**Exploring the Causes of Data Breaches and Strategies for Prevention**

Issam Awamleh and Safaa hriez

1 Al Hussein Technical University, Amman 11831, Jordan

21110241@htu.edu.jo, safaa.hriez@htu.edu.jo

**Abstract**

**Data breaches are one of the most painful cybersecurity challenges facing industries today, especially triggering severe financial and reputational pain. This research then delves into the main causes of data breaches across various sectors — healthcare, finance, and retail — identifying weaknesses of human error, insiders threats, poor password policies, unpatched software, and misconfigured systems. This study proposes to use real world case studies as well as current industry practices to critically examine the most significant gaps in security frameworks. To tackle these weaknesses, the research presents viable tactics, including employee training, strict access controls, consistent software updates, data encryption of highly sensitive data and complete incident response plans. These findings are directed towards the making of actionable knowledge for cybersecurity specialists about this and building the roadmap for enhancing data protection mechanisms. Additionally, the study underlines the importance to develop locally tailored approaches and to continuously innovate to beat new threats and to prevent new incidents.**

**Keywords:** Data Breaches, Cybersecurity Strategies, Vulnerability Mitigation, Industry-Specific Risks, Incident Response

# Introduction

Two paragraphs about your topic (the topic of the dataset).

[1] point to Data breaches or data leak could be one of the serious danger of any business or organization also a source of concern regardless that danger on the data internal or external (hacking attacks). Normally the data consider one of the significant assets of any enterprise. A data breach allows unauthorized individuals to access private, sensitive, or protected information. In a data breach, the files are viewed and/or distributed without authorization.

Data breaches can occur due to various factors, including intentional attacks, oversights, or infrastructure flaws, rather than solely attributed to outside hackers.in this paper study the cases of the breaches that happened to point to the most main reason why these events could have happened. Furthermore, to be this paper as indicator for the enterprise to rearrange their security procedure.

One paragraph about how they use AI in your specific topic.

Jay Trinckes [2] explained that the Cyberattacks are one of the new dangers that AI may bring. Although AI is capable of accurately anticipating and thwarting threats, it may also be used as a weapon for ransomware and sophisticated phishing campaigns. It could be using for in red teaming in blue teaming Cybercriminals are too creating profiling strategies, foreseeing and abusing person behaviors for profoundly personalized assaults. As AI apparatuses gotten to be more complex, organizations must adjust their security measures to ensure delicate information. A comprehensive technique, counting encryption, get to controls, security appraisals, and fix administration conventions, is pivotal for tending to security and protection concerns. AI is being utilized as a vigorous protective instrument in compliance and security instruments, empowering proactive danger chasing, peculiarity location, and prescient approaches in businesses like healthcare. AI is being utilized as a vigorous protective instrument in compliance and security instruments, empowering proactive danger chasing.

One paragraph about your research question, and the objectives for this research, and illustrates your contributions clearly.

What are the main causes of data breaches in different industries, and how smart specialists can overcome these weaknesses and avoid future events? It lays the foundational groundwork for and exploring the complex and dynamic space of cybersecurity threats. This research focuses on identification of the root causes of data breaches with diverse sectors and analysis of the vulnerabilities induced in the industry with the main intent to devise the excellent strategies to mitigate these risks. This study adds to the cybersecurity field by analyzing case studies, examining the patterns of breach occurrences, and reviewing current security practices and offers actionable recommendations appropriate to different industries. As part of the research, methods of proactive measures, including employee training, robust security policies and advanced threat detection mechanisms, are highlighted to further improve data protection. The objective of the study is ultimately to provide practitioners with meaningful insights and tools to proactively reduce the potential for future incidents and ultimately create a safer digital environment.

One paragraph describes the structure of the document. For example, it could be "Section 2 presents the related work. Section 3 discusses the data collection and description. Section 4 illustrates the proposed approach. Section 5 presents the experiments and results. Finally, Section 6 concludes the work and provides recommendation based on the conclusion."

In Section 2, we provide an overview of the related work and summarize the relevant literature, and describe the research gap in understanding the causes and prevention strategies for data breaches. In section 3, the primary survey methodology along with the secondary dataset, and the problems encountered during data acquisition are discussed. In section 4, we outline a proposed research approach including preprocessing (label encoding), and the rationale for using machine learning models (SVM, Random Forest, XGBoost, AdaBoost). The experiments and results are discussed in section 5, including insights from survey responses and visualisations of model performance evaluation. The last section of the research, Section 6 concludes by summarizing these findings and how they correspond to the objectives, and offers recommendations on mitigating data breaches, while offering future research directions.

# Related Work

Two lines introduction about the field of the research papers that will be illustrated in this section. For each paper write 1-3 paragraphs summary that includes the methodology of the paper, its main result, the limitation or weakness of the paper. At the end, add a paragraph about the gap in the literature review that is addressed in your research work.

[3] This paper reviews enterprise data breach causes, challenges, and future prevention strategies, focusing on the threats of intentional or unintentional data leaks that expose sensitive information like employee and customer data, intellectual property, and medical records. It analyzes recent high-profile incidents, such as Target’s 2013 breach and Yahoo’s 2014 breach, showing the substantial financial and reputational damage that data leaks cause. Research indicates that insiders account for over 40% of breaches, with motivations ranging from espionage to accidental errors. The paper explores various state-of-the-art prevention and detection techniques, such as Data Leak Prevention and Detection (DLPD) systems that monitor and control data flows using policies and behavior modeling. However, the big data era presents new challenges, as modern communication tools (e.g., cloud sharing, email, messaging) increase leak risks. Future directions for reducing data breaches include developing privacy-preserving detection systems, cloud service leak detection, and deep learning-based anomaly detection for insider threats, aiming to transition towards more robust and adaptable solutions in enterprise security.

[4] pointed to Computerized change in healthcare has progressed openness to treatment, but has moreover expanded helplessness to information breaches, with hacking/IT occurrences being the foremost common shape, taken after by unauthorized inside divulgence From 2005 to 2019, 249.09 million individuals were influenced and 157.4 million were influenced within the final five a long time alone.In 2018, the industry experienced 536 breaches, compromising 41.2 million records.The normal breach taken a toll in 2019 come to $3.92 million, a worldwide tall, reflecting a 12% increment since 2014. The think about connected straightforward moving normal (SMA) and straightforward exponential smoothing (SES) strategies to foresee data breach patterns utilizing information from the PRC database and HIPAA diaries, and found that SMA was more dependable. The comes about uncovered that 64% of restorative data spills since 2005 were due to hacking, and more than 92% of records have been compromised by hacking in later years.In 2019, hacking occurrences expanded by 73.4%, whereas other spills such as theft/loss and disgraceful transfer sorts diminished. The normal taken a toll per restorative breach rose to $6.45 million, and the fetched per record expanded 5.14%. The consider underscores the require for expanded security measures in healthcare and advocates proactive techniques to address the developing hazard and financial effect of information breaches.

[5] This study examines the prevalence of personal information leaks among Americans in data breaches that lack detailed research, particularly with respect to individual risk. Using a new dataset, the researchers estimated the minimum average number of online accounts compromised per capita by integrating data from a representative YouGov sample with information from Have I Been Pwned (HIBP), which catalogs 293 public information breaches. The results of the analysis of 5,000 e-mails revealed that 14,979 breaches occurred, with an average of three breaches per person and at least 82.84% of Americans experiencing at least one breach. Socioeconomic factors revealed some interesting trends. The frequency of breaches increases with education level, with the average number of breaches for individuals without a high school degree being 2.35, while the more educated are more likely to be breached. The study also found that middle-aged and older accounts face higher risks than younger and older users. Female accounts are 1.12 times more likely to be compromised than male accounts, black accounts are 3.12 times more likely to be compromised than male accounts, and white accounts are 3.16 times more likely to be compromised than white accounts. Of the 15,837 breaches examined, 94.58% were confirmed and one-third were classified as spam listings. The study emphasizes that the relationship between the frequency of breaches and factors such as education and age differs from traditional digital divide concerns, indicating that increased online activity is correlated with a higher risk of breaches, especially among more educated users.

[6] Explain more about Ransomware is a significant cybersecurity threat, causing data breaches and disruptions. However, many studies lack consideration of government strategies, industry guidelines, and cyber intelligence. A study evaluating 212 academic studies found that many were irrelevant to the current reality. The study proposed prioritizing data exfiltration over encryption, considering ransomware in a business-practical manner, and recommending collaboration with the industry to address this evolving threat.

This survey compiled the ransomware evolution history and applied Rogers' Innovation Adoption Curve, predicting the rise of destructive ransomware with espionage. It reviewed 212 academic studies and found that most research has become less relevant in the era of ransomware double extortion with data exfiltration. The survey proposed integrating ransomware risk management into organizational cybersecurity risk management, emphasizing government strategies, industry reports, guidelines, and cyber intelligence. It also discussed innovative research prospects, including generative AI, and suggested future research directions.

[7] Despite significant cybersecurity investments, companies face ongoing data breaches that have significant financial and reputational consequences. To protect digital assets and improve threat visibility, threat intelligence employing AI and machine learning is emerging, moving from reactive to proactive defense strategies. This evolution of threat intelligence aims to predict security threats by analyzing and integrating cyber data to gain insights tailored to an organization's unique risk landscape, enhancing both visibility and incident response. It is critical to distinguish between data loss (unintentional) and data breach (intentional), and data loss prevention (DLP) addresses the former while incident response planning addresses the latter, analyzing system logs for forensic purposes after a breach. Effective machine learning solutions in cybersecurity must be aligned with business goals and security standards, and must focus on specific threat scenarios with high-quality training data sets to maximize predictive accuracy and relevance. However, machine learning also introduces vulnerabilities to hostile attacks, so continuous learning is essential to adapt models to changes in the threat environment and mitigate concept drift. Continuous monitoring, root cause analysis (RCA), and improved mapping of infrastructure and threat interdependencies will continue to be essential in managing organization-specific risks and supporting continuous adaptation to the evolving cybersecurity landscape. Despite significant cybersecurity investments, companies face ongoing data breaches that have significant financial and reputational consequences. To protect digital assets and improve threat visibility, threat intelligence employing AI and machine learning is emerging, moving from reactive to proactive defense strategies.

This evolution of threat intelligence aims to predict security threats by analyzing and integrating cyber data to gain insights tailored to an organization's unique risk landscape, enhancing both visibility and incident response. It is critical to distinguish between data loss (unintentional) and data breach (intentional), and data loss prevention (DLP) addresses the former while incident response planning addresses the latter, analyzing system logs for forensic purposes after a breach. Effective machine learning solutions in cybersecurity must be aligned with business goals and security standards, and must focus on specific threat scenarios with high-quality training data sets to maximize predictive accuracy and relevance. However, machine learning also introduces vulnerabilities to hostile attacks, so continuous learning is essential to adapt models to changes in the threat environment and mitigate concept drift. Continuous monitoring, root cause analysis (RCA), and improved mapping of infrastructure and threat interdependencies will continue to be essential in managing organization-specific risks and supporting continuous adaptation to the evolving cybersecurity landscape.

# Data collection and Description

Define the primary and secondary data and describe the differences between them.

Primary Data: The information gathered does not come very close to the subject of the study unless that is primary information, also called information that has been collated specially for the particular study. It’s original, shaped to the research question (i.e., it might be an interview, survey, focus group, experiment, etc.). For this research, survey was conducted among the cybersecurity professionals in order to understand the causes of data breaches, reporting practice and how mitigate strategies are working.

Secondary Data: Data gained by other sources besides you are called secondary data which includes information that is either already collected, processed and then made available either by other sources (such as research papers, organizational reports, public datasets, or governmental records or by yourself). The second and most utilized secondary data was a dataset which is a preexisting one related to cyber threats that had been preprocessed and used with the use of machine learning models.

**Differences Between Primary and Secondary Data:**

Purpose: Specifically, primary data, whereby the research question is answered by collecting it on its own, is gathered, while secondary data, from existing sources, is repurposed to answer it.

Originality: Primary data is collected by the researcher for the first time while secondary data has been previously collected and have been used for other studies.

Effort and Cost: Primary data is much more effort, time, and resources taken to collect while secondary data is much more cost effective and easily available.

Relevance: Secondary data are more often in need of preprocessing or filtering (removing irrelevant data) to make them relevant to the research objectives, while the primary data is more directly aligned to those objects.

Control: All the control over the data collection lies with researchers but they are obliged to use secondary data in them researches in same or altered structure and limitations.

Mixing firsthand information garnered from experts with statistical reasoning derived from existing data sources, this research brings together the best of both worlds: raw information and hardcore analysis, all to further our knowledge of data breaches.

3.1 Primary Data

Describe your primary data (it could include figures and tables).

Describe your experience in collecting your primary data. Include any issues you faced in this process.

Justify your choice.

Describe the merits and limits of primary data.

**Description of primary data:**

Primary data was obtained from an online self-administered survey distributed to cybersecurity workers and cybersecurity specialized people who have dealt with events such as data breach. The survey was shared through two main channels:

Direct Outreach: The emails were sent to cybersecurity enthusiasts who include cybersecurity academicians and cybersecurity industry professionals.

Social Media Groups: The survey was shared in professional cybersecurity groups, cybersecurity conferences, cybersecurity group in LinkedIn and other such platforms where cybersecurity professionals are active.

The data collected aimed to gather insights on:

* More to the point, the frequency and type of data breach incidents are changing.
* Current defense mechanisms.
* People’s perception of the suitability of primitive measures.
* New risks that contribute to the data breach management.

**Describe about experience in collecting the primary data:**

Outreach Issues:

I faced challenge in accessing cybersecurity personnel through email particularly individuals with field experiences in handling real cyber security instances. A number of these experts could not be reached or chose not to respond, thus reducing the range of input from this focus group.

Trust Barrier on Social Media:

Using survey links on social networks was an issue of concern when it came to trust. Some people refrained from clicking the links presumably because of the phishing fears or personal data protection.

Low Response Rate:

In general, the response rate was not as high as desired even though the survey followed a logical design, some key targeted groups such as high rank professionals might not have found time to be more responsive to the survey questions.

**Justify the choice:**

The survey method was chosen for the following reasons:

* Relevance: The most ideal people to seek insight on data breaches, their underlying reasons and mitigation strategies are cybersecurity experts. Direct input from the target audience was provided by surveys.
* Flexibility: Both qualitative and quantitative data in surveys can be captured, thus surveys can be used to analyze cybersecurity incident trends and patterns.
* Cost and Time Efficiency: Unlike interviews or focus groups surveys are more scalable and quicker to collect a large group.

**Merits of Primary Data:**

Specific to the Research Objective:

When tailored to the research problem, primary data is relevant and specific to the problem.

Up-to-Date Information:

The survey acts as an instant source of current challenges and strategies in cybersecurity.

