Risk Assessment Report

Prepared by: Isabelle Jaber

Date: September 18th, 2024

Table of Contents:

EXECUTIVE SUMMARY

This report provides an overview of key AI-related security risks identified for VISION's operations, focusing on applications and assets such as the Data Integration Platform, Cloud Storage, Data Annotation Application, Customer Relationship Management (CRM) System, and Pre-Trained AI Model Library. Given VISION's involvement in sensitive healthcare data processing and AI development, robust security measures are essential to safeguard data, maintain client trust, and ensure regulatory compliance. The risks identified include Data Breaches, Unauthorized Access, Adversarial Attacks, and Model Poisoning. Recommendations are provided to mitigate these risks and enhance VISION's overall security posture.

HIGH-IMPACT RISK OVERVIEW

The following AI-related risks have been identified as having the highest potential impact on VISION's operations and its healthcare clients:

DATA BREACH

- **Applicable Applications**: Data Integration Platform, Cloud Storage, Data Annotation Application, CRM System and Pre-Trained AI Model Library.
- **Description**: A data breach could expose sensitive healthcare information, leading to financial penalties, loss of client trust, and potential legal action under HIPAA.
- **Impact**: High. The nature of the healthcare data VISION handles makes this a critical risk.

UNAUTHORIZED ACCESS

- **Applicable Applications**: Data Integration Platform, Cloud Storage, CRM System, and Pre-Trained AI Model Library.
- **Description**: Unauthorized access to VISION's or its client's systems could allow attackers to steal or alter sensitive data.
- **Impact**: High. This could compromise patient data integrity and system availability.

ADVERSARIAL ATTACKS

- **Applicable Applications**: Data Ingestion Interface, Data Annotation Application, Pre-Trained AI Model Library.
- **Description**: Adversarial attacks involve intentionally crafted inputs designed to deceive AI models, leading to incorrect outputs or unsafe actions, particularly critical in healthcare contexts.
- **Impact**: Critical. The reliability of AI-driven decisions could be severely undermined.

MODEL POISONING

- **Applicable Applications**: Data Ingestion Interface, Data Annotation Application, Pre-Trained AI Model Library.
- **Description**: Attackers may inject malicious data into the training set, corrupting the model's learning process and resulting in flawed predictions.
- **Impact**: Critical. This poses a direct threat to the integrity of AI systems.

CODE INJECTION/DATA POISONING

- **Applicable Applications**: Data Integration Platform, Data Annotation Application, and Pre-Trained AI Model Library.
- **Description**: Code injection attacks (e.g., SQL injection) could exploit vulnerabilities in AI systems, leading to unauthorized data manipulation.
- **Impact**: High. This can severely compromise system functionality.

MAN IN THE MIDDLE (MITM) ATTACK

- **Applicable Applications**: Data Integration Platform, Cloud Storage, and CRM System.
- **Description**: Attackers could intercept and alter communications between VISION's systems and the client's, compromising data confidentiality.
- **Impact**: Medium-High. Given the healthcare context, secure communication is crucial to protect patient data.

PHISHING

- **Applicable Applications**: Data Integration Platform and CRM System.
- **Description**: Phishing attacks could target VISION's employees or clients, leading to credential theft or malicious access to sensitive systems.
- **Impact**: Medium-High. Phishing remains one of the most common entry points for attackers.

THIRD-PARTY RISKS

- **Applicable Applications**: Cloud Storage, Data Annotation Application, and CRM System.
- **Description**: Reliance on third-party vendors and contractors may introduce vulnerabilities, particularly if security practices are inconsistent.
- **Impact**: Medium-High. Third parties are often weak points in the security chain.

NON-COMPLIANCE WITH DATA PROTECTION POLICIES

- **Applicable Applications**: Data Integration Platform, Cloud Storage, Data Annotation Application, CRM System and Pre-Trained AI Model Library.
- **Description**: Non-compliance with healthcare-specific regulations such as HIPAA could lead to severe penalties and operational restrictions.
- **Impact**: High. Non-compliance can lead to legal issues, especially in the healthcare sector.

PRIVILEGE ESCALATION

- **Applicable Applications:** Data Integration Platform, Cloud Storage, CRM System, and Data Annotation Application.
- **Description**: Attackers could exploit vulnerabilities to gain higher levels of access within VISION's or the client's systems.
- **Impact**: Medium-High. Privilege escalation can lead to widespread system compromise and threaten the confidentiality, integrity, and availability of sensitive data.

OUTDATED SOFTWARE AND UNPATCHED VULNERABILITIES

- **Applicable Applications**: Data Integration Platform, Cloud Storage, Data Annotation Application, CRM System and Pre-Trained AI Model Library.
- **Description**: Outdated software or unpatched vulnerabilities leave systems open to exploitation.
- **Impact**: Medium. This is a common issue in security management but can lead to significant breaches if exploited.

INSECURE CONFIGURATIONS

- **Applicable Applications**: Data Integration Platform, Cloud Storage, Data Annotation Application, CRM System and Pre-Trained AI Model Library.
- **Description**: Improperly configured systems, including weak access controls or default credentials, could be exploited.
- **Impact**: Medium. Misconfigurations are often overlooked but can be easily exploited by attackers.

