**Central Queensland university (CQU) Intellectual property**

**COIT20265: NETWORK AND INFORMATION SECURITY PROJECT (HT2, 2024)**

**RISK ASSESSMENT REPORT**



**PROJECT:**

**GENERATIVE AI: NAVIGATING SHORT-TERM SKEPTISM AND LONG-TERM PROMISE**

**UNDER THE ESTEEMED GUIDANCE OF**

UNIT CO-ORDINATOR: FARIZA SABRINA

           MENTOR: DR. AHMEDI AZRA

**TEAM MEMBERS:**

BASANTA ADHIKARI (12211752)

BHUWAN THAPA (12196590)

KIRAN BHUSAL (12211570)

PRATIK SHINGH DHAMI (12209929)

Table of Contents

[1. Risk Assessment and Mitigation Plan 3](#_Toc179123583)

[1.1 Cyber Security Risk Assessment 3](#_Toc179123584)

[1.2 TVA in Cybersecurity: 4](#_Toc179123585)

[1.3 Assets Table 6](#_Toc179123586)

[1.4 Threat Table 7](#_Toc179123587)

[1.5 Vulnerability 11](#_Toc179123588)

[1.5.1 Vulnerability Table 11](#_Toc179123589)

[1.6 Risk Rank 15](#_Toc179123590)

[1.7 Vulnerability occurrence and Risk AssessmentTable 17](#_Toc179123591)

[1.8 Risk Mitigation Strategies: 20](#_Toc179123592)

[1.8.1 High-Risk Vulnerabilities Mitigation 20](#_Toc179123593)

[1.8.2 Moderate-Risk Vulnerabilities Mitigation 23](#_Toc179123594)

[1.8.3 Low-Risk Vulnerabilities Mitigation 26](#_Toc179123595)

[References 29](#_Toc179123596)

# 1. Risk Assessment and Mitigation Plan

## 1.1 Cyber Security Risk Assessment

Risk Assessment of GenAI application in cybersecurity for Automobile Service Industry is conducted following NIST Special Publication 800-30 Revision 1, titled "Guide for Conducting Risk Assessments," which offers a structured framework for evaluating risks to information systems (Blank and Gallagher 2012). The guide emphasizes a systematic approach starting with the preparation phase, where organizations define the methodology of the risk assessment. This involves identifying the information system boundaries and assets at risk (Fikri et al. 2019). During the assessment phase, identify and analyse potential threats and vulnerabilities, assessing their likelihood and impact to determine overall risk levels. The process includes identifying threats such as cyber-attacks or insider threats and vulnerabilities like unpatched software or misconfigured systems. (Safitri and Kabetta 2023). The guide suggests categorizing likelihood and impact to evaluate the risk comprehensively.

The next steps involve selecting and implementing appropriate controls to mitigate identified risks. NIST CSF 2.0 core functions provides the framework which is used as cybersecurity controls to mitigate Risks. The CSF Core Functions GOVERN, IDENTIFY, PROTECT, DETECT, RESPOND, and RECOVER organize cybersecurity outcomes at their highest level. The NIST Cybersecurity Framework (CSF) outlines six core functions designed to help organizations manage and respond to cybersecurity risks comprehensively. The first function, **Govern (GV)**, focuses on establishing and communicating the organization's cybersecurity risk management strategy, policies, and governance structures. This function ensures that roles and responsibilities are clearly defined and that cybersecurity efforts align with the organization's broader enterprise risk management (ERM) strategy.

Next, the **Identify (ID)** function helps organizations gain a clear understanding of their current cybersecurity risks by identifying critical assets such as data, hardware, software, systems, and people. By understanding these assets and their associated risks, organizations can prioritize their cybersecurity efforts to address the most pressing vulnerabilities.

The **Protect (PR)** function involves implementing safeguards to manage the identified risks, helping to prevent or reduce the likelihood of cybersecurity incidents. These safeguards include measures like access control, identity management, and data security, along with training programs to raise awareness among employees.

The **Detect (DE)** function ensures that organizations have the capabilities to identify and analyze potential cybersecurity threats or incidents in a timely manner. By monitoring for anomalies and indicators of compromise, organizations can quickly detect attacks before they cause significant damage.

The **Respond (RS)** function comes into play, guiding organizations in taking action to contain and mitigate the impact of cybersecurity incidents. This function involves conducting detailed incident analysis, managing the containment process, and ensuring that the incident is communicated appropriately to stakeholders. The goal is to minimize the damage caused by the incident and to take steps to prevent its recurrence.

Finally, the **Recover (RC)** function focuses on restoring affected assets and operations following a cybersecurity incident. Recovery efforts aim to quickly return to normal business operations while reducing the long-term impact of the event. This function also emphasizes clear communication during the recovery process to ensure that all stakeholders are informed, and that the organization can continue to operate smoothly in the aftermath of an incident.