Original Insights:

This is because primary data collection allows the researcher to come up with unique findings by accessing the source directly as you would not necessarily get it in the secondary data.

**Limits of Primary Data:**

Time consuming:

Designing, disseminating and collecting survey responses consumed large amount of time, in particular when response rate was low.

Potential Bias:

This means that respondents may not always give honest or accurate answers when asked and this could lead to biases in the data.

Access Challenges:

One of the biggest constraints was reaching experienced cybersecurity professionals and convincing them to trust us.

Limited Sample Size:

Low participation rates mean that the sample size is no representative of the broader cybersecurity community.

**The questions: note: the yellow highlight is check box question.**

**Section: Data breach trends.**

1. **In your professional experience, what is the most common cause of data breaches in the industries you specialize in?**
   * A) Phishing attacks
   * B) Insider threats
   * C) Misconfigured systems
   * D) Weak or stolen credentials
   * E) Other

**(that is the main cause of the data breaches)**

1. **How frequently do organizations under your expertise report data breaches?**
   * A) Weekly
   * B) Monthly
   * C) Quarterly
   * D) Annually
   * E) Rarely

**(the frequently of the data breaches incidents)**

1. **What type of data is most commonly targeted in breaches you’ve investigated?**
   * A) Personal Identifiable Information (PII)
   * B) Financial data
   * C) Intellectual property
   * D) Healthcare records
   * E) Other

(the consequence of the breaches and most targeted get effected)

**Section: data breach prevention.**

1. **Which security measure do you find most effective in preventing data breaches?**
   * A) Regular security audits
   * B) Multi-factor authentication (MFA)
   * C) Encryption of sensitive data
   * D) Continuous monitoring and detection systems
   * E) Other
2. **In your professional opinion, how effective are organizations at identifying breaches in their early stages?**
   * A) Very effective
   * B) Moderately effective
   * C) Somewhat effective
   * D) Ineffective
   * E) Other

**Section: data breach impact and response.**

1. **What do you observe as the biggest consequence of data breaches for organizations?**
   * A) Financial loss
   * B) Reputational damage
   * C) Regulatory fines
   * D) Loss of customer trust
   * E) Other
2. **How prepared are organizations to respond effectively to a major data breach?**
   * A) Very prepared
   * B) Moderately prepared
   * C) Somewhat prepared
   * D) Not prepared

**Section: data breach impact and response.**

1. **What area should organizations invest in most to reduce the risk of data breaches?**
   * A) Employee training on cybersecurity awareness
   * B) Advanced threat detection tools
   * C) Incident response planning
   * D) Regular vulnerability assessments and penetration testing
   * E) Other

.

3.2 Secondary Data

Describe your secondary data (it could include figures and tables).

Justify your choice.

Describe the merits and limits of secondary data.

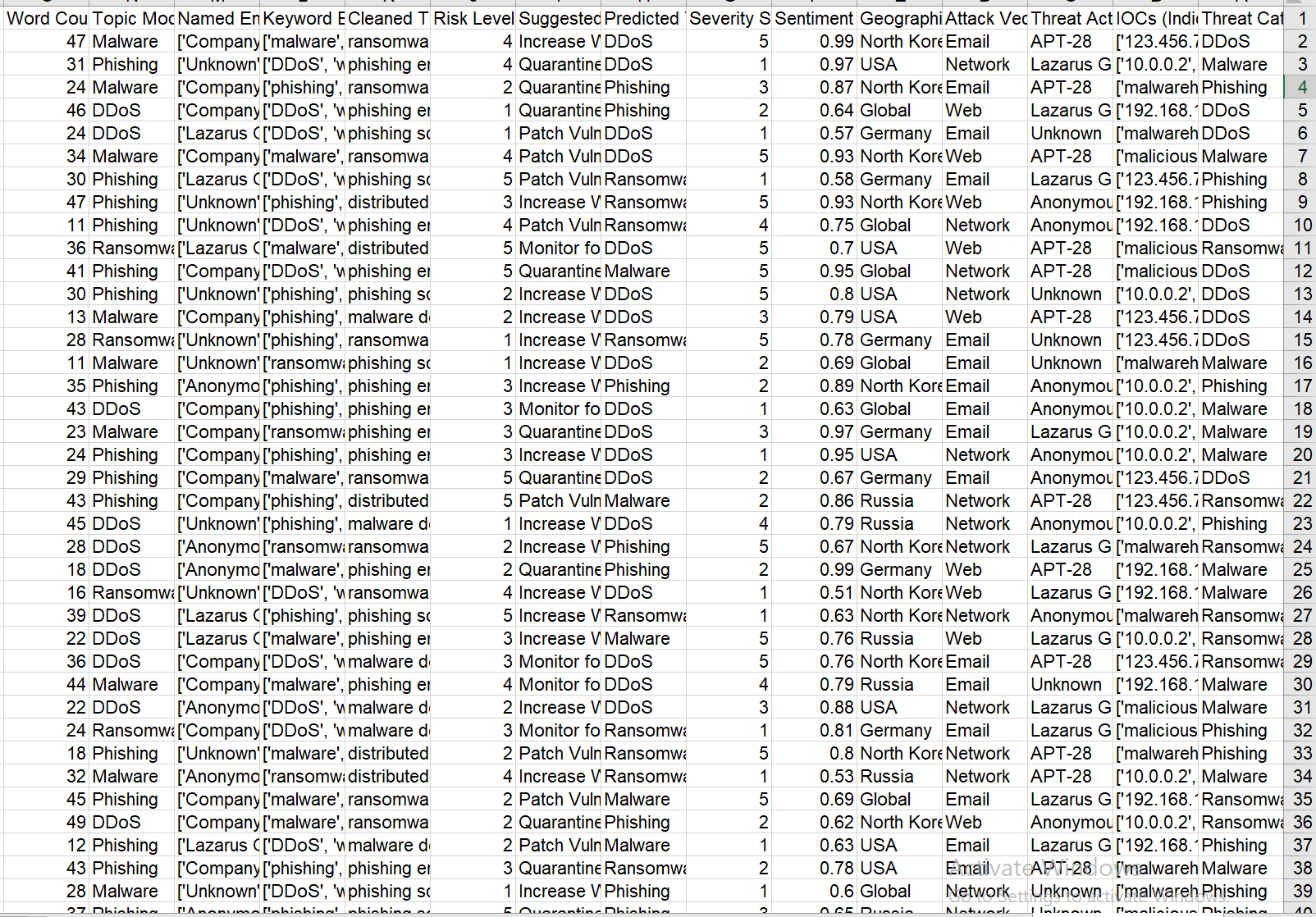
About the datasets:

[8] explained In order to facilitate Cyber Threat Intelligence (CTI) research, this dataset comprises 1100 examples of cyber threat reports that have been augmented with Natural Language Processing (NLP) features. The collection consists of textual descriptions, indications of compromise (IOCs), and several danger classifications. With features to recognize, categorize, and rate the seriousness of cybersecurity risks, each event is a distinct danger report.

|  |  |  |
| --- | --- | --- |
| Feature Name | Feature Description | Expected Values |
| Threat Category | Describes the type of threat | ['Phishing', 'Ransomware', 'DDoS', 'Malware'] |
| IOCs (Indicators of Compromise) | Lists known indicators of compromise | List of strings, e.g. ['192.168.1.1', 'malicious.com'] |
| Threat Actor | The group or entity behind the attack | ['Unknown', 'APT-28', 'Anonymous', 'Lazarus Group'] |
| Attack Vector | The method used to execute the attack | ['Email', 'Network', 'Web'] |
| Geographical Location | The country or region where the attack originated or was targeted | ['USA', 'Russia', 'Global', 'North Korea', 'Germany'] |
| Sentiment in Forums | Sentiment analysis score (0.5 to 1.0) from hacker forums | Float value between 0.5 and 1.0 |
| Severity Score | A severity rating from 1 (low risk) to 5 (critical) | Integer between 1 and 5 |
| Predicted Threat Category | The predicted category of the threat based on model analysis | Same as Threat Category feature |
| Suggested Defense Mechanism | The recommended action to mitigate the threat | ['Monitor for Phishing', 'Patch Vulnerability', 'Increase Web Security', 'Quarantine'] |
| Risk Level Prediction | A prediction of the overall risk level (1 to 5) | Integer between 1 and 5 |
| Cleaned Threat Description | A cleaned and preprocessed description of the threat | String |
| Keyword Extraction | Important keywords extracted from the description | List of strings, e.g. ['phishing', 'link', 'email'] |
| Named Entities (NER) | Entities identified within the description (organizations, malware names) | List of strings, e.g. ['CompanyX', 'APT-28'] |
| Topic Modeling Labels | Labels assigned based on topic modeling (e.g., LDA) | Same as Threat Category feature |
| Word Count | The number of words in the cleaned threat description | Integer |

**Figure (1) (Table for the secondary data)**

Screenshot of the datasets:



**Figure (2) (preview for the dataset)**

[The datasets link](https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset)

**Choosing justification:**

1\_ Diversity of Threats: This dataset also contains many types of threats like DDoS, phishing, malware, and ransomware, which can be used to analyze different attack vector and pattern.

2\_Rich Metadata: It includes those key attributes such as:

* IOCs.
* Attack vector and threat actors.
* Sentiment analysis and geographical impact.
* Predicted risk levels and suggested defense mechanisms.

3\_Actionable Insights:

* It offers clear links between threats to Java programs and suggested mitigation strategies.
* It enables detect threat severity sequences and geographical trends.

4\_Relevance to Research:

* It provides support to understand attack methodologies and organization vulnerabilities.
* It contributes significantly in shaping the practical response to the problem of threat prevention and response.

**Merits of Secondary Data**

1\_Cost-Effective:

Pre collected and organized, secondary data does not have to be gathered at resource intensive levels.

2\_Time-Saving:

Immediate analysis is possible with fast access to a well-structured dataset.

3\_Broad Scope:

Data covers the different aspects of a breach covered in full.

4\_High Reliability:

It looked like some of the data is collected systematically in respect to technical and contextual details.

5\_Data Availability:

It contains a mix of quantitative (severity score, risk levels) and qualitative (descriptions and keywords cleaned).

**Limitations of Secondary Data**

1\_Data Quality Issues:

The fields such as "Named Entities" contain vague entries like "Unknown," that makes analysis less precise.

2\_Lack of Context:

With secondary data, you can never be fully sure to know how the data gets collected, and that can create gaps in your understanding.

3\_Limited Customization:

Specific research needs may not match pre-defined structures and variables.

4\_Outdated Information:

The dataset is updated only if the dataset doesn’t match with recent trends or new attack methods.

5\_Bias in Data Collection:

Possible biases in selection of threat types, sources, or regions; lack of applicability to global scale.

# Research Approach and Methodologies

1 line illustrates what you will include in this section. talk about the onion mode

To present the methodological research approach and tools employed in the study, the Research Onion Model will be used and What, Why, how approaches will be outlined, as well as several layers, such as research philosophy, research strategy, time horizon, and data collection techniques.

### Onion Model

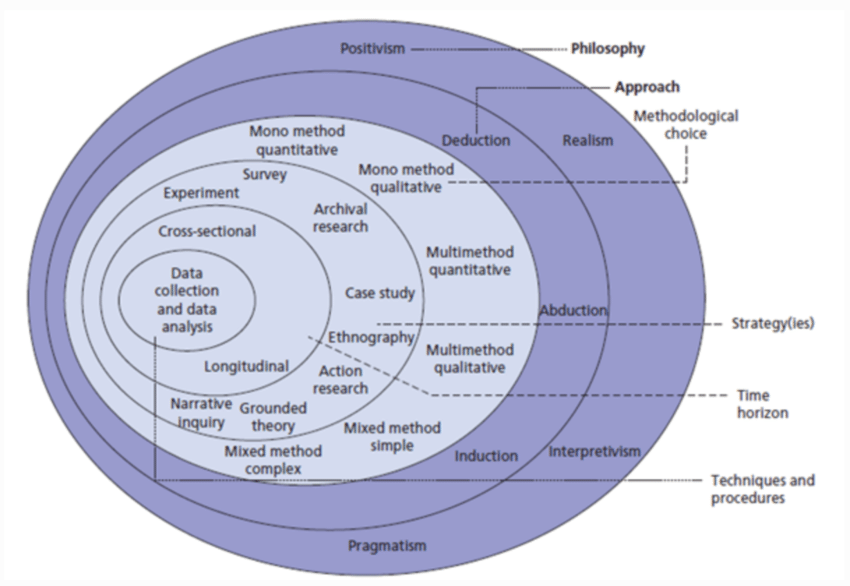


Figure 1: The structure of the Onion Research Model [1].

* + 1. **Philosophy**

Discuss different research approaches and methodology for the Philosophy layer. Provide justification of your choice.

**Positivism**: Designed with objective data collection and the responsibility of measurable outcomes, making it a good candidate for analyzing cybersecurity incidents (such as data breaches).

**Justification and reasons:**

Objective Approach: Positivism is something that seeks data to measure, analyze, and study. This data breach survey is also appropriate because it seeks to gather factual data from cybersecurity professionals.

Quantitative Nature: As is the case of numerical data (e.g., survey results and the performance metrics from machine learning models), positivism supports their usage because it involves statistical analysis and objective truths.

Scientific Testing: Hypothesis testing can be endorsed by positivism (in deploying inquiry into the causes or patterns of data breaches, for example). Collecting data and analysis through research is your holy grail here so you are hoping to validate these hypotheses.

Generalizability: Positivism enables you to analyze data systematically, and reach conclusions of general applicability that can be applied to all of cyber security best practice, like recommendations, for example.

* + 1. **Theory Development Approach**

Discuss different research approaches and methodology for the Development Approach layer.

Provide justification of your choice.

**How is the Deductive Approach used?**

Data breaches are initially fallen into existing studies on cybersecurity vulnerabilities, human errors, or technical flaws with deductive approach, under general theories or frameworks about data breaches.

Then you run these theories against your collected data (your survey responses) to find out which exact causes apply in your case.

**Justification for Deductive Approach in Your Research:**

Existing Theories: There is such a lot of literature and theoretical frameworks about data breaches ranging from the topics area of IT risk management, Information risk management, Security risk management, Information security management, Data security management, Data privacy management, Information privacy management, and Data encryption management.

Human factor causes, as in (e.g., phishing, poor password hygiene).

Related to system vulnerabilities (old software or no encryption).

Testing Hypotheses: For instance, you can postulate that

"Human error is the leading cause of data breaches."  
"Organizations with outdated software are more prone to data breaches."

Survey Data: As the deductive approach allows you to systematically test these hypotheses with your survey responses, it allows you to test them systematically.

Quantifiable Results: You can quantitatively validate or refute the proposed causes of your problem by analyzing your survey results.

