends_data <- breach_data %>%
 group_by(Year) %>%
 summarise(
  Breach_Count = n(),
  Total_Records = sum(Records_Compromised),
  .groups = "drop"
 )

# Create line chart
p_trends <- ggplot(trends_data, aes(x = Year)) +

 geom_line(aes(y = Breach_Count,
           colour = "Number of Breaches"),
         linewidth = 1.2) +

 geom_point(aes(y = Breach_Count,
           colour = "Number of Breaches"),
         size = 3) +

 geom_line(aes(y = Total_Records/1000000,
           colour = "Records Compromised (Millions)"),
         linewidth = 1.2) +

 geom_point(aes(y = Total_Records/1000000,
           colour = "Records Compromised
(Millions)"),
         size = 3) +

 scale_x_continuous(breaks = 2021:2024) +

 scale_y_continuous(
  name = "Number of Breaches",
  sec.axis = sec_axis(~.*1000000,
                name = "Records Compromised")
 ) +

 labs(
  title = "Figure 1. Data Breach Trends (2021–2024)",
  x = "Year",
  colour = "Metric"
 ) +

 theme_minimal(base_size = 13) +

 theme(
  plot.title = element_text(face = "bold", hjust = 0.5),
legend.position = "bottom"
```

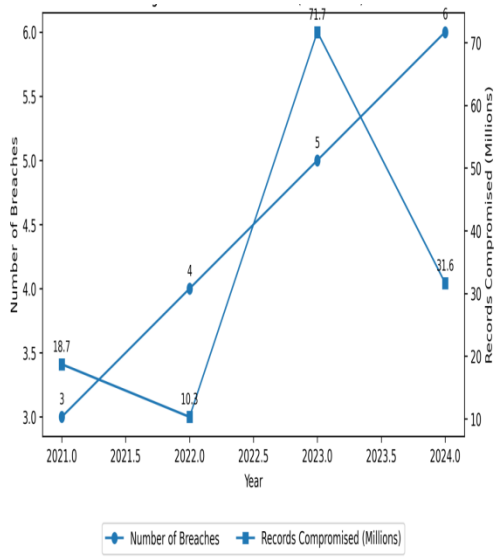

Figure 1. Data Breach Trends from 2021 to 2024

4.2 Industry-Specific Patterns (RQ2).

Table 3: Total Data Breaches by Industry

Industry	Number of Breaches
Healthcare	4
Technology	3
Government	4
Other	7

Figure 2 provides a visual comparison of these numbers, clearly showing that industries handling sensitive personal, and health information are the most attractive targets for attackers.

```
# --- R Code for Figure 2: Horizontal Bar Chart ---
library(dplyr)library(ggplot2)
library(plotly)
# 1. Prepare the dataindustry_data <-
breach_data %>%
  group_by(Industry) %>%
  summarise(Breach_Count = n(), .groups =
"drop") %>%
  arrange(Breach_Count) %>%
  mutate(Industry = factor(Industry, levels = Industry))

# 2. Create the horizontal bar chart
p_industry <- ggplot(industry_data,
                     aes(x = Breach_Count,
                         y = Industry,
                         fill = Industry)) +
  geom_col() +
  geom_text(aes(label = Breach_Count),
            hjust = -0.2,
            size = 4) +
  labs(
    title = "Figure 2. Distribution of Breach Incidents by
Industry (2021–2024)",
    x = "Number of Breach Incidents",
    y = "Industry"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
# 3. Interactive version
interactive_industry_plot <- ggplotly(
  p_industry,
  tooltip = c("y", "x")
)
interactive_industry_plot
filter(Industry != "Financial") %>%
```

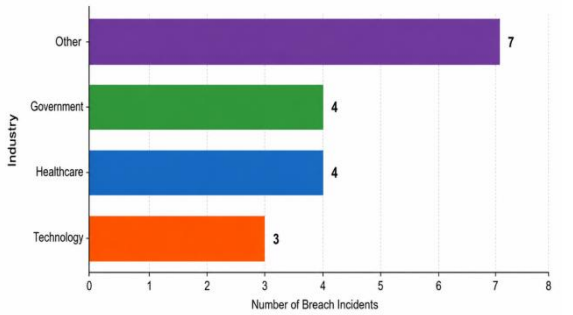

Figure 2. Industry-wise Distribution of Data Breaches (2021-2024)

4.3 C. Dominant Attack Types (RQ3).

Table 4: Frequency of Attack Types

Attack Type	Number of Incidents	Percentage of Total
Hacking	8	44.4%
Ransomware	4	22.2%
Phishing	4	22.2%
Insider Threat	2	11.1%
Total	18	100%

Figure 3 presents the overall frequency of cyber attack types identified in the analyzed dataset from 2021 to 2024. As shown in Table 4, hacking was the most frequently reported attack type, accounting for 8 incidents (44.4%) of all recorded breaches. Ransomware and phishing each accounted for 4 incidents (22.2%). Insider threats showed the smallest proportion with 2 incidents (11.1%). These findings shows distribution of attack types observed during the study period, followed by ransomware and phishing.