These controls can be technical, administrative, or physical and should be continuously monitored to ensure their effectiveness. Regular reviews and adjustments are crucial as the risk environment evolves (Möller, 2023). Overall, NIST SP 800-30 Rev. 1 and NIST CSF 2.0 provides a detailed methodology for identifying, analysing, and managing risks, helping organizations protect their information systems and data effectively. In cybersecurity, understanding the relationships between threats, vulnerabilities, and assets is essential for developing robust protection strategies. The Threat-Vulnerability-Asset (TVA) model helps in identifying and managing these elements systematically.

Threats are potential events or actions that can exploit vulnerabilities to cause harm or damage to an organization. They represent the external or internal forces that pose risks. Examples include cyberattacks, insider threats, and natural disasters. (Asaad and Saeed 2022)

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats. They represent gaps in security that could be used to breach or damage the system. Vulnerabilities can be due to software bugs, misconfigurations, lack of updates, or weak authentication mechanisms. (Asaad and Saeed 2022)

Assets are the resources or components of an organization that need protection. They can include hardware, software, data, and processes. Assets are valuable because they store, process, or transmit sensitive information or are crucial for operations. (Asaad and Saeed 2022)

## 1.2 TVA in Cybersecurity:

Threat-Vulnerability-Asset Relationships:

1. Threat: A cyberattack aiming to exploit a weakness.
2. Vulnerability: The weakness in the system, such as unpatched software.
3. Asset: The system or data that the threat is targeting.

For example, if a threat is a malware attack, the vulnerability could be an unpatched software, and the asset could be a server storing critical data.

Using TVA for Risk Management:

1. Identify Threats: Understand what potential threats exist (e.g., phishing attacks, data breaches).
2. Assess Vulnerabilities: Identify weaknesses that could be exploited by these threats (e.g., weak passwords, outdated software).
3. Protect Assets: Focus on securing assets that are at risk (e.g., sensitive data, critical systems).

TVA Example:

1. Threat: Software attack (e.g., code injection).
2. Vulnerability: Poorly secured API endpoints.
3. Asset: GenAI application that processes and stores customer data.

In this case, the threat of a software attack could exploit the vulnerability of poorly secured API endpoints to compromise the asset, which is the GenAI application. This would lead to potential data breaches and operational disruption.

Understanding these relationships, risk analysis is conduct and organizations can prioritize security efforts to address the most critical vulnerabilities and protect their valuable assets from potential threats. This systematic approach helps in developing effective risk management and mitigation strategies (Blank and Gallagher 2012). Table below shows the Assets which are associated with the organization.

## 1.3 Assets Table

|  |  |  |
| --- | --- | --- |
| **Asset No.** | **Asset Type** | **Asset** |
| **1.** | **Hardware** | Servers (Cloud-based) that host the GenAI application and website |
|  |  | Web Server for hosting the company's website |
|  |  | Routers for network connectivity |
|  |  | Firewalls to protect the network |
|  |  | Switches to connect network devices |
|  |  | Workstations/Computers used by staff |
|  |  | Telecommunications Equipment (phones, intercoms) |
|  |  | Load Balancers (Cloud-based) to manage website traffic |
|  |  | Point of Sale (POS) Systems for handling customer payments |
|  |  | Carwash Equipment (automated carwash machines, sensors, etc.) |
| **2.** | **Software** | GenAI Application |
|  |  | Content Management System (CMS) for website updates |
|  |  | Operating Systems on servers and workstations |
|  |  | Security Tools (e.g., antivirus, threat detection software) |
|  |  | APIs for integrating GenAI with the website and other systems |
|  |  | SSL/TLS Certificates for secure web communication |
|  |  | Point of Sale (POS) Software |
|  |  | Customer Relationship Management (CRM) Software |
|  |  | Business Management Software (inventory, scheduling, etc.) |
|  |  | Monitoring Tools for network and application performance |
|  |  | Logging Systems for tracking activities and troubleshooting |
| **3.** | **Data** | Training Data for the GenAI model |
|  |  | Customer Data (personal details, service history, payment info) |
|  |  | Operational Data (logs, configurations, business metrics) |
|  |  | Proprietary Information (business strategies, pricing models) |
|  |  | Financial Data (transaction records, financial statements) |
|  |  | Compliance Documentation Data (data protection, safety regulations) |
|  |  | Legal Agreements Data (customer terms, supplier contracts, privacy policies) |
| **4.** | **Process** | Network Setup Process |
|  |  | Device Configuration Process |
|  |  | GenAI Tool Development and Deployment Process |
|  |  | Customer Interaction Process (in-person and online) |
|  |  | Service Booking Process (online and on-site) |
|  |  | Payment Processing Process |
|  |  | Change Management Process |
|  |  | Incident Response Process |
|  |  | Data Management Process |
|  |  | Access Control Process for IT systems and physical premises |
|  |  | Employee Training Process (for using IT systems and carwash equipment) |
|  |  | Disaster Recovery Plan for IT infrastructure and carwash operations process |
| **5.** | **Cloud Services** | Cloud Storage for data backups |
|  |  | Cloud Compute Instances for running the GenAI model |
|  |  | Cloud Security Services to protect cloud-based resources |