**Steps in the Deductive Approach for Your Research:**

Step 1: Approach to Data Breach Causes: Identify theorize existent theories for causes of data breach.

Example: It could be from ISO 27001 standards, Verizon Data Breach Investigations Report or from academic research on cybersecurity incidents.

Step 2: Defining Hypotheses: According to these theories it is possible to define hypotheses, for instance:

"The human errors are the source of most data breaches."

"In fact, the most commonly cited cause of breaches is technical vulnerabilities."

Step 3: Hypothesis: These hypotheses are validated through survey design of a survey aimed at cybersecurity experts.

Step 4: Statistical Methods of Data Collection and Analysis: How to use statistical methods to interpret survey data and focus on which causes are most important.

Step 5: Findings are used to conclude to either confirm or reify the hypotheses and provide a more nuanced understanding of the data breach causes.

* + 1. **Methodological Choice**

Discuss different research approaches and methodology for the Methodological layer. Provide justification of your choice.

**Qualitative vs. Quantitative Research: A Comprehensive Comparison**

Qualitative and doing Quantitative research methods serve as important tools of various fields such as Science, Social Science, and Business. Each provides somewhat different ways to collect and analyze data, strengths and limitations. Here's a comprehensive comparison:

|  |  |  |
| --- | --- | --- |
| Feature | Qualitative Research | Quantitative Research |
| Purpose | Exploration, understanding | Measurement, testing |
| Data Type | Non-numerical (words, images, etc.) | Numerical |
| Methods | Interviews, observations, case studies | Surveys, experiments, statistical analysis |
| Analysis | Interpretation, thematic analysis | Statistical analysis |
| Sample Size | Small | Large |
| Focus | In-depth understanding, context | Generalizability, objectivity |

**Figure (3) (comparison between Qualitative and Quantitative Research)**

**Mixed Methods Research:**

However, it's important to note here that the research is not exchangeable or cannot be done using only one or the other types. As a combination of approaches, mixed methods research integrates the two in order to provide a more holistic view of a research problem. This approach promises the depth of the qualitative data as it also offers generalizability of the quantitative data.

**Mono Method Quantitative**: The use of surveys for the purpose of numerical data collection and statistical analysis.

**Justification**: It provides the measured insights of the trends and patterns for cybersecurity incidents.

Because it is efficient and easily scalable, and can collect both quantitative and qualitative data, it is the perfect survey method to research the causes of data breach. It allows for data collection of structured and consistent type from a large crowd thus allowing for diverse and reliable insights. Anonymity in surveys also encourages response to a particularly sensitive topic like that. To overcome trust constraints and accessibility challenges, I distributed surveys online to social media and professional platforms where experts on the topic participated. Despite this, surveys are advantageous due to the fact that they are time saving and generalize more extensively than interviews or case studies and are convenient to do.

* + 1. **Research Strategy**

Discuss different research approaches and methodology for the Strategy layer.

Provide justification of your choice.

**Research Strategy: Survey**

Firsthand data on the causes of data breaches was collected using the survey strategy as the most proper research method from cybersecurity professionals and experts. With a structured and efficient approach to ask questions on a small sample yet a large and diverse audience, surveys enable you to understand common trends, opinions and experiences on data breaches.

**Surveys are particularly effective for this research for several reasons:**

* Efficiency in Data Collection: One of the benefits of surveys over one on one interviews or focus groups is the ability to upscale the questions and have them collected from the same number of people, saving an enormous amount of time. If you want to reach a big group of cybersecurity professionals around the globe, it’s important to do this.
* Anonymity and Honesty: With anonymity, participants are more likely to give honest feedback on such topics as data breaches. With this, we are able to collect unbiased and authentic data.
* Scalability: Surveys can be conducted with even larger sample size, and better data, through online platforms. For example, it’s particularly useful to get insights from across the board in various regions and business areas.
* Structured Insights: If you use predefined questions, you have consistency in how your data was collected and they are easier to analyze and compare the results. In addition, the structured format allows for the capture of quantitative information (e.g., number of breaches), and qualitative information (e.g., causes of breaches) in an efficient manner.
* Cost-Effective: When resource constraints are an issue, conducting surveys (especially online) is far simpler than, say, a case study or an interview.
* Relevance to the Study's Objective: Based on the research aim to identify common causes and patterns in data breaches, the survey directly entails the research aim. It allows the great real world insights from the people that actually have experienced it firsthand.
  + 1. **Time Horizons**

Discuss different research approaches and methodology for the Time Horizons layer. Provide justification of your choice.

**Time Horizons: Cross-Sectional**

In the research, a cross sectional time horizon is assumed which is based upon collecting information in one point in time. That being said, this method of data breach analysis is best suited to determine what is causing data breaches today.

**Justification:**

* Understanding Current Trends: A cross-sectional study captures immediate data breaches experienced by cybersecurity professionals, providing insight into contemporary challenges, attack patterns, and organizational vulnerabilities.
* Efficiency: Longitudinal studies suffer from long observation to correlate with present day and cross sectional data collection reduces time and focus on present day issues.
* Resource Constraints: Cross sectional surveys are less time consuming than longitudinal studies because the later requires time, manpower and capital to accomplish therefore practical for research that has limited capital and time to accomplish research projects.
* Alignment with Survey Methodology: Surveys fit well with cross-sectional research since they establish a point-in-time data point that the participants can discuss, which is the goal of the study since it seeks to identify the causes of data breaches.
* Focused Analysis: Concentrating on a specific timeframe enables research to create clear and definitive, and therefore actionable, results, which correspond to the state of play in the cybersecurity field at a particular point in time. It helps in finding action able patterns and recommendation without complexities of monitoring changes over time.
  + 1. **Techniques and Procedures**

Discuss different research approaches and methodology for the Techniques and Procedures layer.

Provide justification of your choice.

Techniques Used:

Survey Creation and Distribution:

Data breaches are gathered from surveys designed to tap directly into the experiences of cybersecurity professionals and subject matter experts.

It is distributed to platforms like social media and expert groups for very wide reach, and the responses are extremely varied.

Data Exploration with Power BI:

Why Power BI?

The collected data is explored with tools in Power BI to visualize trends, correlations and understand the reasons why data breaches happen. With its easy to use dashboard and interactive features, it’s simple to identify patterns, anomalies and relationships in the data.

Benefits:

It gives detailed charts and graphs to break down factors like geographical trends, common attack vectors and activities of threat actor.

It simplifies the identification of key issues in large datasets and brings immediate visual insights.

Modeling with Orange (Machine Learning):

Why Orange?

Orange is a powerful, user friendly data mining and machine learning tool, which, when applied to survey data analysis, allows one to discover the common causes of data breach. This allows for easy to use visual workflow, with robust machine learning tools for predictive analysis.

Process:

It cleans and pre processes the data.

Common causes of data breach are classified and predicted with various algorithms such as Random Forest, SVM, XGBoost.

Output:

Insights into what are likely causes are offered by models, which help to tell us the things that may trend, and what ways we can defend against them.

Justification of Choice:

Survey Creation and Distribution:

Surveys are considered to be efficient tools for collecting targeted, primary data from the right people, i.e., those with relevant experience.

This all enables fast and scalable data collection, with a wide audience of cybersecurity professionals.

Exploration with Power BI:

* Ease of Use: The visualization is simplified by Power BI, even the users who are not technical can use to visualize.
* Actionable Insights: The immediate identification of common themes through visualizations forms a basis for more in depth analysis.
* Customizable Reports: It can also be filtered and segmented based on attributes such as demographics, attack vector.

Modeling with Orange:

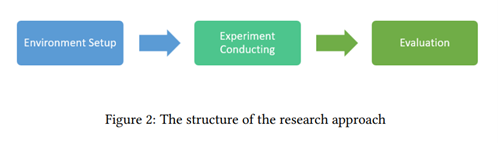
* User-Friendly Interface: Orange adopts drag & drop workflows that allow you to build a machine learning model with no extensive coding knowledge.
* Predictive Power: However, the paper goes beyond descriptive analysis and utilizes machine learning to devise predictions about patterns and causes of data breaches.
* Reproducibility: Orange models are also highly replicable, which means that future researchers can easily built upon this work.

Comprehensive Analysis:

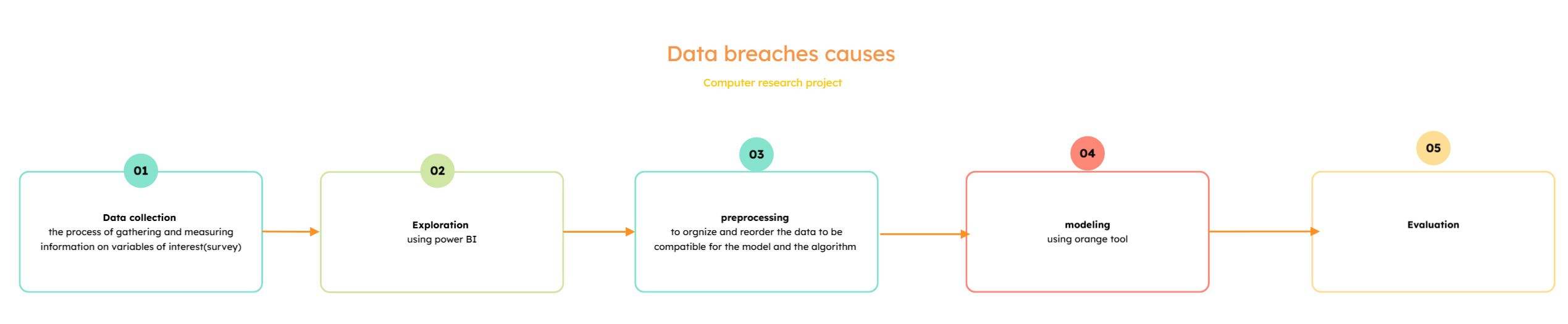
This approach utilizes survey data combined with the exploration techniques powered by Power BI and machine learning with Orange, thereby resulting in a structured and complete methodology. With this, you will get both descriptive and predictive insights on why data breaches exist.

### Research Methodology

Use a chart to show how your work is completed and discuss each part of the chart. An example of a chart is shown in Figure2.



Critically evaluate each part of the chart to justify chosen research methods and analysis.



**Figure (4) (here's the chart of the research methodology)**

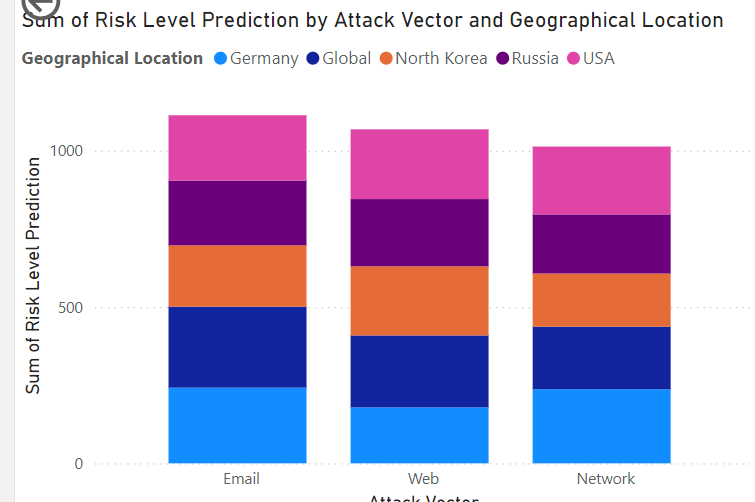
**Stage 1: Data collection**

Explained in section 3.

**Stage 2: exploration**

Using power Bi I choose the following charts:

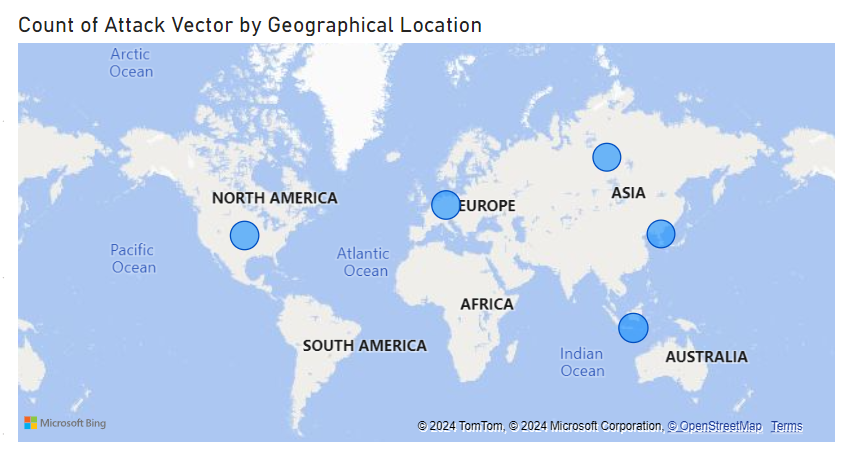
1\_stacked column chart



**Figure (5) (Sum of Risk Level Prediction by Attack Vector and Geographical Location)**

The attack vectors are plotted on the stacked column chart by location with the highest predicted risk levels. It enables to determine which locations are the most vulnerable and from which attack vectors they are exposed. The chart then breaks down the risk by categories, determining which areas are now critical and must be addressed, and prioritizing risk mitigation strategies for each vulnerability in each location.

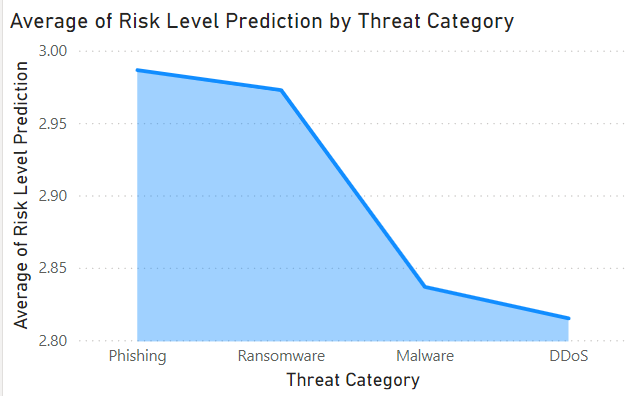
2\_Map



**Figure (6) (** **Count of Attack Vector by Geographical Location)**

The geographic map displays the sum of attack vectors for each location, providing a visual representation of the distribution of attacks across different regions. Additionally, it highlights the most prevalent attack vector in each location, enabling a deeper understanding of regional vulnerabilities and helping to focus security measures on the most common threats in each area.