```
# --- R Code for Figure 3: Stacked Bar Chart ---

# 1. Prepare the data: Count each attack type for each
year.
attack_type_data <- breach_data %>%
  group_by(Year, Attack_Type) %>%
  summarise(Count = n(),.groups = 'drop')

# 2. Create the stacked bar chart.
```

```
p_attack_type <- ggplot(attack_type_data, aes(x =
Year, y = Count, fill = Attack_Type)) +
 geom_bar(stat = "identity", position = "stack") +
 labs(
  title = "Figure 3: Composition of Attack Types Over
Time",
   x = "Year",
   y = "Number of Breaches",
   fill = "Attack Type"
 ) +
 theme_minimal() +
 theme(legend.position = "bottom")

# 3. Make the plot interactive.
interactive_attack_plot <- ggplotly(p_attack_type)

# Display the interactive plot.
interactive_attack_plot
```

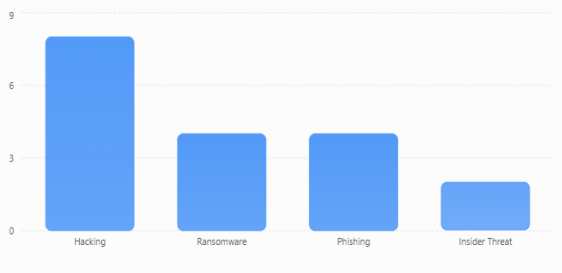

Figure 3. Distribution of Cyber Attack Types (2021–2024)

4.4 Geographic Distribution (RQ4)

The geographic analysis shows that reported data breaches are not spread out evenly across the world. The United States had the largest number of incidents in our dataset, as shown in Table 5.

Table 5: Data Breaches by Country

Country	Number of Breaches
USA	10
United Kingdom	2
Canada	2
Germany	2
Australia	1
France	1

This is likely because the U.S. is a major target and has strong laws that require breaches to be reported publicly. Figure 4 provides an interactive map that visualizes this distribution, showing that while the U.S. has the most incidents, cyber threats are a global problem affecting many developed countries.

```
# --- R Code for Figure 4: Interactive Map ---

# 1. Prepare the data: Group data by location and get
counts.
map_data <- breach_data %>%
 group_by(Location, lat, lon) %>%
 summarise(
  Breach_Count = n(),
  Total_Records = sum(Records_Compromised),
  .groups = 'drop'
 )

# 2. Create the interactive map with leaflet.
breach_map <- leaflet(data = map_data) %>%

addProviderTiles(providers$CartoDB.Positron) %>%
 addCircleMarkers(
  lng = ~lon,
  lat = ~lat,
  radius = ~log(Total_Records) * 2, # Circle size based
on records
  color = "navy",
  stroke = FALSE,
  fillOpacity = 0.6,
  popup = ~paste(
   "<b>Location:</b>", Location, "<br>",
   "<b>Number of Breaches:</b>", Breach_Count,
"<br>",
   "<b>Total Records Compromised:</b>",
format(Total_Records, big.mark = ",")
  )
 ) %>%
 addLegend(
  "bottomright",
  title = "Data Breach Hotspots",
  labels = "Circle size shows records compromised",
  colors = "navy",
  opacity = 1
 )