Table-1 Assets Table

## 1.4 Threat Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Threat No** | **Threat Name** | **Description** | **Related Assets** |
| 1 | **Software Attacks** | Includes Malware, DoS (Denial of Service), DDoS (Distributed Denial of Service), MITM (Man in the Middle), Sniffing, DNS Poisoning, Spoofing, Code Injection, Backdoors, and Ransomware. | GenAI Application, Operating System, Security Tools, APIs, Servers (Cloud-based) |
| 2 | **Human Errors** | Refers to Phishing, Server Misconfiguration, Social Engineering, and Accidental Data Deletion. | All Assets (Hardware, Software, Data, Processes) |
| 3 | **Trespass** | Unauthorized Access via password attacks, such as Dictionary Attack, Brute Force Attack, Shoulder Surfing, and Physical Tampering. | Servers, Router, Switch, Computer, Customer Data, Access Control Process, APIs |
| 4 | **Information Extortion** | Ransomware attacks leading to Blackmail or Information Disclosure for Financial Gain. | Customer Data, Proprietary Information, Operational Data, GenAI Application |
| 5 | **Hardware Failures/Errors** | Includes Drive Failures, Server Crashes, and Network Hardware Malfunctions. | Servers, Router, Firewall, Switch, Computer |
| 6 | **Wi-Fi Eavesdropping** | Capturing Sensitive Information, Passwords, or Other Confidential Data through Unauthorized Wireless Network Monitoring. | Router, Switch, Access Control Process, Customer Data |
| 7 | **Software Failures/Errors** | Bugs, Code Performance Issues, Loopholes, and Unpatched Vulnerabilities. | GenAI Application, Operating System, APIs, Security Tools, Chatbot Development Process, Service Booking Process |
| 8 | **Sabotage and Vandalism** | Deliberate Destruction or Tampering of Assets, including Physical Damage to Hardware and Digital Destruction (e.g., Data Wiping, Defacement). | Servers, Firewall, Proprietary Information, Data Management Process |
| 9 | **Forces of Nature** | Natural Disasters such as Cyclones, Fires, Floods, and Lightning Strikes, leading to Data Loss, Hardware Damage, or Operational Downtime. | All Physical Hardware (Servers, Router, Firewall, Switch, Computer), Training Data, Customer Data |
| 10 | **Data Breaches** | Unauthorized access to sensitive data, resulting in information leakage and compliance violations. | Customer Data, Proprietary Information, Training Data, Operational Data |
| 11 | **Insider Threats** | Malicious actions taken by employees or contractors, such as data theft, sabotage, or misuse of access privileges. | All Assets (Hardware, Software, Data, Processes) |
| 12 | **Supply Chain Attacks** | Attacks targeting the organization's suppliers, leading to compromised components or software used within the infrastructure. | GenAI Application, APIs, Security Tools, Operating System |
| 13 | **API Abuse** | Unauthorized use or exploitation of API endpoints, leading to data breaches, service disruption, or unauthorized actions. | APIs, GenAI Application, Operating System |
| 14 | **Credential Theft** | Theft of authentication credentials through phishing, malware, or brute-force attacks, leading to unauthorized access. | Servers, Router, Firewall, Switch, GenAI Application, Customer Data, Access Control Process |
| 15 | **Cloud Security Misconfigurations** | Improper configuration of cloud services, leading to exposure of sensitive data or unauthorized access to cloud-based systems. | Servers (Cloud-based), GenAI Application, Customer Data, Proprietary Information |
| 16 | **Zero-Day Exploits** | Exploitation of unknown vulnerabilities in software before they can be patched, leading to unauthorized access or system compromise. | Operating System, GenAI Application, APIs, Security Tools |
| 17 | **Regulatory Non-Compliance** | Failure to comply with relevant data protection and cybersecurity regulations, leading to fines, legal action, and reputational damage. | All Data (Customer Data, Training Data, Operational Data, Proprietary Information), Processes related to Data Management and Compliance |
| 18 | **Data Poisoning** | Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs. | GenAI Application, Training Data |
| 19 | **Model Inversion** | Attacker’s reverse-engineering the model to infer sensitive information from the training data. | GenAI Application, Training Data |
| 20 | **Adversarial Attacks** | Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions. | GenAI Application |
| 21 | **Model Drift** | Performance degradation over time as the model encounters data differing from its training set. | GenAI Application |
| 22 | **Hallucination** | Generation of nonsensical or incorrect outputs that appear plausible. | GenAI Application |
| 23 | **Bias and Discrimination** | Inherited biases from training data leading to unfair or discriminatory outputs. | GenAI Application |
| 24 | **Explain ability** | Difficulty in understanding and interpreting model outputs, obscuring errors and reducing trust. | GenAI Application |
| 25 | **Dependency on Training Data Quality** | Inadequate or biased training data leading to flawed models and erroneous outputs. | GenAI Application |
| 26 | **Intellectual Property Violations** | Unintentional plagiarism or copyright infringement due to the replication of protected content. | GenAI Application, Proprietary Information |
| 27 | **Overfitting** | The model performs well on training data but poorly on new, unseen data. | GenAI Application |
| 28 | **Malicious Use of GenAI** | GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks. | GenAI Application |
| 29 | **Privacy Invasion** | The application inadvertently exposing sensitive personal or organizational data through its outputs. | GenAI Application, Customer Data |
| 30 | **Algorithmic Manipulation** | Attackers manipulating the GenAI’s algorithm to produce biased or harmful outputs. | GenAI Application |
| 31 | **Unauthorized Access** | Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs. | GenAI Application |
| 32 | **Ethical Violations** | GenAI producing outputs that violate ethical standards, such as generating offensive content. | GenAI Application |