3\_Stacked area chart

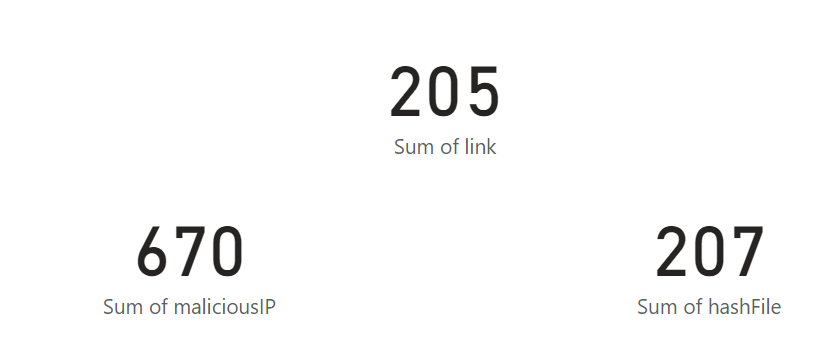


**Figure (7) (Average of Risk Level Prediction by Threat Category)**

The stacked area chart illustrates the average risk level predictions categorized by threat type. This

visualization helps identify which threat categories are associated with the highest average risk levels, offering insights into areas requiring prioritized attention for enhanced security measures.

4\_Card



**Figure (8) (sample of the cards and numerical values)**

cards can be used to quickly display the sum of different attack categories and the average severity score. The data would be put on each card so that we have more space and clarity on data for better emphasis. Here's the breakdown:

Scams: 274

Executable Files: 445

Vulnerability: 194

Malicious IP: 670

Phishing: 452

Network: 194

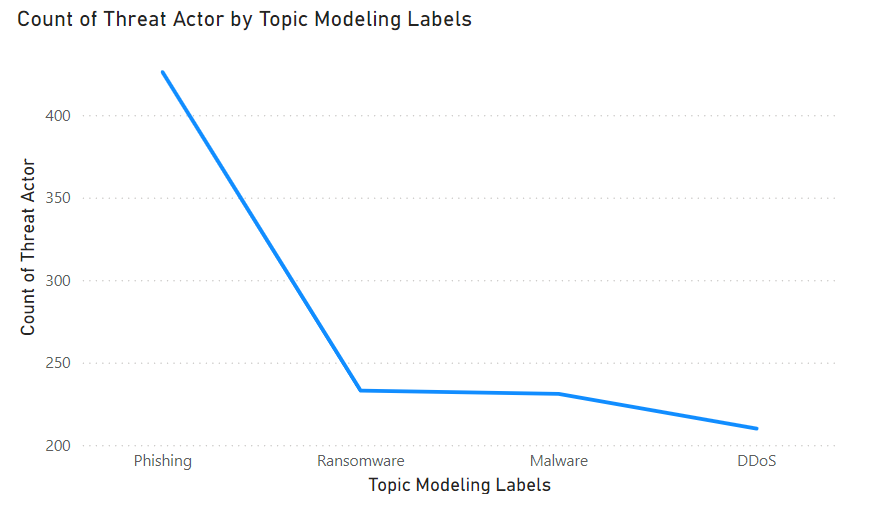
Average Severity Score: 2.99

Hash Files: 207

Link: 205

They give you an immediate sneak peek into the key data points so that you can do some focused analysis.).

5\_ line chart



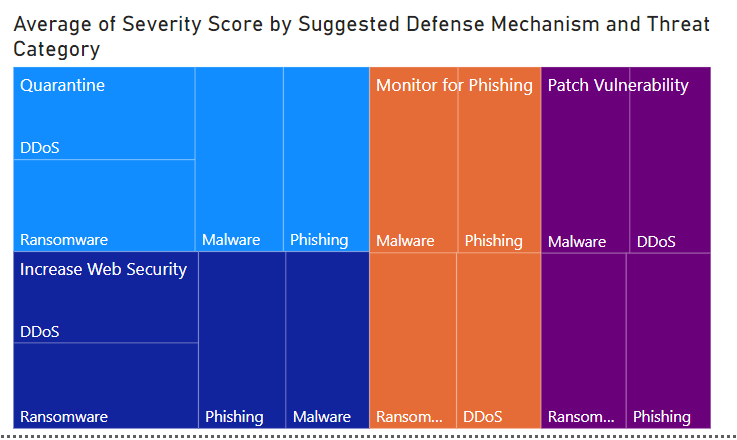
**Figure (9) (The figure shows the most targeted topic modeling label by the attackers)**

The topics targeted by threat actors will be visualized by count over time or across categories, respectively, using the line chart.

To understand which topics (i.e. vulnerabilities, phishing, malicious IPs) are the most common themes attacked by threat actors. Trends that will be shown include topics affected by a consistent or rapid increase in threat actor interest, facilitating in proactive measures.

A line chart is a great way to clearly see fluctuations and patterns, enabling you to spot the already most popular and least targeted topics to tweak security priorities as needed.

6\_treemap



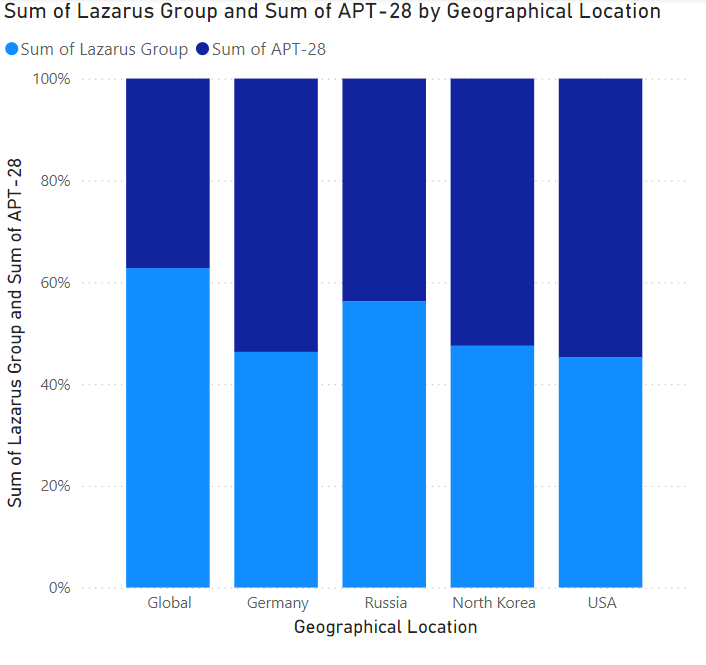
**Figure (10) (the figure shows the defense mechanism effectiveness)**

A detailed breakdown of the defense mechanisms impact on average severity score per threat category comparison will be provided in the treemap by threat category.

We want to visually explore how defense effectiveness is related to severity scores and threat categories. Per threat category, the frequency or importance will be represented by the size of each treemap block; color intensity will represent the average severity score. Through this visualization, one can understand which threat categories have higher risk levels, and whether defense mechanisms are working in reducing them.

Although not pretty, a treepmap is really good at showing relationships and letting you pick the top few areas to focus on improving their defense.

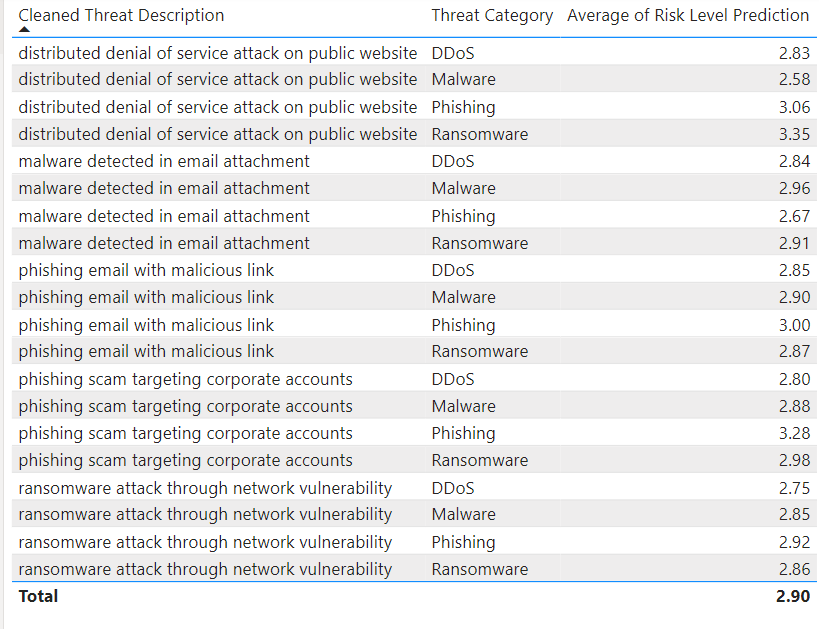
7\_Stacked bar chart



**Figure (11) (the chart pointed to the most geographic location get attacked)**

The sum of attacker groups’ attacks over different locations is visualized in the stacked bar chart. Each location is represented by a bar, and each segment of a bar is a contribution by an attacker group towards its location’s total attacks. This chart can help you identify hotspots that certain attacker groups have been targeting, and get an insight to regional patterns of attack so one can strategize a more focused defense along the way.

8\_Table



**Figure (12) (the table explained the attacks details the risk level prediction)**

However, the table shows the descriptions for the attack types for which the most average risk level predictions are shown in the table. Each row contains a risk level, the assigned threat category, and only a short explanation of what the attack entails to bind the risk level to the type of incident. This format allows easy comparison and understanding of how attack types contribute to elevated risk levels in other aspects of the EMI tool.

**Stage 3 preprocessing:**

**This is a method called Label Encoding** used to convert a text such categorical values with text into numerical codes. This assigns a unique integer to each unique value of a column. For instance, if a column contains the values ["Phishing", "DDoS", "Malware"], Label Encoding would assign: 0 for "DDoS" 1 for "Malware" and 2 for "Phishing". The idea is to make the categorical data machine processable while maintaining information.

**Application to How Label Encoding is implemented**

We first identified the columns of the dataset that needed encoding to start with. These consisted of "Threat Category", "Threat Actor", "Attack Vector", "Geographical Location", "Predicted Threat Category", "Suggested Defense Mechanism", and "Topic Modeling Labels". Each of these columns were transformed using Python’s LabelEncoder from the sklearn.preprocessing library. The encoder extracted all the unique (unique text on the column side) text values for each column, assigned an integer to each unique category, and replaced the original text values in the dataset with the corresponding numerical codes. For instance, "Threat Category" can contain a value of 'DDoS' which becomes 0, and 'Malware' becomes 1.

**Why Label Encoding Was Chosen**

Since our categorical features have a limited number of unique values, this was effective and simple enough to use label encoding. While in One-Hot Encoding we could have blow up the dimensionality of the dataset by adding a tons of new columns, Label Encoding supported the original structure of the dataset. This is especially so when the categorical data does not have an inherent order, which was the case here.

**How the Encoders Were Stored**

In order to future consistency and traceability, the label encoders were stored in a single dictionary. It makes it easy to decode the numerical values in to their proper categories if that is required. For example, if we take the saved encoder for "Threat Category": [0, 1, 2] we could transform it back into ["DDoS", "Phishing", "Malware"]. Moreover, saved encoders are applicable if the model is used later on something new or unseen data.

**Outcome and Benefits**

The best thing about that is that the final transformed dataset includes only numerical values also dropping the textual columns such as attacks description and this is perfect for machine learning models and the targeted column (which is threat castigatory) make it categorical value to avoid orange recognize it as aggression model. Now the dataset is fully compatible with such algorithms that require numerical datasets since we encode categorical data in this way. This serves two purposes: model compatibility, which in theory can improve the model or its accuracy, and makes the data preprocessing a simpler task.

**Stage 4: Modeling**

Machine learning algorithms are used to analyze the collected data and find patterns in the data, in order to predict the most common data breach causes. In this research, three machine learning models are employed: In the other words, we use XGBoost, Random Forest and Support Vector Machine (SVM). Reasons for that are these algorithms could handle complex datasets and help identify key factors that lead to the data breaches.

**1. Extreme Gradient Boosting(XGBoost)**

The algorithm of XGBoost is an ensemble learning algorithm based on decision trees. Gradient boosting techniques were used for optimization to minimize error. It generally does well with handling missing values.

It is accurate and performing, especially in structured data.

It offers feature importance metrics which highlights the factors that contribute the most to data breaches.

**2. Random Forest**

Beyond voting controlled ensemble, another ensemble learning method is Random Forest, which builds multiple decision trees to make prediction. It is simply the output of these trees combined together to improve general accuracy and robustness. It does a great job with higher dimensionally large data sets.

It reduces the risk of over predicting because it is a randomized nature.

Also offers explain ability ranking the importance of features related to data breaches.

3. **Support Vector Machine (SVM)**

SVM is a supervised learning algorithm for classification task. What it does is by identifying the hyperplane that will best separate given data points into different categories. It works well on linear as well as nonlinear datasets.

Common in cybersecurity research, and well suited to smaller or imbalanced datasets.

It produces robust results despite high variability in the data.

4. **AdaBoost (Adaptive Boosting)**

AdaBoost can be thought as the combination of multiple weak classifiers (e.g. decision stumps) to create one strong classifier. Rather it gives higher weights to the harder to predict data points during training. It focuses on misclassified instances to improve accuracy.

It works well with the structured datasets, and also offers the generalization with better precision. It is computationally simple and appropriate for use in iterative research processes.

Modeling Process:

At this stage, Machine learning algorithms are applied to the prepared dataset, in order to discover the typical causes of data breaches. Random Forest, XGBoost, SVM, and AdaBoost were selected for the algorithms because they are capable of classifying problems.

**Algorithms and Justifications**

* Random Forest: Since it is robust to handling noisy datasets and captures complex relationships. It does well with categorical and numerical fields.
* XGBoost: Its efficiency and effectiveness in handling large datasets, with built in feature selection, was selected. Regularization lends high accuracy [and] does not let the model overfit.
* SVM (Support Vector Machine): It is preferred for its ability to perform well in high dimensional spaces, for its capability to classify data effectively. The kernel trick allows us to model boundaries that are too complex.
* AdaBoost: This is included for its ability to comprised of iterative combination of weak learners to improve classification performance. It is useful to deal with imbalanced class.

**Cleaning and Processing justification:**

The combination of which algorithms are suitable is the most important point, which means that preprocessing and cleaning steps are vital to achieve this compatibility. For SVM and tree based models to work optimally, both categorical variables needed to be encoded and numbering any numerical variables needed to be standardized. Missing data was addressed to avoid any valuable information being lost and validated dataset to be used safely for making true predictions. It helps to make the models more reliable in terms of recognizing data breach causes.

**Stage 5: evaluation**(indicators of the model)

1\_ AUC (Area Under the Curve):

What it is: Receiver Operating Characteristic (ROC) curve is the area under the curve. ROC curve shows the relationship (true positive rate) vs (false positive rate) at different threshold levels.

Purpose: A model is assessed for discriminating those classes.

Range:

AUC = 1: Perfect classifier.

AUC = 0.5: Random guess.

AUC < 0.5: Worse than random guessing.