# Display the map.
breach_map
```

Figure 4. Data Breaches by Country Distribution

5. DISCUSSION

5.1 Interpretation of Findings

To begin with, the digital attack surface is growing exponentially due to popularizing cloud computing, the normalization of remote and mix-hybrid work, and the spread of Internet of Things (IoT) devices [47]. Every new point of connection is a possible entry point. Second, cybercrime has entered the mature phase in its economics, and an exchange of attack tools is commoditized. Lastly, the professionalization of cybercrime syndicates, some of which in turn operate with the tacit or direct endorsement of nation-states, has added in a certain level of organization, resourcefulness as well as tenacity that most organizations are ill-equipped to deal with [48]. Healthcare and government sectors are the best targeted because of the value and the non-portability of the information they contain. Healthcare organization organize and maintain the personal health information (PHI) are valued at high prices on black markets and may be applied to commit identity fraud in the long term [49].

This provides attackers with plenty of bargaining power; a hospital that is unable to access patient records or a bank that is unable to make a payment is under extreme pressure to fix an incident within a short time, in many cases, by paying a ransom [50]. Ransomware poses several points of pressure on the victim organization by combining data encryption (denial of access) with data exfiltration and extortion (threat of disclosure), which significantly raises the risk of payment. The fact that phishing has remained a popular first form of access makes it clear that it is more reliable and cost-effective compared to other technical protection strategies, demonstrating that exploiting human psychology is more usually easier than breaking the technical one.

5.2 Practical Implications

The obvious tendency towards the cybersecurity breaches identified during the study period [2021-2024] proves that cybersecurity is a business risk rather than a technical issue in IT [51]. Organizations should strengthen their cybersecurity policies by implementing effective security pprocedures, regular monitoring, and timely incident response strategies. The findings also explain the importance of regular safety awareness and training programs, as human-related attacks such as phishing/hacking continue to represent a significant threat. The high amount of cost reduction due to quick incident response shows that one of the best methods to prevent security losses is the investment in developing an incident response (IR) plan that is well practiced [52]. The persistence of phishing as an entry point requires constant and extensive security awareness education that is frequently challenged with realistic training to create a robust human firewall. The simultaneous rise in vulnerabilities exploitation requires a vigorous and effective patch management initiative by focusing on critical vulnerabilities in systems facing the internet to decrease the window of attacker opportunity [53].

In addition, security teams ought to lead the use of current architectural concepts such as the zero-trust model, which presupposes that no user and device is inherently trusted, and every request to access is verified, and this can provide the attacking side with limited ways to propagate in a network even after the initial compromise. That is a good reason why harmonized federal or international standards of breach notification should be enforced that mandates essential information to be disclosed in a timely manner, including the exact method of the attack [54]. The international character of the menace similarly clarifies the necessity of augmenting the level of international collaboration in the regulation of cybercrime syndicates, which frequently function with immunity in jurisdictions where the laws are loosely acknowledged [55].

6. LIMITATIONS