Table-2 Threats Table

## 1.5 Vulnerability

Vulnerabilities are weaknesses or flaws within a system that can be exploited by threats to cause harm or unauthorized access. In the context of Implementing GenAI application for the Automobile Service Industry several critical vulnerabilities have been identified and are in table below.

### 1.5.1 Vulnerability Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Threat** | **Asset** | **Vul.No** | **TVA** | **Vulnerability** |
| **Software Attacks** | Firewall | 1 | T1V1A1 | Weak authentication, such as easily guessable passwords. |
| **Software Attacks** | GenAI Application | 2 | T1V2A2 | Unpatched vulnerabilities in the GenAI codebase. |
| **Software Attacks** | APIs | 3 | T1V3A3 | Poorly secured API endpoints prone to injection attacks. |
| **Human Errors** | Server (Cloud-based) | 4 | T2V4A1 | Misconfiguration during server setup. |
| **Human Errors** | Operating System | 5 | T2V5A2 | Lack of timely updates and patches. |
| **Trespass** | Router | 6 | T3V6A1 | Default credentials not changed, leading to unauthorized access. |
| **Information Extortion** | Customer Data | 7 | T4V7A3 | Lack of encryption, leading to data exposure during extortion. |
| **Hardware Failures/Errors** | Servers (Cloud-based) | 8 | T5V8A1 | Single point of failure in server hardware. |
| **Wi-Fi Eavesdropping** | Switch | 9 | T6V9A1 | Insufficient encryption on wireless communication. |
| **Software Failures/Errors** | GenAI Application | 10 | T7V10A2 | Bugs and code performance issues leading to system crashes. |
| **Sabotage and Vandalism** | Servers (Cloud-based) | 11 | T8V11A1 | Lack of physical security, leading to tampering. |
| **Forces of Nature** | All Physical Hardware | 12 | T9V12A1 | Lack of disaster recovery plans for natural disasters. |
| **Data Breaches** | Customer Data | 13 | T10V13A3 | Weak access controls leading to unauthorized data access. |
| **Insider Threats** | All Assets | 14 | T11V14A1-5 | Insufficient monitoring and auditing of employee activities. |
| **Supply Chain Attacks** | GenAI Application | 15 | T12V15A2 | Compromised third-party software or components used within the application. |
| **API Abuse** | APIs | 16 | T13V16A2 | Insufficient rate limiting and authentication mechanisms on APIs. |
| **Credential Theft** | Access Control Process | 17 | T14V17A4 | Weak password policies leading to credential theft. |
| **Cloud Security Misconfigurations** | Servers (Cloud-based) | 18 | T15V18A1 | Improper configuration of cloud services, leading to exposure of sensitive data. |
| **Zero-Day Exploits** | Operating System | 19 | T16V19A2 | Unpatched zero-day vulnerabilities exploited by attackers. |
| **Regulatory Non-Compliance** | Customer Data, Training Data, Proprietary Information | 20 | T17V20A3 | Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties. |
| **Data Poisoning** | GenAI Application | 21 | T18V21A2 | Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs. |
| **Model Inversion** | GenAI Application | 22 | T19V22A2 | Attacker’s reverse-engineering the model to infer sensitive information from the training data. |
| **Adversarial Attacks** | GenAI Application | 23 | T20V23A2 | Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions. |
| **Model Drift** | GenAI Application | 24 | T21V24A2 | Performance degradation over time as the model encounters data differing from its training set. |
| **Hallucination** | GenAI Application | 25 | T22V25A2 | Generation of nonsensical or incorrect outputs that appear plausible. |
| **Bias and Discrimination** | GenAI Application | 26 | T23V26A2 | Inherited biases from training data leading to unfair or discriminatory outputs. |
| **Dependency on Training Data Quality** | GenAI Application | 28 | T24V27A2 | Inadequate or biased training data leading to flawed models and erroneous outputs. |
| **Intellectual Property Violations** | Proprietary Information | 29 | T25V28A3 | Unintentional plagiarism or copyright infringement due to the replication of protected content. |
| **Overfitting** | GenAI Application | 30 | T26V29A2 | The model performs well on training data but poorly on new, unseen data. |
| **Malicious Use of GenAI** | GenAI Application | 31 | T27V30A2 | GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks. |
| **Privacy Invasion** | GenAI Application and Data | 32 | T28V31A3 | The application inadvertently exposing sensitive personal or organizational data through its outputs. |
| **Algorithmic Manipulation** | GenAI Application | 33 | T29V32A2 | Attackers manipulating the GenAI’s algorithm to produce biased or harmful outputs. |
| **Unauthorized Access** | All Assets | 34 | T30V33A1-5 | Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs. |
| **Ethical Violations** | GenAI Application | 35 | T31V34A2 | GenAI producing outputs that violate ethical standards, such as generating offensive content. |