Why it matters: AUC values that are higher indicate that the model is more adept at giving positive cases a higher ranking than negative ones.

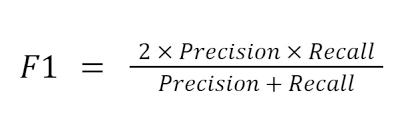
2\_ F1-Score:

What it is: Precision \* Recall / 2 It balances both metrics.

Purpose: This is useful when your dataset has an imbalance and you’re as concerned about the precision as you are about the recall.

Why it matters: It punishes very unbalanced problems (e.g high Precision and low Recall or vice versa).

Formula:



**Figure (13) (F1 formula)**

3\_ CA (Classification Accuracy):

What it is: It's the ratio of the correct labelled samples to total number of samples.

Where:

TP = True Positives.

TN = True Negatives.

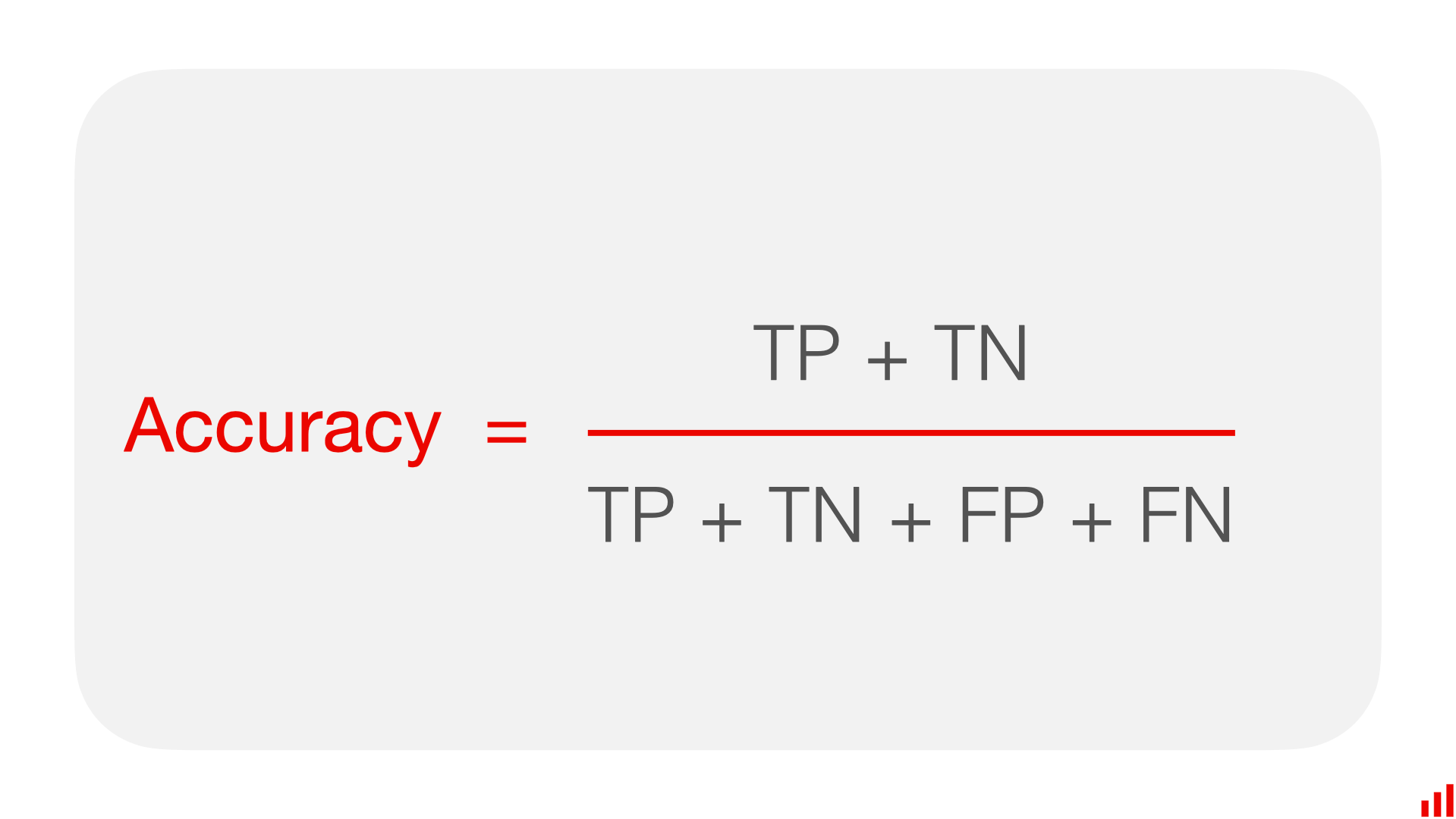
FP = False Positives.

FN = False Negatives.

Purpose: It measures overall correctness of the model.

Why it matters: But relatively simple and intuitive, and can potentially be misleading if the dataset is imbalanced (i.e., high accuracy but poor performance on minorities class).

Formula:



**Figure (14) (Accuracy formula)**

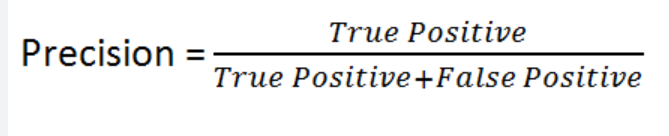
4\_precision:

What it is: Measures how many predicted positive cases are actually positive.

Purpose: Focuses on the relevance of positive predictions.

Why it matters: Important in situations where false positives are costly (e.g., spam detection, fraud prevention).

Formula:



**Figure (15) (precision formula)**

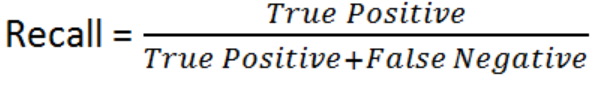
5\_Recall:

What it is: It tells us how many actual positive cases have been identified and we are correct in doing so.

Purpose: It mainly records all positive cases.

Why it matters: Also needed if false negatives have severe consequences (e.g. malicious threats in medical tests or security applications, etc.).

Formula:



**Figure (16) (recall formula)**

6\_MCC (Matthews Correlation Coefficient):

What it is: It is a balanced metric taking all confusion matrix values: TP, TN, FP, FN.

Purpose: It provides a single score for binary classification even for imbalanced dataset.

Why it matters:

MCC = +1: Perfect prediction.

MCC = 0: Random prediction.

MCC = -1: Complete disharmony between actual and prediction.

# Results and Discussion

Discuss the results. Include all tables and figures.

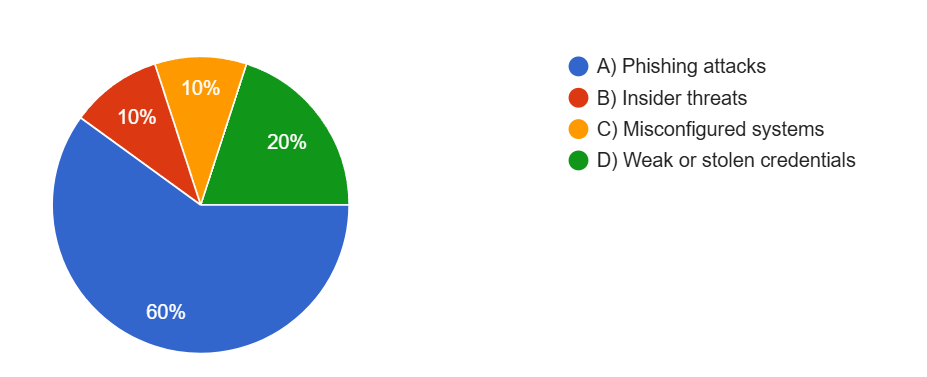
Explain how the results meet the research question and objectives.

Describe merits and limits of the analysis.

The result of the primary data (survey for the cyber security experts about the reasons of the data breaches):

Note that the survey got 10 responds from the expertise

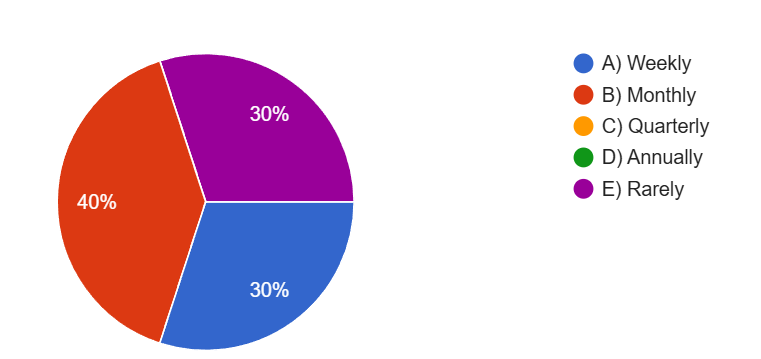
**Q1:**



**Figure (17) (In your professional experience, what is the most common cause of data breaches in the industries you specialize in?)**

Figure [17] highlight the most common causes of data breaches in specific industries based on professional experiences. The majority of respondents, 60%, identified phishing attacks as the primary cause, underscoring their prevalence as a significant cybersecurity threat. Weak or stolen credentials were the second most common cause, accounting for 20% of responses, reflecting the critical importance of robust authentication practices. Insider threats and misconfigured systems each contributed to 10% of breaches, emphasizing the need for proper system configurations and effective internal security measures. These findings demonstrate that human and technical vulnerabilities remain key factors in data breaches.

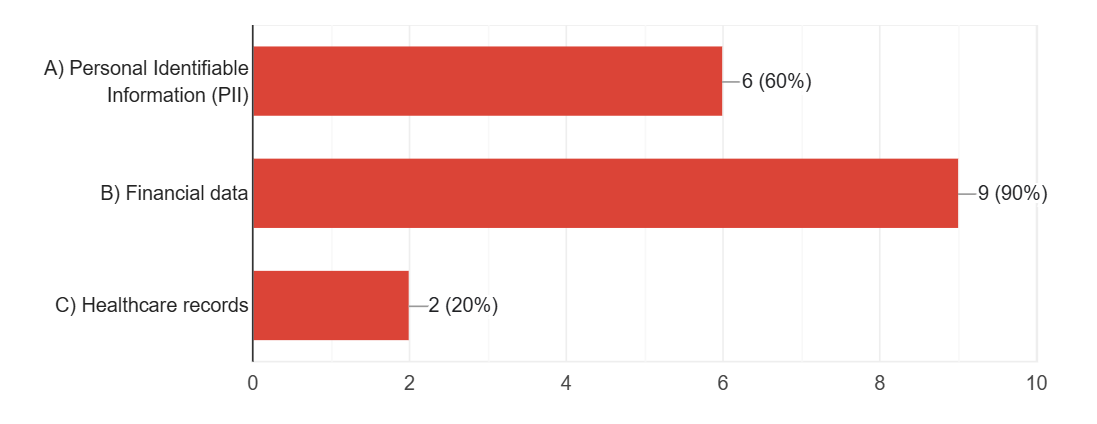
**Q2:**

****

**Figure (18) (How frequently do organizations report data breaches?)**

Figure (18)provide insight into how frequently organizations report data breaches. The majority of respondents (40%) indicated that data breaches are reported on a monthly basis, suggesting a structured approach to addressing such incidents. Interestingly, 30% of respondents noted that breaches are reported weekly, reflecting a need for more immediate response mechanisms in certain organizations. Another 30% indicated that breaches are rarely reported, highlighting a possible lack of transparency or reporting protocols in some cases. Notably, there were no responses for quarterly or annual reporting, suggesting that these intervals may be less practical or less prioritized for organizations managing data breaches. These findings emphasize the varying strategies organizations employ when dealing with data breach incidents.

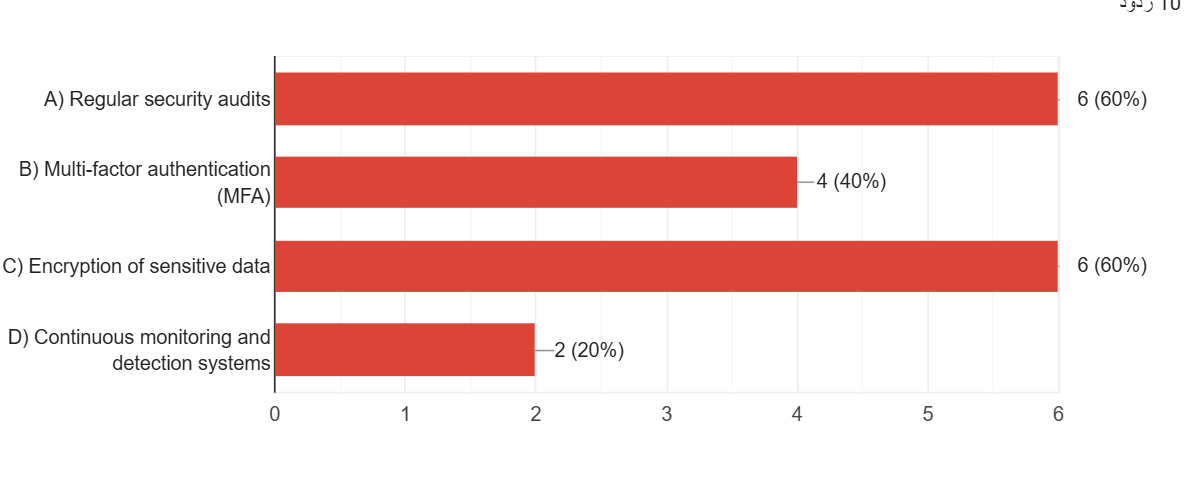
**Q3:**

****

**Figure (19) (What type of data is most commonly targeted in breaches you’ve investigated?)**

Figure [19] indicate what data they most commonly investigate in breaches themselves. Financial data is the most common target of breaches, with 90 per cent of respondents declaring it was a major focus. The second most targeted category (60 percent of respondents) is Personal Identifiable Information (PII), which clearly has a high value to bad actors. Leaving healthcare records under further attack — while still accounting for 20 per cent of breaches — the fact that they are a key aspect of the picture in certain situations underscores their relevance. This shows that attackers consider financial data and PII as prime targets that need more secure protection against data leakage and makes it clear why incredibly robust security measures are required to protect these highly sensitive data types.