Cyber incidences in many cases or most cases are never publicized. It is also possible that organizations do not disclose a breach because it would cause reputational harm, regulatory fines, and even litigation, and there is no legal requirement to disclose breaches in their jurisdiction [56]. According to the academic sources on this phenomenon, the amount of cyber incidents that public databases capture is as low as 3 percent of the overall cyber incidents, which means that the real rate and magnitude of the issue are likely many times higher than those indicated by the results of this study. Second, reported data is usually incomplete and may not be accurate. Breach notifications often contain important information, including the actual attack, the chronology of the breach, or the data that was compromised [57].

The count of the affected individuals is commonly preliminary data that is released by the breached organization, and that organization might have a motive to decrease the number of incidents that were reported. This is restrictive to the richness of the analysis that can be conducted. Lastly, there is a high geographic bias to the dataset. This has been heavily biased in favor of the incidents in the United States, which is a direct outcome of the state-level required notification of breaches in the country, which results in a public record, which is lacking in most other countries [58].

7. RECOMMENDATIONS

Future research must strive to include private data in order to remove the restrictions of the publicly available data, including anonymized cyber insurance claims data or incident response data of cybersecurity companies. It would enable creating more precise statistical descriptions of the threat picture that can be used to measure and address the identified underreporting bias in the public sources. It is impossible to develop a culture of security with only single year training. Second, there should be an destructive and proactive vulnerability management program. The rate at which attackers are now weaponizing vulnerabilities as they are disclosed is forcing organizations to have mechanisms in place that enable them to quickly detect, prioritise and fix critical vulnerabilities, especially those on systems facing the internet.

Third, all organizations need to prepare, write and practice a formal incident response (IR) plan. The statistics indicate that the pace of detection and containment is one of the main factors that define the financial burden of a breach, and an established IR plan is an important cost-reduction measure. Such laws must demand some specific, standardized information, such as the attack vector to enhance the quality of public data and give all stakeholders a superior situational awareness. Moreover, since cybercrime is transnational, it is important to governments to implement and reinforce the collaboration of international law enforcement to effectively study, interrupt, and convict cross-border cybercrime syndicates.

8. CONCLUSION

This secondary data research examined publicly disclosed data concerning cybersecurity breaches, in order to determine and project prominent trends and patterns of the world threat environment. The discussion aimed to address four main questions of the research and the results mostly supported the suppositions based on the available literature. The findings reveal that there is a high and continuous growth in the reported instances of data breaches between 2021 and 2024 and the magnitude of such cases varies depending on the prevalence of mega breaches. The analysis proved that the Healthcare and Government Services sectors are overrepresented, which is typical of the high value of the data that they contain and the sensitivity of their operations to interference. Regarding attack vectors, the results demonstrate the ongoing preeminence of phishing as an initial access technique, as well as the drastic and recent escalation of ransomware as a major danger.

The analysis also presented that healthcare and other critical sectors remain attractive targets because they achieve