 Table-3 Vulnerability Table

## 1.6 Risk Rank

In risk assessment, the risk rank is a crucial metric used to prioritize vulnerabilities based on their likelihood of exploitation and the potential impact if they are exploited. The likelihood measures how probable it is that a vulnerability will be exploited, categorized as High, Moderate, or Low. Impact evaluates the potential consequences of exploitation, also categorized as High, Moderate, or Low. By combining these two factors, the overall risk is determined (Blank and Gallagher 2012).

For example, a vulnerability with High likelihood and High impact is deemed a High-risk issue, requiring immediate attention and remediation. Conversely, a vulnerability with Low likelihood and Low impact is considered Low risk and is given a lower priority. The risk rank assigns a numerical value or category, such as Rank 1 for Critical High risks that demand urgent action, rank 2 for Moderate risks that should be addressed soon, and Rank 3 correspond to risks that are less critical and can be addressed as resources permit. This prioritization ensures that the most significant threats are managed first, effectively reducing the potential for substantial damage or disruption to the organization.

|  |  |  |  |
| --- | --- | --- | --- |
| **Likelihood** | **Impact** | **Combined Risk** | **Risk** |
| High | Low | High /Low | Low |
| Moderate | Low | Moderate /Low | Low |
| Low | Low | Low /Low | Low |
| High | Moderate | High /Moderate | Moderate |
| Moderate | Moderate | Moderate /Moderate | Moderate |
| Low | Moderate | Low/Moderate | Low |
| High | High | High /High | High |
| Moderate | High | Moderate /High | Moderate |
| Low | High | Low /High | Low |

Table-4 Risk Rank Value Table

## 1.7 Vulnerability occurrence and Risk AssessmentTable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Vul No.** | **Vulnerability** | **Likelihood** | **Impact** | **Risk** | **Rank** |
| 1 | Weak authentication, such as easily guessable passwords. | Moderate | High | Moderate | 2 |
| 2 | Unpatched vulnerabilities in the GenAI codebase. | High | High | High | 1 |
| 3 | Poorly secured API endpoints prone to injection attacks. | High | Moderate | Moderate | 2 |
| 4 | Misconfiguration during server setup. | Moderate | High | Moderate | 2 |
| 5 | Lack of timely updates and patches. | Moderate | High | Moderate | 2 |
| 6 | Default credentials not changed, leading to unauthorized access. | High | High | High | 1 |
| 7 | Lack of encryption, leading to data exposure during extortion. | Moderate | High | Moderate | 2 |
| 8 | Single point of failure in server hardware. | Low | High | Moderate | 3 |
| 9 | Insufficient encryption on wireless communication. | Moderate | Moderate | Moderate | 3 |
| 10 | Bugs and code performance issues leading to system crashes. | High | High | High | 1 |
| 11 | Lack of physical security, leading to tampering. | Moderate | High | Moderate | 2 |
| 12 | Lack of disaster recovery plans for natural disasters. | Low | High | Moderate | 3 |
| 13 | Weak access controls leading to unauthorized data access. | High | High | High | 1 |
| 14 | Insufficient monitoring and auditing of employee activities. | Moderate | High | Moderate | 2 |
| 15 | Compromised third-party software or components used within the application. | Moderate | High | Moderate | 2 |
| 16 | Insufficient rate limiting and authentication mechanisms on APIs. | High | Moderate | High | 2 |
| 17 | Weak password policies leading to credential theft. | High | High | High | 1 |
| 18 | Improper configuration of cloud services, leading to exposure of sensitive data. | Moderate | High | Moderate | 2 |
| 19 | Unpatched zero-day vulnerabilities exploited by attackers. | Moderate | High | Moderate | 2 |
| 20 | Lack of adherence to data protection and cybersecurity regulations, leading to legal and financial penalties. | Moderate | High | Moderate | 2 |
| 21 | Malicious actors introducing corrupted data into the training datasets, leading to harmful or inaccurate outputs. | High | High | High | 1 |
| 22 | Attackers reverse-engineering the model to infer sensitive information from the training data. | Moderate | High | Moderate | 2 |
| 23 | Small, crafted inputs causing the GenAI model to make incorrect or harmful predictions. | High | High | High | 1 |
| 24 | Performance degradation over time as the model encounters data differing from its training set. | Moderate | Moderate | Moderate | 3 |
| 25 | Generation of nonsensical or incorrect outputs that appear plausible. | Moderate | Moderate | Moderate | 3 |
| 26 | Inherited biases from training data leading to unfair or discriminatory outputs. | Moderate | High | Moderate | 3 |
| 27 | Difficulty in understanding and interpreting model outputs, obscuring errors and reducing trust. | Moderate | Moderate | Moderate | 3 |
| 28 | Inadequate or biased training data leading to flawed models and erroneous outputs. | Moderate | High | Moderate | 2 |
| 29 | Unintentional plagiarism or copyright infringement due to the replication of protected content. | Moderate | High | Moderate | 2 |
| 30 | The model performs well on training data but poorly on new, unseen data. | Moderate | Moderate | Moderate | 3 |
| 31 | GenAI being leveraged by attackers to create deepfakes, phishing content, or other forms of social engineering attacks. | High | High | High | 1 |
| 32 | The application inadvertently exposing sensitive personal or organizational data through its outputs. | Moderate | High | Moderate | 2 |
| 33 | Attackers manipulating the GenAI’s algorithm to produce biased or harmful outputs. | High | High | High | 1 |
| 34 | Insufficient access controls allowing unauthorized users to manipulate the GenAI application or its outputs. | High | High | High | 1 |
| 35 | GenAI producing outputs that violate ethical standards, such as generating offensive content. | Moderate | High | Moderate | 2 |