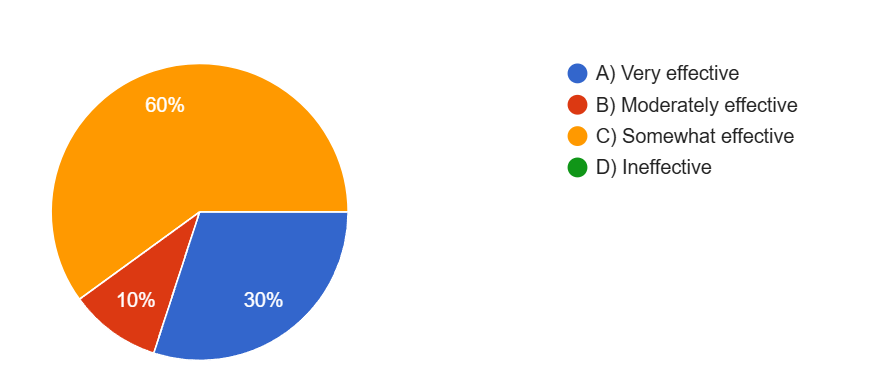
**Q4:**



**Figure (20) (**Which**security measure do you find most effective in preventing data breaches?)**

According to Figure [20], the top two most effective ways to keep data breaches from happening, according to the respondents (6 out of 10) were 'Regular security audits' and 'Encryption of sensitive data', which each received 60 percent of the votes (6 out of 10 respondents). "Multi factor authentication, (MFA)" also 40%, "Continuous monitoring and detection system," also 20%. The survey yielded 10 responses.

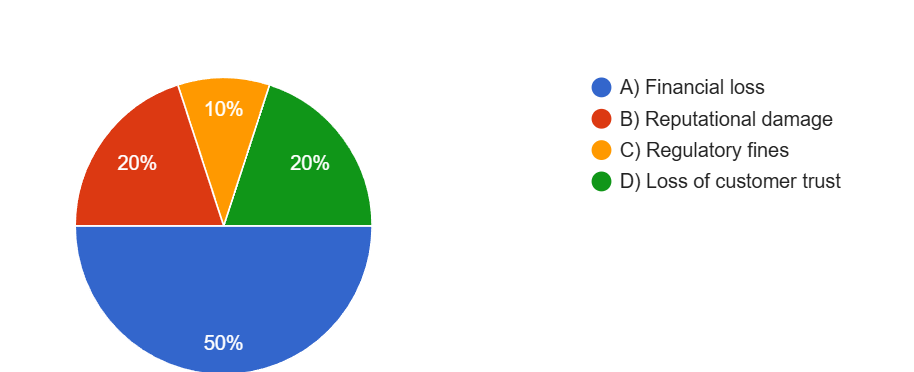
**Q5:**



**Figure (21) (How effective are organizations at identifying breaches in their early stages?)**

According to the Figure [21], 60% of respondents believe organizations are "Somewhat effective" at identifying breaches in their early stages. Meanwhile, 30% think organizations are "Very effective," and 10% view them as "Moderately effective." None of the respondents rated organizations as "Ineffective." A total of 10 responses were recorded in the survey.

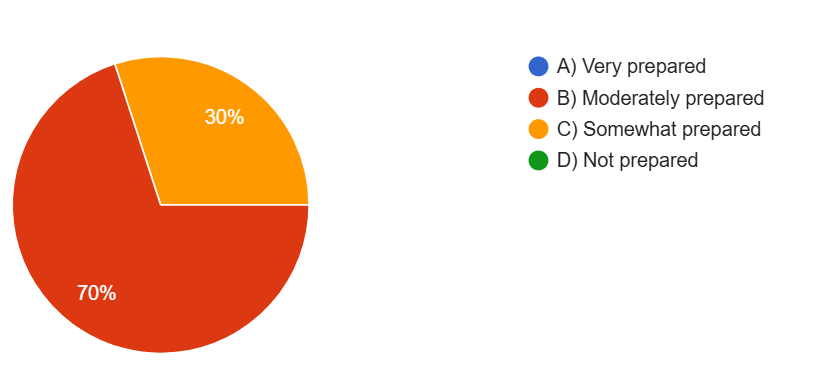
**Q6:**

****

**Figure (22) (What do you observe as the biggest consequence of data breaches for organizations?)**

The chart shows that the biggest consequence of data breaches for organizations, as perceived by respondents, is "Financial loss," which was selected by 50% of respondents. "Reputational damage" and "Loss of customer trust" were each identified by 20% of respondents. Meanwhile, "Regulatory fines" were considered the biggest consequence by 10% of respondents. A total of 10 responses were recorded in the survey.

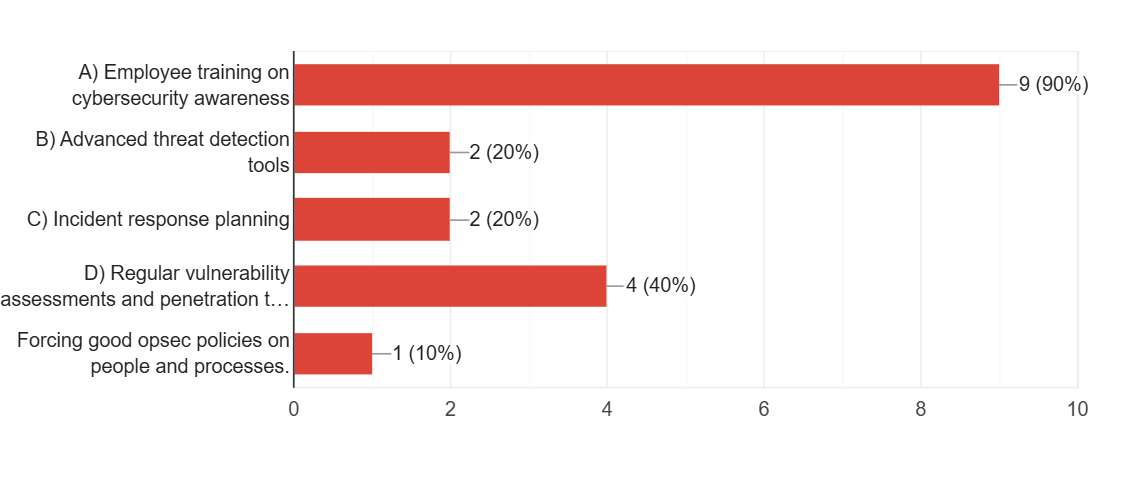
**Q7:**

****

**Figure (23) (How prepared are organizations to respond effectively to a major data breach?)**

Organizations are illustrated as to how ready they are to tackle things in a hand in glove manner in the event of a huge data breach. 70% of participants responded that they are 'Moderately prepared', and 30% responded 'Somewhat prepared'. None was recorded in "Very prepared" or "Not prepared."

**Q8:**



**Figure (24) (What area should organizations invest in most to reduce the risk of data breaches?)**

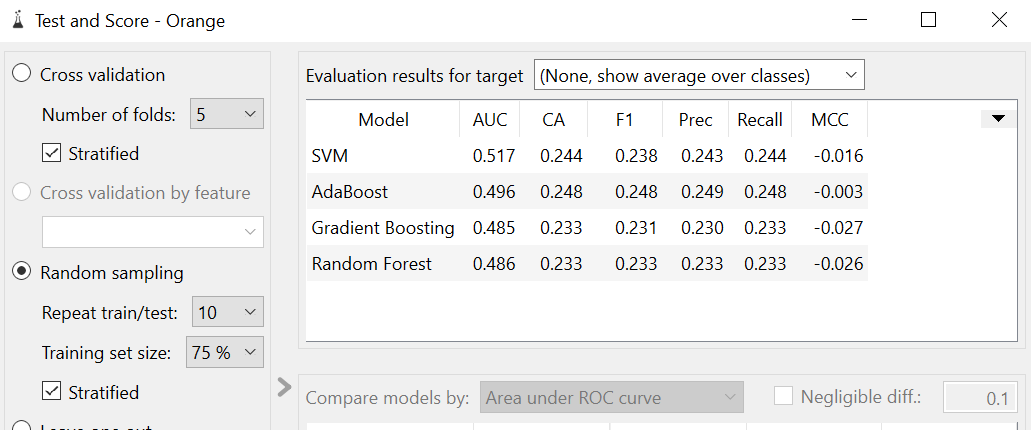
Figure [24] illustrates the areas organizations should prioritize investing in to reduce the risk of data breaches. The majority, 90%, selected "Employee training on cybersecurity awareness." Other responses included "Regular vulnerability assessments and penetration testing" at 40%, "Advanced threat detection tools" and "Incident response planning," both at 20%. Finally, "Forcing good operational security (opsec) policies on people and processes" was chosen by 10% of respondents.

**Final result of the primary data:**

While data breaches present a challenge for organizations, phishing attacks rank number one as the chief cause at 60%. 20 percent of the breaches take place because of either weak or stolen credentials, and for that reason you need to have robust practices for authentication. It is surprising that 10% of all breaches are caused by insider threats and misconfigured systems, which requires conservative measures to be taken internally and properly configured systems. Organizations report data breach at different frequencies, 40% report daily, 30% weekly and 30% rarely. The most prominent data loss targets are financial data which is followed by something called Personally Identifiable Information, or PII, along with healthcare records. Organizations have in place systems for regular security audits, encrypting sensitive data, enabling multi factor authentication, and ongoing monitoring, detection, and response. Early breach detection effectiveness ranges, with 60 percent respondents believing that organizations are somewhat effective, 30 percent very effective, and 10 percent moderately effective. Financial loss, reputational damage, loss of customer trust, and just plain old regulatory fines. Organization are moderately prepared for major data breaches with 90% prioritizing employee training on cybersecurity awareness. Other areas which are given priority are to perform regular vulnerability assessments, penetration tests, advanced threat detection tools, incident response planning and enforce strong operational security policy.

Secondary data result (using data preprocessing and modeling then visualize it):

Orange tool:



**Figure (25) (The general accuracy comparison of the algorithms)**

Below is summarized the evaluation results for the target, as seen in the image. Table compares performance of four machine learning models—SVM, AdaBoost, Gradient Boosting, Random Forest and with respect to four different metrics—AUC, CA, F1, Precision, Recall, MCC.

In the end SVM had an AUC of 0.517, CA of 0.244, F1 score of 0.238, Precision of 0.243, Recall of 0.244 and MCC of -0.016.

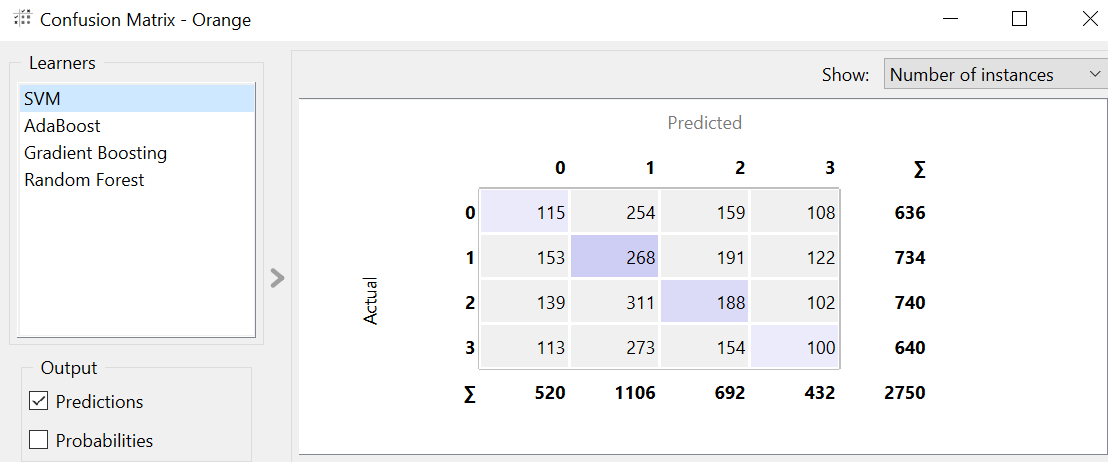
The AUC from AdaBoost was 0.496, CA of 0.248, F1 score of 0.248, Precision of 0.249, Recall of 0.248 and MCC of -0.003.

The AUC of Gradient Boosting was 0.485, CA 0.233, F1 score 0.231, Precision 0.230, Recall 0.233, and MCC was -0.027.

The AUC, CA, F1 score, Precision, Recall, and MCC were respectively 0.486, 0.233, 0.233, 0.233, and -0.026 in Random Forest respectively.

Although all models have overall performance metrics which are relatively close and can be improved, the SVM model has the highest AUC value (0.517).

SVM (confusion matrix):



**Figure (26) (confusion matrix of SVM)**

By using the confusion matrix for the SVM model to summarize the predicted versus actual classifications for four classes (0, 1, 2, 3), we can easily explain the differences in the loss functions for the same input. Here is the interpretation of the numbers:

**Key Numbers:**

Diagonal values (correct predictions): These instances were correctly classified by the model.

Class 0: 115 instances.

Class 1: 268 instances.

Class 2: 188 instances.

Class 3: 100 instances.

Off-diagonal values (misclassifications): The other shows how often the model classified instances erroneously.

Row totals (actual instances): Number of instances for each actual class.

Class 0: 636 instances.

Class 1: 734 instances.

Class 2: 740 instances.

Class 3: 640 instances.

Column totals (predicted instances): Number of instances for each classes that are predicted.

Predicted as Class 0: 520 instances.

Predicted as Class 1: 1,106 instances.

Predicted as Class 2: 692 instances.

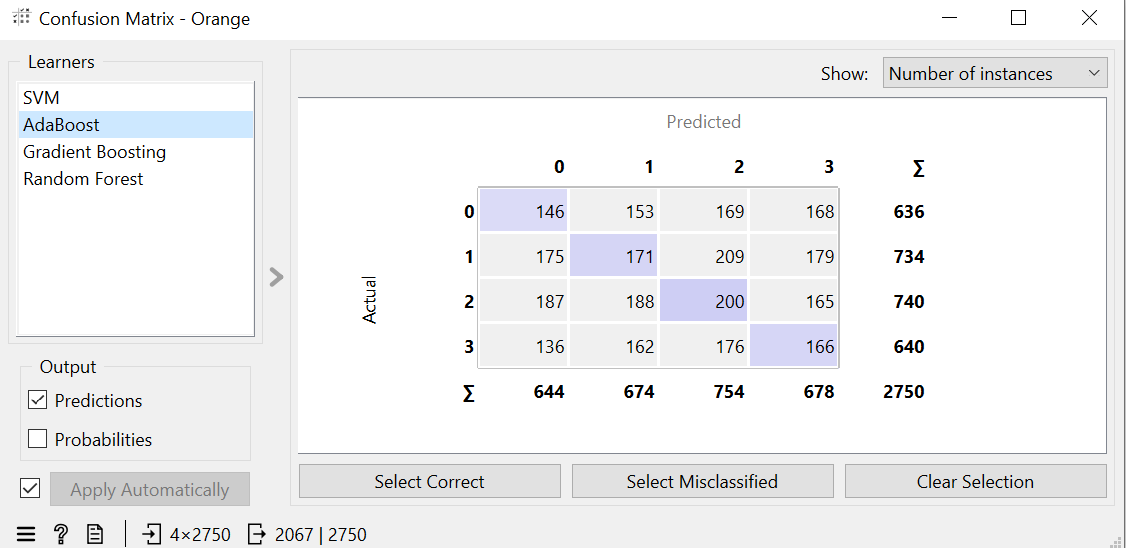
Predicted as Class 3: 432 instances.