sensitive personal and organizational information. These findings highlight the need for organizations to support their cybersecurity posture through regular vulnerability valuations, timely software updates, security awareness training, and well-established incident response techniques. Executing these measures can improve an organization's capability to detect, respond to, and recover from cybersecurity incidents.

Although this study delivers valuable insights into publicly reported cybersecurity breaches, it has several limitations. The analysis was restricted to publicly available breach information covering the period from **2021 to 2024**, and therefore may not signify all cybersecurity incidents worldwide. In addition, the study focused on descriptive trend analysis and did not estimate financial losses or technical characteristics associated with individual breach events.

Future research should develop the analysis by incorporating larger and more recent datasets, additional cybersecurity variables, and advanced analytical techniques such as machine learning and predictive demonstrating. These methods may provide deeper insights into emerging cyber threats and support the development of more effective cybersecurity approaches for protecting organizations against developing digital attacks.

9. REFERENCES

[1] D. Molitor, W. Raghupathi, A. Saharia, and V. Raghupathi, "Exploring Key Issues in Cybersecurity Data Breaches: Analyzing Data Breach Litigation with ML-Based Text Analytics," Information, vol. 14, no. 11, p. 600, 2023.

[2] SentinelOne, "Cyber Security Statistics for 2025 (and Beyond)," 2025. [Online]. Available: https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-statistics/

[3] AAG IT, "The Latest Cyber Crime Statistics," 2022. [Online]. Available: https://aag-it.com/the-latest-cyber-crime-statistics/

[4] IBM, "Cost of a Data Breach 2024: Financial Industry," 2024. [Online]. Available: https://www.ibm.com/think/insights/cost-of-a-data-breach-2024-financial-industry

[5] CyberPilot, "Key Insights from the 2024 Verizon Data Breach Investigations Report," 2024. [Online]. Available: https://www.cyberpilot.io/cyberpilot-blog/key-insights-from-the-2024-verizon-data-breach-investigations-report

[6] Secureframe, "45+ Data Breach Statistics for 2026 [Updated]," 2025. [Online]. Available: https://secureframe.com/blog/data-breach-statistics/

[7] Federal Trade Commission, "Data Breach Response: A Guide for Business," 2023. [Online]. Available: https://www.ftc.gov/business-guidance/resources/data-breach-response-guide-business

[8] Council on Economic Advisers, "The Cost of Malicious Cyber Activity to the U.S. Economy," 2018. [Online]. Available: https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/02/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf

[9] European Union Agency for Cybersecurity (ENISA), "ENISA Threat Landscape 2024," 2024. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

[10] B. Madnick, K. Huang, and S. Madnick, "The Evolution of Global Cybersecurity Norms in the Digital Age: A Longitudinal Study of the Cybersecurity Norm Development Process," Information Security Journal: A Global Perspective, vol. 33, no. 3, pp. 204–225, 2024.

[11] L. Hafiz and T. Hidayat, "Unveiling the Cybercrime Ecosystem: Impact of Ransomware-as-a-Service (RaaS) in Indonesia," International Journal of Science Education and Cultural Studies, vol. 4, no. 1, pp. 11–21, 2025.

[12] S. Berrios, D. Leiva, B. Olivares, H. Allende-Cid, and P. Hermosilla, "Systematic Review: Malware Detection and Classification in Cybersecurity," Applied Sciences, vol. 15, no. 14, p. 7747, 2025.

[13] H. Kryshtal, M. Samofalova, L. Sakhno, V. Fedyna, T. Mokiienko, and M. Yermolaieva, "Cyber Risks in the Financial and Banking System: Analysis of Direct and Systemic Losses," 2025.

[14] S. Navani and G. T. Cirella, "Cybercrimes in the Cryptocurrency Domain: Identifying Types, Understanding Motives and Techniques, and Exploring Future Directions for Technology and Regulation," Journal of Geography, Politics and Society, vol. 14, no. 2, pp. 1–22, 2024.

[15] S. Gallagher, B. Gelman, S. Taoufiq, T. Vörös, Y. Lee, A. Kyadige, and S. Bergeron, "Phishing and Social Engineering in the Age of LLMs," in Large Language Models in Cybersecurity: Threats, Exposure and Mitigation, Cham, Switzerland: Springer Nature, 2024, pp. 81–86.