Table-5 Vulnerability occurrence and Risk rank Table

## 1.8 Risk Mitigation Strategies:

To address the identified vulnerabilities and associated risks effectively, a comprehensive risk mitigation strategy is developed. According to the NIST Cybersecurity Framework (CSF) 2.0 To organize vulnerabilities and corresponding mitigation strategies, each vulnerability is categorized by NIST core functions (Identify, Protect, Detect, Respond, Recover and Govern) with clear mappings to the respective mitigation strategies (Dimakopoulou and Rantos 2024). This strategy includes preventative measures, detection mechanisms, and response plans tailored to the specific vulnerabilities and their risk rankings (Katsumata et Al. 2010), (Blank and Gallagher 2012). Robust mitigation strategies categorized by NIST CSF category and sub-category risk rank are as Follows:

### 1.8.1 High-Risk Vulnerabilities Mitigation

1. **Unpatched Vulnerabilities (Vul No. 2):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).

Implement a robust patch management process. Regularly update and patch all software components, including third-party libraries. Conduct periodic security audits and vulnerability scans to identify and address potential issues. Default.

1. **Credentials Not Changed (Vul No. 6):**

* **Category:** PROTECT
* **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
* **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

Enforce strong password policies and ensure that default credentials are changed during initial setup. Use multi-factor authentication (MFA) to enhance access security.

1. **Bugs and Code Performance Issues (Vul No. 10):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-06 (Secure software development practices are integrated, and their performance is monitored throughout the software development life cycle).

Establish a rigorous code review and testing process. Implement automated testing tools to detect bugs and performance issues early in the development cycle. Conduct regular performance evaluations and stress testing.

1. **Weak Access Controls (Vul No. 13):**

* **Category:** PROTECT
* **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
* **Mitigation:** Align with PR. AA-05 (Access permissions, entitlements, and authorizations are defined in a policy, managed, enforced, and reviewed, and incorporate the principles of least privilege and separation of duties).

Implement granular access controls and enforce the principle of least privilege. Regularly review and update access permissions. Use MFA and strong authentication mechanisms for sensitive systems.

1. **Weak Password Policies (Vul No. 17):**

* **Category:** PROTECT
* **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
* **Mitigation:** Align with PR. AA-01 (Identities and credentials for authorized users, services, and hardware are managed by the organization).

Develop and enforce strong password policies, including complexity requirements and regular password changes. Implement MFA for critical systems and accounts.

1. **Malicious Data in Training Datasets (Vul No. 21):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Implement data validation and cleansing procedures to detect and remove corrupted or malicious data. Use secure data sources and regularly review training data for integrity.

1. **Reverse-Engineering of the Model (Vul No. 22):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-05 (Installation and execution of unauthorized software are prevented).