Grand total: There are 2,750 overall instances.

The diagonal values represent instance classifications correctly by the model and off-diagonal values reflect where the model misclassified instances.

All classes are seen to have noticeable misclassifications; as most predicted instances differ from the actual class. This supports the fact that the SVM doesn’t work that well classifying all categories as our distribution demonstrates.

Adaboost (confusion matrix):



**Figure (27) (confusion matrix of AdaBoost)**

The AdaBoost model confusion matrix gives us the distribution of predicted versus actual classifications on four classes (0, 1, 2 and 3). Below is the interpretation of the numbers:

**Key Numbers:**

Diagonal values (correct predictions): They show how many instances we correctly classified.

Class 0: 146 instances.

Class 1: 171 instances.

Class 2: 200 instances.

Class 3: 166 instances.

Off-diagonal values (misclassifications): These are the number of such instances that were misclassified to other classes.

Row totals (actual instances): Per Actual class number of total instances.

Class 0: 636 instances.

Class 1: 734 instances.

Class 2: 740 instances.

Class 3: 640 instances.

Column totals (predicted instances): Prediction of total instances for each class:

Predicted as Class 0: 644 instances.

Predicted as Class 1: 674 instances.

Predicted as Class 2: 754 instances.

Predicted as Class 3: 678 instances.

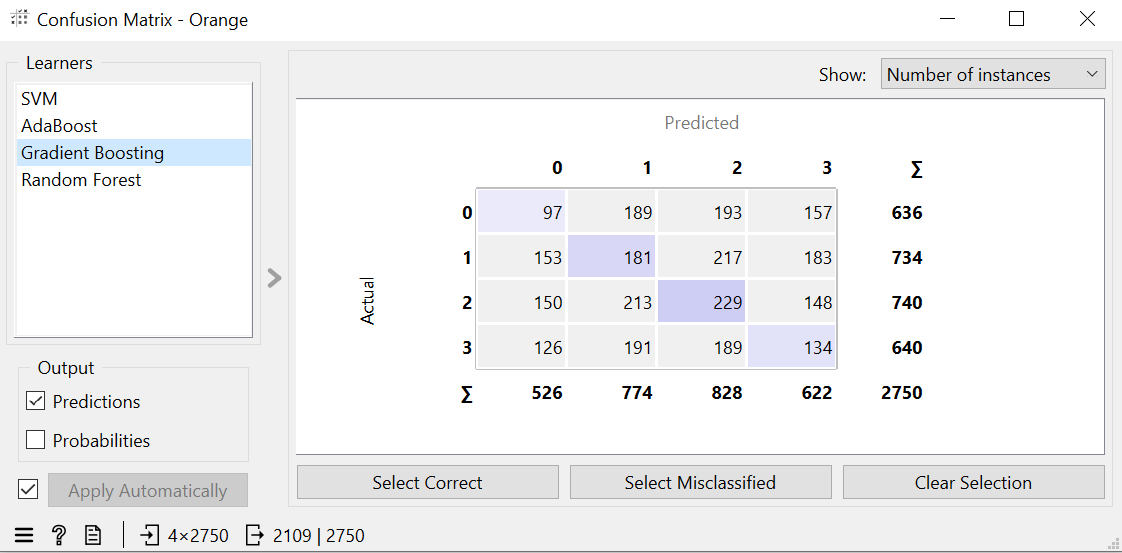
Grand total: There are 2,750 total instances.

It correctly classified 146, 171, 200, and 166 instances for classes 0, 1, 2 and 3, respectively.

It is clear that there are significant misclassifications; many are incorrectly predicted for other classes, meaning the model's performance is spread unevenly over the classes.

This is a relatively balanced dataset, but the class totals they are predicting with a high degree of accuracy and the class totals that actually occur show problems with accurate classification.

xgboost (confusion matrix):



**Figure (28) (confusion matrix of xgboost)**

The Four Classes (0, 1, 2, and 3) predicted vs estimated by the Gradient Boosting Model is summarized in the confusion matrix. Below is the interpretation of the numbers:

**Key Numbers:**

Diagonal values (correct predictions): These are the number of instances correctly classified.

Class 0: 97 instances.

Class 1: 181 instances.

Class 2: 229 instances.

Class 3: 134 instances.

Off-diagonal values (misclassifications): These give numbers of instances that were classified as non-other class.

Row totals (actual instances): Per actual class:

Class 0: 636 instances.

Class 1: 734 instances.

Class 2: 740 instances.

Class 3: 640 instances.

Column totals (predicted instances): Each class has been predicted total instances.

Predicted as Class 0: 526 instances.

Predicted as Class 1: 774 instances.

Predicted as Class 2: 828 instances.

Predicted as Class 3: 622 instances.

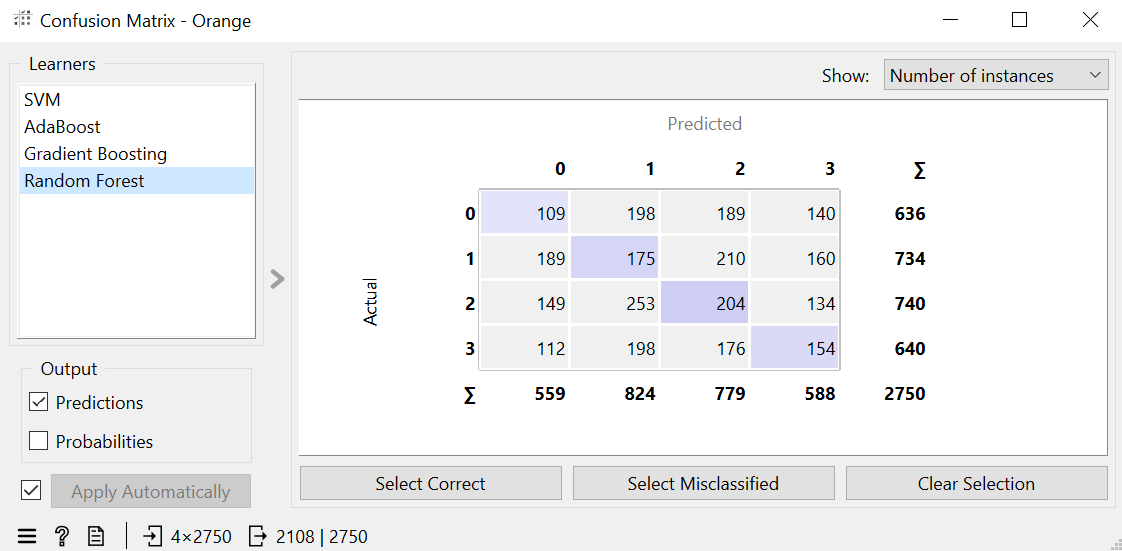
Grand total: There are 2,750 instances in total.

For classes 0, 1, 2, and 3, Gradient Boosting successfully classified 97, 181, 229, and 134 instances respectively.

Not only did we find misclassification topping, but class 0 and class 1 were often misclassified to other categories.

Some bias in the predictions can be seen by using the predicted class totals and actual class totals for the predicted class totals indicate higher class 2 predicted class totals than other classes (e.g. class 2). The model’s performance suggests that it is difficult to accurately differentiate the four classes.

Random Forest (confusion matrix):



**Figure (29) (confusion matrix of Random Forest)**

When it was compiled on four classes (0,1,2,3) the Random Forest model confusion matrix gave this. Below is the interpretation of the numbers:

**Key Numbers:**

Diagonal values (correct predictions): The following are the instances correctly classified.

Class 0: 109 instances.

Class 1: 175 instances.

Class 2: 204 instances.

Class 3: 154 instances.

Off-diagonal values (misclassifications): For example, these are misclassified instances into other classes.

Row totals (actual instances): Number of actually seen instances in each class:

Class 0: 636 instances.

Class 1: 734 instances.

Class 2: 740 instances.

Class 3: 640 instances.

Column totals (predicted instances): Number of predicted instances for each class.

Predicted as Class 0: 559 instances.

Predicted as Class 1: 824 instances.

Predicted as Class 2: 779 instances.

Predicted as Class 3: 588 instances.

Grand total: The data set has 2,750 instances in total.

The Random Forest model correctly classified:

109/636 instances of Class 0.

175/734 instances of Class 1.

204/740 instances of Class 2.

154/640 instances of Class 3.

It produces lots of misclassification in all classes, especially for neighboring classes.

Overall, the model does poorly with some distinction, with Class 1 and Class 2 having additional instances predicted as neighbors classes.

Power BI:

Explained in the previous section (research method).

Results in Relationship to Research Questions and Objectives

**Explanation**

This research aimed at explaining causes, breach reporting practices and breach impact from a breach data analysis, identifying effective breach mitigation strategies, and evaluating machine learning models for predicting breach risk. The findings correspond with the goals of the research, elucidating essential information about data breaches and areas for enhancement.

**Primary Data Analysis:**

The survey findings provide valuable insights:

Causes of Data Breaches: Phishing attacks (60%, weak credentials (20%) were the leading cause, and since robust authentication and employee training was a necessity.

Reporting Practices: This has resulted in the variance in breach reporting frequency, 40% of companies reporting monthly, 30% reporting weekly or rarely, which relies on varying organizational approaches, and potentially broken reporting mechanisms.

Impact Analysis: Financial data and PII were the prime targets, with financial loss at number 1 (50%) It is exactly the motivation that makes protecting sensitive data so crucial.

Mitigation Strategies: Plans to perform regular audits, encryption, MFA, and continuous monitoring were deemed effective as this is what is seen as Best Practises for a company's cybersecurity.

Preparedness: Although 70% of respondents think organizations are 'moderately prepared' for breaches, it also shows there is still work to do to beef up incident response readiness.

**Secondary Data Analysis:**

We evaluated machine learning models using AUC, F1 score and MCC. The results revealed limitations in model performance:

SVM Model: Although their results had the highest AUC (0.517), their overall classification accuracy was low, due to high misclassification.

Other Models (AdaBoost, Gradient Boosting, Random Forest): Models maintained a similar, but suboptimal performance when predicting breach scenario.

Confusion Matrices: Results across all classes suggest that better data preprocessing and feature selection is needed in order to improve predictive accuracy.

Analysis Merits and Limitations

**Merits:**

Comprehensive Insights: This survey described the causes of and impact from data breaches in detailed manner, and provides actionable recommendations for the organizations.

Integration of Machine Learning: Predictive models evaluation give a forward looking aspect about the potential use of technologies in cybersecurity.

Mixed-Method Approach: A combination of qualitative and quantitative analysis verifies a full encounter of the research questions.

**Limitations:**

Sample Size: Findings are not generalizable due to the surveyed responses of people (10 participants). More robust insights can be yielded from a larger sample.

Machine Learning Model Performance: All models show a low predictive accuracy which implies that data preprocessing, feature engineering, and model selection need to be improved further.

Focus on Select Metrics: Specific performance metrics used (e.g. AUC, MCC) may not be what really matters in practical cases.

Lack of Real-World Testing: Although the findings were not validated against real world breach data, their practical applicability in machine learning was limited.

# Conclusion and Recommendations

State the aim of the study and summarize the findings.

Based on the findings, discuss the recommendations that you have.

Propose future work (optional).

**Aim of the Study:**

The purpose of this study was to examine the causes, frequency, and implications of data breaches, and to uncover the data most often accessed in those breaches as well as to evaluate the efficacy of existing data protection mechanisms. Additionally, it aimed to examine the level to which organizations were prepared to reduce and counter data breaches through taking into consideration both primary and secondary data for an analysis of patterns and trends in organizational cybersecurity practice.

**Summary of Findings:**

It found the biggest cause behind data breaches will be phishing attacks — with 60 percent of incidents, followed by weak or stolen credentials, at 20 percent. 10% caused insider threats and 10% caused misconfigured systems. Of the respondents, 40% reported breaches monthly, 30% weekly, and 30% reported them rarely. The big three types of ‘at risk’ information, in order according to the numbers supplied by respondents, were financial data at 90%, Personally Identifiable Information (PII) at 60%, and healthcare records at 20%.

The most effective security audit and encryption were considered when making security measure evaluation, 60% of respondents agreed. Respondents were also asked about MFA and continuous monitoring, with 40 percent supporting MFA and 20 percent supporting continuous monitoring. About 60 percent of participants thought organizations were “somewhat effective” at early detection, while 30 percent thought that they were very effective. According to 50 per cent of respondents, the majority of consequences were financial losses resulting from data breaches, with 20 per cent citing reputational damage and loss of customer trust in each case and 10 per cent reporting regulatory fines.

In this discussion, only 30% considered organizations it had spoken with to be ‘somewhat prepared’ whereas 70% rated organizations as ‘moderately prepared’. Timely corporate cybersecurity awareness training for employees ranked as its top investment priority and was emphasized by 90% of respondents. Regular vulnerability assessments (40%), advanced threat detection tools (20%) and incident response planning (20%) were also other priorities. Thirdly, results from the analysis of machine learning model performance showed that Support Vector Machines (SVM) had high area under curve (AUC) value (0.517) but overall accuracy was horrible having many misclassifications across all models.

**Recommendations:**

To help address human induced breaches, organizations should prioritize employee training on preventative of phishing, credential management and cybersecurity best practice. To tackle weak passwords or stolen credentials, stronger authentication measures, including well implemented robust multi factor authentication (MFA) and frequent password audits are paramount. Security practices have to be advanced, including things like mandatory regular security audits and vulnerability assessments. In addition, increased use of advanced threat detection tools plus proper system configurations can substantially decrease the odds of breaches.

It’s important to improve incident response protocols. Regularly setting up and testing incident response plans will make organizations better prepared and quick to contain breaches. It is also important to encourage transparency in breach reporting, timely and transparent communication with the regulatory bodies and affected stakeholders can contribute to maintaining accountability and building trust. As financial data and PII are most protected by the attackers, they should be given encryption, masking of data and strict access controls. Finally, we need to support the AI investments to improve machine learning models using better accuracy and effectiveness. In other words, we can lower the misclassification rates by refinding data preprocessing methods and exploring ensemble learning methods.

**Future work:**

More work in this space should be focused around building predictive tools based on AI capabilities in order to predict and prevent data breaches based on behavioral patterns and threat indicators. One could dedicate a study to small and medium enterprises (SMEs) to glean the kind of unique challenges that small and medium enterprises engage in when confronting cybersecurity threats in order to make suggestions more specific. Analysis of data breach patterns over a long time frame for longitudinal evidence could identify and help learn about evolving threats and emerging attack vectors. Furthermore, combining behavioral analytics with traditional security techniques may contribute to insight around the reduction of insider threats, as well as anomaly detection.