[16] R. A. Grimes, Fighting Phishing: Everything You Can Do to Fight Social Engineering and Phishing. Hoboken, NJ, USA: John Wiley & Sons, 2024.

[17] F. William, Machine Learning, Fuzzy Logic and Zero Trust-Based Quishing Prevention Solution for Android Smartphones. Rochester, NY, USA: Rochester Institute of Technology, 2024.

[18] CrowdStrike, "What Is a Phishing Attack?" [Online]. Available: https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/phishing-attack/

[19] S. N. Vulpe, R. Rughiniș, D. Țurcanu, and D. Rosner, "AI and Cybersecurity: A Risk Society Perspective," Frontiers in Computer Science, vol. 6, Art. no. 1462250, 2024.

[20] T. T. T. Horn, "A Study of Cybersecurity Landscape in the United States: Trend, Regional Variations, and Socioeconomic Factors," Ph.D. dissertation, Indiana State University, Terre Haute, IN, USA, 2024.

[21] N. Naik, P. Jenkins, P. Grace, D. Naik, S. Prajapat, and J. Song, "An Introduction to Threat Modelling: Modelling Steps, Model Types, Benefits and Challenges," in Proc. International Conference on Computing, Communication, Cybersecurity & AI,

Cham, Switzerland: Springer Nature, Jul. 2024, pp. 260–270.

[22] A. C. M. Day, AL Report 2023, 2024.

[23] J. F. N. González, A. Ortiz, A. C. Cuenca, M. M. Barón, L. Segú, B. Pimentel, et al., "Projection of the Clinical and Economic Burden of Chronic Kidney Disease Between 2022 and 2027 in Spain: Results of the Inside CKD Project," Nefrología (English Edition), vol. 44, no. 6, pp. 807–817, 2024.

[24] Identity Theft Resource Center, "2025 H1 Data Breach Report," 2025. [Online]. Available: https://www.idtheftcenter.org/wp-content/uploads/2025/07/ITRC-H1-2025-Data-Breach-Analysis.pdf

[25] S. M. Obisesan, "Integrating Artificial Intelligence and Cybersecurity Frameworks: Challenges and Opportunities in E-Commerce Cybersecurity Management," SSRN Electronic Journal, 2024, Art. no. 5070108.

[26] S. Metta, I. Chang, J. Parker, M. P. Roman, and A. F. Ehuan, "Generative AI in Cybersecurity," arXiv preprint arXiv:2405.01674, 2024.

[27] H. Collier, "AI: The Future of Social Engineering!," in Proc. European Conference on Cyber Warfare and Security, vol. 23, no. 1, Jun. 2024.

[28] N. A. H. S. Ikram, "Data Breaches Exit Strategy: A Comparative Analysis of Data Privacy Laws," Malaysian Journal of Syariah and Law, vol. 12, p. 135, 2024.

[29] A. Harkai and C. E. Ciurea, "Economic Impact of IoT and Conventional Data Breaches: Cost Analysis and Statistical Trends," Control and Cybernetics, vol. 53, 2024.

[30] P. Kumar, D. Y. Gowda, and A. M. Prakash, "Machine Learning in Cybersecurity: A Comprehensive Survey of Data Breach Detection, Cyber-Attack Prevention, and Fraud Detection," in Pioneering Smart Healthcare 5.0 with IoT, Federated Learning, and Cloud Security, pp. 175–197, 2024.

[31] E. V. Cobos, S. Cakir, S. Straub, C. Z. Qiang, and C. Torgusson, A Review of the Economic Costs of Cyber Incidents. Washington, DC, USA: World Bank, 2024.

[32] P. Mahadevappa, R. Al-Amri, G. Alkawsi, A. A. Alkahtani, M. F. Alghenaim, and M. Alsamman, "Analyzing Threats and Attacks in Edge Data Analytics Within IoT Environments," IoT, vol. 5, no. 1, pp. 123–154, 2024.

[33] S. A. Moamin, M. K. Abdulhameed, R. M. Al-Amri, A. D. Radhi, R. K. Naser, and L. G. Pheng, "Artificial Intelligence in Malware and Network Intrusion Detection: A Comprehensive Survey of Techniques, Datasets, Challenges, and Future Directions," Babylonian Journal of Artificial Intelligence, pp. 77–98, 2025.

[34] F. Romanosky, "Examining the Costs and Causes of Cyber Incidents," Journal of Cybersecurity, vol. 2, no. 2, pp. 121–135, 2016, doi: 10.1093/cybsec/tyw001.

[35] K. Ben Yahia, "The Dark Side of Cryptocurrency Adoption in an Emerging Market: Perspectives of Tunisian Users vs. Professionals," International Journal of Emerging Markets, vol. 21, no. 5, pp. 1479–1502, 2026.

[36] U. Ehsan, S. Passi, K. Saha, T. McNutt, M. O. Riedl, and S. Alcorn, "From Future of Work to Future of Workers: Addressing Asymptomatic AI Harms to Foster Dignified Human-AI Interaction," in Proc. CHI Conference on Human Factors in Computing Systems, Apr. 2026, pp. 1–21.