Employ model protection techniques, such as encryption and obfuscation, to safeguard intellectual property. Regularly review and update security measures around model deployment

1. **Small, Crafted Inputs Causing Harmful Predictions (Vul No. 23):**

* **Category:** DETECT
* **Subcategory:** Continuous Monitoring (DE.CM)
* **Mitigation:** Align with DE.CM-09 (Computing hardware and software, runtime environments, and their data are monitored to find potentially adverse events).

Implement robust input validation and sanitization processes. Use anomaly detection mechanisms to identify and mitigate malicious inputs.

1. **GenAI Used for Social Engineering Attacks (Vul No. 31):**

* **Category:** DETECT
* **Subcategory:** Adverse Event Analysis (DE.AE)
* **Mitigation:** Align with DE.AE-02 (Potentially adverse events are analyzed to better understand associated activities).

Monitor and restrict the use of GenAI for generating sensitive or potentially harmful content. Implement detection mechanisms to identify misuse of the model.

1. **Algorithmic Manipulation (Vul No. 33):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Apply security controls to prevent unauthorized access and modifications to the GenAI algorithm. Conduct regular audits to detect and address potential manipulations.

1. **Unauthorized Access to GenAI Application (Vul No. 34):**

* **Category:** PROTECT
* **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
* **Mitigation:** Align with PR. AA-06 (Physical access to assets is managed, monitored, and enforced commensurate with risk).

Strengthen access controls and implement logging and monitoring to detect unauthorized access attempts. Use encryption and secure authentication mechanisms.

### 1.8.2 Moderate-Risk Vulnerabilities Mitigation

1. **Weak Authentication (Vul No. 1):**

* **Category:** PROTECT
* **Subcategory:** Identity Management, Authentication, and Access Control (PR. AA)
* **Mitigation:** Align with PR. AA-02 (Identities are proofed and bound to credentials based on the context of interactions).

Strengthen authentication mechanisms by enforcing strong password policies and implementing MFA. Regularly review and update authentication methods.

1. **Poorly Secured API Endpoints (Vul No. 3):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Secure API endpoints with proper authentication and authorization mechanisms. Implement rate limiting and input validation to protect against injection attacks.

1. **Lack of Timely Updates and Patches (Vul No. 5):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-02 (Software is maintained, replaced, and removed commensurate with risk).
* Develop a patch management strategy to ensure timely application of security updates. Monitor for new vulnerabilities and apply patches as needed (Huang et al., 2024).

1. **Lack of Encryption for Data Exposure (Vul No. 7):**

* **Category:** PROTECT
* **Subcategory:** Data Security (PR.DS)
* **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).

Implement strong encryption for data at rest and in transit. Regularly review encryption practices to ensure they meet current security standards.

1. **Insufficient Monitoring and Auditing (Vul No. 14):**

* **Category:** DETECT
* **Subcategory:** Continuous Monitoring (DE.CM)
* **Mitigation:** Align with DE.CM-01 (Networks and network services are monitored to find potentially adverse events).

Enhance monitoring and auditing processes to detect and respond to suspicious activities. Implement comprehensive logging and regular reviews of access and activity logs.

1. **Compromised Third-Party Components (Vul No. 15):**

* **Category:** GOVERN
* **Subcategory:** Cybersecurity Supply Chain Risk Management (GV.SC)
* **Mitigation:** Align with GV.SC-07 (The risks posed by a supplier, their products and services, and other third parties are understood, recorded, prioritized, assessed, responded to, and monitored over the course of the relationship).

Evaluate and vet third-party software and components for security. Regularly update and patch third-party components and use trusted sources.

1. **Insufficient Rate Limiting on APIs (Vul No. 16):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Implement rate limiting and API gateway controls to manage and restrict API traffic. Monitor API usage for unusual patterns (Huang et al., 2024).

1. **Improper Configuration of Cloud Services (Vul No. 18):**

* **Category:** PROTECT
* **Subcategory:** Platform Security (PR.PS)
* **Mitigation:** Align with PR.PS-01 (Configuration management practices are established and applied).

Regularly review and audit cloud service configurations to ensure compliance with best practices. Implement security controls for data protection and access management.

1. **Unpatched Zero-Day Vulnerabilities (Vul No. 19):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-02 (Cyber threat intelligence is received from information sharing forums and sources).

Stay informed about emerging vulnerabilities and apply updates as soon as patches become available. Use threat intelligence to anticipate and prepare for zero-day threats (Krishnamurthy 2023).

1. **Lack of Adherence to Regulations (Vul No. 20):**

* **Category:** GOVERN
* **Subcategory:** Organizational Context (GV.OC)
* **Mitigation:** Align with GV.OC-03 (Legal, regulatory, and contractual requirements regarding cybersecurity — including privacy and civil liberties obligations — are understood and managed).

Ensure compliance with data protection and cybersecurity regulations through regular audits and updates to policies and procedures. Provide training to employees on regulatory requirements.