# Reflections

Avoid generalization and focus on personal development and the research journey in a critical and objective way.

However, this line of research has been education and beneficial in general as part of personal and professional growth. The process went to various phases that are peculiarities and lessons involved which gave me realization of the research process.

At first the key challenge was noted to be in the survey component of the work. While developing and administering survey, attention was paid to what kind of data is relevant and sufficient which has to be collected. Some aversive incidents that occurred in relation to the people were the following The issue of attracting the people’s response and, in fact, ensuring that the people understood the presented questions. At this stage I found how operationally the term clarity is used in survey and also about the patience and tolerance level while talking about response rate and missing data.

The preprocessing stage was also an area of difficulty in the areas of problem-solving. The very selection of methods to be applied for preparing the data, for example label encoding, eliminated the possibility of considering benefits and drawbacks of different techniques. For instance, while label encoding preserved the dataset’s original structure, decisions about dropping columns such as textual descriptions posed a dilemma: as knowing the right way to approach the issue of feature selection and at the same time possibly discarding some relevant information. The fact that the need to match data with machine learning algorithms was considered more critical than the quality of data also the challenge in this stage is to getting acceptable accuracy indicators for the model.

Major issues encountered at this stage of modeling included the inability to find high accuracy algorithms that fit the type of dataset used as well as picking algorithms that were appropriate for the study. When models were tried out in the field and their success was compared with the ideal the suitability of models for certain applications was discovered to be not so easily determined. Even after the very thorough preprocessing of the dataset in different aspects, the performance of the model was not as good as what I expected originally. This compelled me to explore further various aspects related to models such as features selection, tuning of parameters and constraint of datasets. While the lower accuracy was a disappointment, it provided an important lesson: when getting to the results, research is not about obtaining the best outcomes but about learning the outcomes.

Thus, reviewing the entire methodology of this research, it is also important to note that this has been, in equal measure, a journey of personal growth as well as of technical evolution. I improved and refined my data gathering, cleaning and analysis, and modeling techniques while getting a better sense of research as cycle. The difficulties I experienced on the job underlined such values as the ability to learn, creativity, and conscientiousness. Furthermore, it also introduced the understanding that, in fact, each failure can be a way to improve modes of work and increase one’s knowledge about related matters.

In the future, I will continue to enhance this experience by studying other techniques or technologies applying to the increment of model accuracy and other aspects regarding algorithm choice, and model validation. This reflection confirms the cyclical and complex process of doing research, stating that each and any difficulty faced in the process of the research contributes to development and advancement.

* 1. **Selected Research Methodology**

Reflection of the research process.

Reflection on the merits, limitations, and potential pitfalls of the selected methods.

The method employed for this study was selected considering the research aims and objectives as well as the nature of the data set and the scalability of the research. This involved survey data collection, data preprocessing, modeling, and the last phase of evaluation. All the stages used appropriate methods and instruments and shared the best practices balanced with the references to the theoretical background.

In data collection phase, surveys were used to get raw data from respondents about cyber threats, threat actors, threats delivery methods and protective measures. They applied such techniques as label encoding whereby categorical information is translated into numerical type so as to conform with the models of machine learning. Modeling entailed practicing on different algorithms and algorithms testing to see which was best for the dataset , while evaluation tested the model and its ability to be resilient.

**Looking back at the Research Process**

It proved to be informative and difficult at the same time partly due to navigating and synthesizing a vast amount of information received. At the survey stage it was such activities as creating interesting and appropriate questions and motivators to fill it in. Maintaining accuracy of the answers was important, meaning that the survey process was refined across several cycles. Technical decisions of which columns to encode or drop were part of this stage of preprocessing, and this process highlighted the idea of ensuring that decision making was based on the need to maintain the simplest model whilst passing information.

There are certain difficulties and they were associated with the specifics of modeling while choosing the appropriate algorithms, which were consistent with the structure of the selected dataset and the objectives of the work. Exploration was important and as the model would not be the most accurate the goal was to highlight advantages and shortcomings of the approach chosen. This process established the fact that flexibility and learning is key at each step of the process.

**Alongside, an evaluation of various merits, limitations, and potential pitfalls of the selected methods is presented as part of the reflection.**

The selected methods provided several advantages. Structured survey was adopted to offer overall, real time information required for the research goals. Label encoding of the features made the dataset machine learning ready while at the same time preserving the quality of the dataset. The iterative modeling approach enabled a subjective choice of the suitable algorithms and their performance to be assessed.

However, those methods were not so flawless as some limitations were observed. A limitation to the study was that survey data was limited by the availability of participants and response rates and therefore may have contained bias or missing values. Even though dropping textual columns for example, made the dataset simpler to work with there may have been invaluable contextual data eliminated. The long process of modeling was slightly affected by the dataset used and did not obtain relatively high accuracy despite the available sources; therefore, more enhancement or application of other techniques is needed.

Some risks include overemphasizing features that help to strip data of rich detail but sacrifice depth and the nearly exclusive consideration of model’s accuracy while excluding other factors like precision or recall. Further, one possible cause for the seemingly inferior performance of the models is the modest amount of feature engineering done for the models.

* 1. **Alternative Research Methodologies**

Alternative research methodologies in view of outcomes.

Lessons learned in view of outcomes.

**Other Research Methodologies**

With the research findings in mind, different tools could have been applied to substantially enhance the depth and quality of insights. Although structured and scalable data could be collected using the survey method, interviews would have been both an alternative and a complementary way in which to obtain more detailed and nuanced information. Interviews would enable more in depth discussions with cybersecurity professionals to learn more of the causes of data breaches and effectiveness of mitigation strategies. This method could also capture unique perspectives, personal experience and challenges that are challenging to capture with standardized survey question.

The other potential alternative is to use focus groups, where we can convene a panel of cybersecurity experts and talk through current trends, share experiences and look at ways that we can collaborate together. This approach would encourage a dynamic discussions in which common patterns, or other ideas not captured in individual responses to the question might be revealed. Other ways to get a more granular perspective on data breaches include conducting case studies of particular organizations that have been hit by data breaches. The research will have practical value if this method is used, providing detailed real world examples.

**Lessons Learnt from Outcomes**

The survey method was effective if the breadth of audience can be covered but the focus is to engage participants and in eliciting deeper responses. Many responses were missing the context or detail that would give an understanding of underlying issues. If they had been interviewing, they might have provided much more in the way of a comprehensive explanation of why certain vulnerabilities manifest or how any particular mitigation strategy works, or fails to work, in practice.

As in machine learning models, provided quantitative insights into data patterns, but a more detailed and contextual data input could have strengthened their performance, for example interviews or case studies. Sometimes you might know the organizational culture, leadership priorities, and some specific resource limitations, and that could have added some valuable context to the data analysis.

Looking back, the lesson learned is that a survey combined with interviews and case studies is the way to go, as it could provide a richer data set and, in return, a deeper understanding of both quantitative trends and the qualitative idiosyncrasies. Consequent future research can incorporate alternative methodologies to greatly improve the depth, accuracy, and relevance to practice of the resulting outcomes.

* 1. **Recommended Actions and Future Considerations**

Use reflection to inform future considerations.

As an outcome of the research process, several actionable recommendations and considerations for future studies are discussed. Despite this, this study has successfully identified such causes, along with potential means to mitigate them, but the encountered challenges serve to indicate the need for future improvements in the methodology, analysis and practical implementation.

**Recommended Actions:**

Enhance Data Collection Methods: To overcome limits of survey responses, future research should use surveys to be combined with either interviews or focus groups. These methods offer richer, context data and help fill in some of the aspects of understanding behind data breach causes. In addition, case studies of affected organizations could provide a very rich account of breach scenarios and recovery strategies.

Expand Participant Pool: We should try to involve broader participation in ranging from IT professionals to incident response and executives from diverse sectors. A consequence of this would be to obtain diverse perspectives and increasingly generalize findings.

Improve Dataset Quality for Modeling: Future studies should acquire more complete datasets for the use in machine learning models. This could include including organization, characteristics or behavioral indicators to increase model performance and accuracy of the predictions.

Develop and Test AI Solutions: According to the research findings, more accurate predictive tools for revealing the existence of vulnerabilities and reducing data breaches should be built based on the use of advanced machine learning techniques like deep learning and ensemble models.

Strengthen Organizational Practices: Regular cybersecurity training, vulnerability assessments, and full incident response plan implement should be done by organizations in order to ensure security at work. These actions are consistent with the study’s recommendations and may be assessed further for their real world usefulness.

**Future work** takes a need for a more interdisciplinary approach. Insights from psychology, sociology, and organizational behavior could aid in explaining the human factors for breaches like insider threats and employee negligence. In addition, researchers could then analyze longitudinal studies and analyze how data breach patterns evolve over time and how it informs former threats and how well mitigation strategies are actually effective.

Future studies should experiment with other machine learning algorithms that will help to improve predictive model accuracy and will make it more accurate, including **Catboost, LightGBm, and Neural Network. Catboost and LightGBM** are gradient boosting algorithms for categorical data and imbalanced data, that often works better than traditional tools in terms of accuracy. Understanding complex non linear relationships in the data is possible using neural networks including deep learning architectures and can lead into enhanced classification and prediction outcomes. These advanced techniques may help overcoming performance limitations as we observed in this study, and provide more robust tools for cybersecurity applications.

Future research should also encompass global perspectives, analyzing data breaches across various regions due to different regulations, technologies, and threat landscapes. Thirdly, for the adoption of cybersecurity solutions, addressing ethical consideration, for example, privacy implications of the use of advanced AI tools would serve as a necessary step to secure the responsible deployment of such cybersecurity solutions.

While this study provides a basis for future research to address these recommended actions and considerations, deeper insights and more effective response strategies to prevent data breaches are offered in a continuously evolving threat environment.

* 1. **Recommended Methodology**

Updated version of paper Methodology (flowchart, block diagram) with discussion.

**Data Collection (Stage 01):**

The study suggests introducing real-time or dynamic data feeds (focus group and interviews) and replacing traditional survey methods with automated data scraping from reliable sources like breach reports or threat intelligence platforms to provide more diverse and accurate insights into emerging threats.

**2. Automated Machine Learning (Stage 02):**

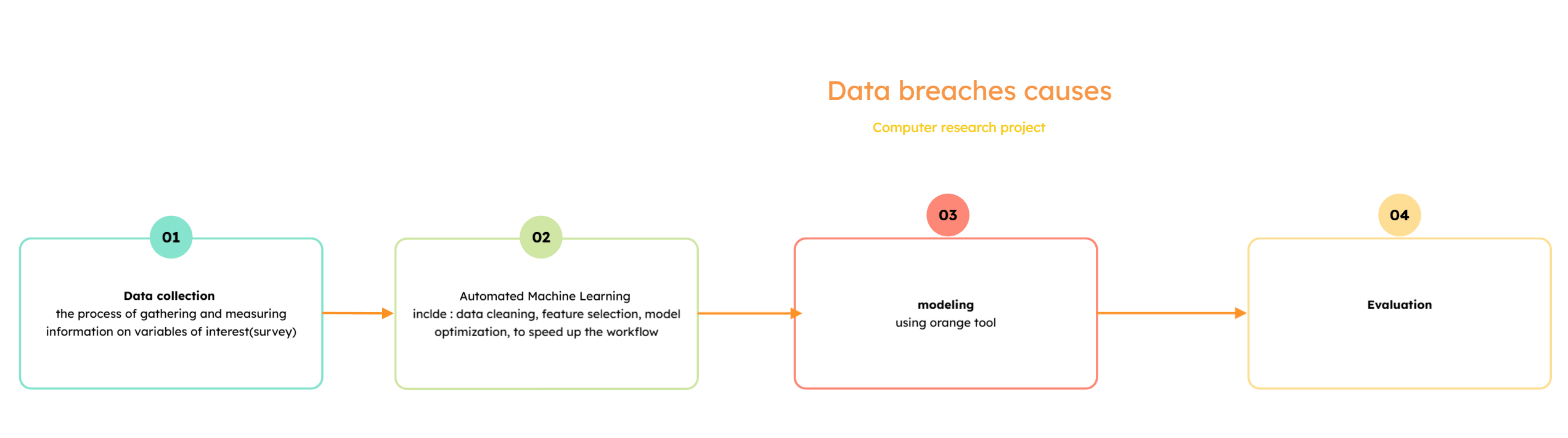
And auto ML frameworks can do the tedious tasks for you: data cleaning, feature selection, model optimization, to speed up the workflow. in this way, manual effort could be reduced, more time could be invested on interpretation and strategic insights.

3. **Modeling (Stage 04)**:

In this proposed change, we replace the conventional machine learning algorithms with the ensemble approach or deep learning; yet, it is a hybrid approach which involves combining the rule-based with machine learning for giving higher accuracy in predicting the complex breaches.

**4. Evaluation (Stage 05):**

These changes solve the problem by replacing the accuracy evaluation with metrics such as F1 score, AUC-RO and recall and introducing exploitability tools like SHAP or LIME to provide a better overview of the model's efficiency and trustworthiness.

****

**References**

[1] KasperSky (2022). How Data Breaches Happen. [online] Kaspersky. Available at: <https://www.kaspersky.com/resource-center/definitions/data-breach>.

[2] Trinckes, J. (2024). AI data breach: Understanding their impact and protecting your data - Thoropass. [online] Thoropass. Available at: <https://thoropass.com/blog/compliance/ai-data-breach/>.

[3] Long Cheng, F. L. D. (. Y., 2017. *Enterprise data breach: causes, challenges, prevention, and future directions.* [Online]   
Available at: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1211>

[4]Adil Hussain Seh 1ORCID, M. Z. 2. A. 2. K. S. 1. A. 1. K. 1. a. A. K. 1., 2020. *Healthcare Data Breaches: Insights and Implications.* [Online]   
Available at: <https://www.mdpi.com/2227-9032/8/2/133>

[5] Gaurav Sood, K. C., 2019. *Pwned: The Risk of Exposure From Data Breaches.* [Online]   
Available at: https://dl.acm.org/doi/10.1145/3292522.3326046

[6] Timothy McIntosh, T. S. T. L. D. X. P. W. D. L. Y. H. A. N. M. H. I. &. C., 2024. *Ransomware Reloaded: Re-examining Its Trend, Research and Mitigation in the Era of Data Exfiltration.* [Online]   
Available at: <https://dl.acm.org/doi/10.1145/3691340>

[7] Amani Ibrahim, D. T.-G. S. A., 2020. *The Challenges of Leveraging Threat Intelligence to Stop Data Breaches.* [Online]   
Available at: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2020.00036/full>

[8] Hussain Afzaal 03 (2024). NLP Based Cyber Security Dataset. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset.

‌