[37] H. Wickham et al., "Welcome to the Tidyverse," Journal of Open Source Software, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.

[38] G. Ireri, C. Abungu, J. Cheptumo, S. Dounia, M. Gitau, S. Kasaon, et al., "Assessing the Case for Africa-Centric AI Safety Evaluations," arXiv preprint arXiv:2602.13757, 2026.

[39] S. Woods, "Quantifying Under-Reporting in Cyber Incident Data," Journal of Cybersecurity, vol. 8, no. 1, 2022, doi: 10.1093/cybsec/tyac001.

[40] S. Harting, D. Gonzales, M. J. Mazarr, and J. Schmid, Comparative Analysis of US and PRC Efforts to Advance Critical Military Technology. Santa Monica, CA, USA: RAND Corporation, 2024.

[41] D. DiPersio, "Data Protection, Privacy and US Regulation," in Proc. Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-identification of Sensitive Data in Language Resources within the 13th Language Resources and Evaluation Conference, Jun. 2022, pp. 9–16.

[42] A. Hmaidi, "Here to Stay: Chinese State-Affiliated Hacking for Strategic Goals," MERICS Report, 2023.

[43] G. Versluis, Xamarin.Forms Essentials: First Steps Toward Cross-Platform Mobile Apps. New York, NY, USA: Apress, 2017.

[44] Privacy Rights Clearinghouse, Privacy Rights Clearinghouse, 2019. [Online]. Available: https://privacyrights.org/

[45] L. Li, B. Pal, J. Ali, N. Sullivan, R. Chatterjee, and T. Ristenpart, "Protocols for Checking Compromised Credentials," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), Nov. 2019, pp. 1387–1403.

[46] The Devastator, "Data Breaches (A Comprehensive List)," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/data-breaches-a-comprehensive-list

[47] G. Kotronoulas, S. Miguel, M. Dowling, P. Fernández-Ortega, S. Colomer-Lahiguera, G. Bağçivan, et al., "An Overview of the Fundamentals of Data Management, Analysis, and Interpretation in Quantitative Research," Seminars in Oncology Nursing, vol. 39, no. 2, Art. no. 151398, Apr. 2023.

[48] L. Sardinha, M. Maheu-Giroux, H. Stöckl, S. R. Meyer, and C. García-Moreno, "Global, Regional, and National Prevalence Estimates of Physical or Sexual, or Both, Intimate Partner Violence Against Women in 2018," The Lancet, vol. 399, no. 10327, pp. 803–813, 2022.

[49] Bluefin, "2024 ITRC Data Breach Report: Record-Breaking Breaches," 2024. [Online]. Available: https://www.bluefin.com/bluefin-news/2024-itrc-data-breach-report-record-breaking-breaches/

[50] Bank of America, "The Costs of a Cyber Incident Can Ripple Through the Economy," 2024. [Online]. Available: https://www.privatebank.bankofamerica.com/articles/cyber-incident-costs.html

[51] C. Elendu, E. K. Omeludike, P. O. Oloyede, B. T. Obidigbo, and J. C. Omeludike, "Legal Implications for Clinicians in Cybersecurity Incidents: A Review," Medicine, vol. 103, no. 39, Art. no. e39887, 2024.

[52] R. A. Alsharida, B. A. S. Al-Rimy, M. Al-Emran, and A. Zainal, "A Systematic Review of Multi Perspectives on Human Cybersecurity Behavior," Technology in Society, vol. 73, Art. no. 102258, 2023.

[53] S. Backman and T. Stevens, "Cyber Risk Logics and Their Implications for Cybersecurity," International Affairs, vol. 100, no. 6, pp. 2441–2460, 2024.

[54] A. I. Sani, A. O. Olajide, O. V. Abosede, D. F. Oyebode, R. Vayyala, J. G. Alfred, et al., "Cybersecurity Challenges in Digitizing Government Administration," Proceedings of the International Academy of Sciences, vol. 2, no. 1, pp. 1–13, 2025.

[55] W. Tuleun, "Analysis of Cybercrimes, Major Cyber Security Attacks and the Overall Economic Impact on Nigeria," World Journal of Advanced Research and Reviews, vol. 16, no. 1, pp. 1196–1221, 2022.

[56] M. Alshaikh, S. Mehmood, R. Amin, and F. S. Alsubaei, "The Impact of Cybersecurity Through Knowledge Sharing Practices: Limitations, Analysis of Current Trends and Future Research Directions," International Journal of Advanced Computer Science and Applications, vol. 16, no. 3, 2025.

[57] S. Backman and T. Stevens, "Cyber Risk Logics and Their Implications for Cybersecurity," International Affairs, vol. 100, no. 6, pp. 2441–2460, 2024.