1. **Inherited Biases in Training Data (Vul No. 26):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
* Regularly review and audit training data for biases. Implement techniques to detect and mitigate bias in model training.

1. **Inadequate or Biased Training Data (Vul No. 28):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).
* Use diverse and representative datasets for training. Regularly review data quality and make necessary adjustments to improve model accuracy and fairness (Krishnamurthy 2023).

1. **Unintentional Plagiarism (Vul No. 29):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Implement measures to detect and avoid replication of protected content. Use plagiarism detection tools and adhere to copyright laws.

1. **Privacy Invasion Through Outputs (Vul No. 32):**

* **Category:** PROTECT
* **Subcategory:** Data Security (PR.DS)
* **Mitigation:** Align with PR.DS-01 (The confidentiality, integrity, and availability of data-at-rest are protected).

Implement data anonymization techniques and review outputs to ensure they do not expose sensitive information.

1. **Ethical Violations (Vul No. 35):**

* **Category:** GOVERN
* **Subcategory:** Organizational Context (GV. OC)
* **Mitigation:** Align with GV. OC-01 (The organizational mission is understood and informs cybersecurity risk management).

Establish ethical guidelines for GenAI usage and ensure compliance through regular reviews and audits. Implement mechanisms to detect and address unethical outputs.

### 1.8.3 Low-Risk Vulnerabilities Mitigation

1. **Single Point of Failure in Server Hardware (Vul No. 8):**

* **Category:** PROTECT
* **Subcategory:** Technology Infrastructure Resilience (PR.IR)
* **Mitigation:** Align with PR.IR-03 (Mechanisms are implemented to achieve resilience requirements in normal and adverse situations).

Implement redundancy and failover solutions to mitigate the impact of hardware failures. Regularly test disaster recovery plans.

1. **Insufficient Encryption on Wireless Communication (Vul No. 9):**

* **Category:** PROTECT
* **Subcategory:** Data Security (PR.DS)
* **Mitigation:** Align with PR.DS-02 (The confidentiality, integrity, and availability of data-in-transit are protected).
* Use strong encryption protocols for wireless communications. Regularly review and update encryption practices.

1. **Lack of Disaster Recovery Plans (Vul No. 12):**

* **Category:** RESPOND
* **Subcategory:** Incident Recovery Plan Execution (RC.RP)
* **Mitigation:** Align with RC.RP-01 (The recovery portion of the incident response plan is executed once initiated from the incident response process).

Develop and test comprehensive disaster recovery plans. Ensure plans address various scenarios and include regular updates.

1. **Performance Degradation Over Time (Vul No. 24):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).
* Implement monitoring tools to track model performance and conduct regular evaluations. Update models as necessary to address performance issues.

1. **Generation of Nonsensical Outputs (Vul No. 25):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Continuously review and refine the model to improve output quality. Implement feedback mechanisms to detect and address nonsensical outputs.

1. **Difficulty in Interpreting Model Outputs (Vul No. 27):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-01 (Vulnerabilities in assets are identified, validated, and recorded).

Enhance model explain ability and provide clear documentation on output interpretation. Use visualization tools to aid in understanding outputs.

1. **The Model Performs Poorly on New Data (Vul No. 30):**

* **Category:** IDENTIFY
* **Subcategory:** Risk Assessment (ID.RA)
* **Mitigation:** Align with ID.RA-04 (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded).

Implement continuous learning and adaptation strategies to improve model performance on new data. Regularly update the training data to reflect changing patterns.

# References

Asaad, R.R. and Saeed, V.A., 2022. A Cyber Security Threats, Vulnerability, Challenges and Proposed Solution. *Applied computing Journal*, pp.227-244.

Blank, R. and Gallagher, P., 2012. Nist special publication 800-30 revision 1 guide for conducting risk assessments. *National Institute of Standards and Technology*.

Dimakopoulou, A. and Rantos, K., 2024. Comprehensive Analysis of Maritime Cybersecurity Landscape Based on the NIST CSF v2. 0. *Journal of Marine Science and Engineering*, *12*(6), p.919.

Huang, K., Yeoh, J., Wright, S. and Wang, H., 2024. Build your security program for genai. In *Generative AI Security: Theories and Practices* (pp. 99-132). Cham: Springer Nature Switzerland.

Katsumata, P., Hemenway, J. and Gavins, W., 2010, October. Cybersecurity risk management. In *2010-MILCOM 2010 Military Communications Conference* (pp. 890-895). IEEE.

Möller, D.P., 2023. NIST cybersecurity framework and MITRE cybersecurity criteria. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices* (pp. 231-271). Cham: Springer Nature Switzerland.

Safitri, E.H.N. and Kabetta, H., 2023, August. Cyber-Risk Management Planning Using NIST CSF V1. 1, ISO/IEC 27005: 2018, and NIST SP 800-53 Revision 5 (A Study Case to ABC Organization). In *2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)* (pp. 332-338). IEEE.