RECOMMENDATIONS

To mitigate the identified risks, the following recommendations are provided:

DATA BREACH MITIGATION

- **Encrypt Model and Training Data**: Ensure that all data (both in training sets and models) are encrypted at rest and in transit to prevent unauthorized access and to ensure that even if the data is accessed it remains unreadable.
    - **Field-Level Encryption**: Encrypt specific, sensitive fields to ensure that even in the event of a breach, the most critical data remains protected.
    - **Data-at-Rest and Data-in-Transit Encryption**: Apply full-disk encryption for data at rest and Transport Layer Security (TLS 1.3) encryption for data in transit to ensure that unauthorized parties cannot intercept or access sensitive data.
    - **Data Masking**: Apply data masking techniques to hide sensitive information in environments where full access is not required. This can reduce the exposure of sensitive data without compromising functionality for testing, training, or analytics purposes.
    - **Tokenization**: Replace sensitive data with unique tokens during storage and processing, ensuring that real data is only accessible to authorized systems. This mitigates risks even in the event of a data breach, as the tokenized data has no exploitable value without access to the tokenization system.
- **Secure Model Predictions**: Use encryption for communication between Machine Learning (ML) models and their consumers, ensuring that predictions and results remain secure.
- **Model Access Control**: Apply strict access control to ML models to prevent unauthorized tampering or inference from sensitive datasets. Role-based access control (RBAC) should extend to model access.
- **Federated Learning & Differential Privacy**: Incorporate privacy-preserving technologies like federated learning and differential privacy to minimize risks from data exposure during model training.
- **Data Loss Prevention (DLP) Policies**: Tailor DLP policies based on data classification and context. Implement DLP solutions that automatically detect, monitor, and restrict sensitive data transfers. This helps prevent unauthorized access or sharing of sensitive data like patient health information or intellectual property.
- **Security Audits**: Establish regular security audits and penetration testing to identify vulnerabilities.

UNAUTHORIZED ACCESS PREVENTION

- **Multi-Factor Authentication for ML Systems**: Deploy MFA, not only for general access but, specifically for systems where models are trained, stored, or deployed.
- **Attribute-Based Access Control (ABAC)**: Instead of simple RBAC, implement ABAC to ensure access decisions are made based on attributes like user roles, data sensitivity, and context.
- **Model Usage Monitoring**: Implement continuous monitoring and logging for any model-related access to detect and respond to abnormal access patterns.

ADVERSARIAL ATTACK MITIGATION

- **Adversarial Training**: Integrate adversarial examples into the training data to enhance model robustness.
- **Input Validation**: Ensure all inputs are validated and sanitized to prevent adversarial manipulation.

MODEL POISONING DEFENSE

- **Anomaly Detection**: Utilize anomaly detection methods to identify potential data poisoning attempts during training.
- **Robustness Testing**: Regularly test models against adversarial examples to identify vulnerabilities.

CODE INJECTION/DATA POISONING DEFENSE

- **Input Sanitization for Model Training Data**: Ensure that data being used to train models is sanitized and validated to prevent malicious data injections that could alter model performance.
- **Model Poisoning Defense**: Use anomaly detection techniques to monitor for poisoning attempts during model training (e.g., detecting outlier or adversarial samples).
- **Code Reviews for ML Frameworks**: Conduct regular code reviews not only for applications but for machine learning frameworks and libraries to mitigate vulnerabilities like backdoors in models.
- Use input validation and implement **parameterized queries** to prevent SQL injection.

MAN IN THE MIDDLE (MitM) ATTACK PREVENTION

- **Encrypt Model API Endpoints**: Use strong encryption (TLS 1.3) for all API endpoints that serve ML models to ensure that model predictions and communications are secure.

- **Model Integrity Verification**: Use cryptographic signatures to verify the integrity of machine learning models when being transmitted or deployed to prevent MITM tampering.

PHISHING DEFENSE (For DS/ML Platform Operators)

- **Train Data Scientists and ML Engineers on Phishing Threats**: Provide specific training focused on phishing and social engineering risks that target personnel who have access to sensitive model or training data.
- **Anti-Phishing for Data Pipelines**: Implement phishing-resistant measures for critical components of data pipelines (e.g., authenticating datasets received from external sources to prevent rogue data sources).

THIRD-PARTY RISK MITIGATION

- **Security Due Diligence for Third-Party ML Models**: Conduct thorough security evaluations and testing on any third-party pre-trained models or libraries before integrating them into production systems, ensuring no backdoors or malicious components are present.
- **Secure ML-as-a-Service Providers**: Ensure third-party cloud providers hosting ML services (like training or inference) adhere to stringent security standards and provide robust isolation for data and models.

COMPLIANCE WITH DATA PROTECTION POLICIES

- **Anonymize Data for ML**: Use anonymization techniques/applications to de-identify personally identifiable information (PII) in training datasets, especially for healthcare data, to ensure compliance with regulations like HIPAA.
- **Model Auditing**: Ensure that all models can be audited for how they were trained and what data was used, allowing full transparency for compliance purposes.
- **Data Retention Policies for Training Sets**: Establish clear policies on the retention and deletion of training data to ensure compliance with data protection laws (e.g., GDPR, HIPAA).

PRIVILEGE ESCALATION PREVENTION

- **Implement Least Privilege for ML Tools**: Ensure that data scientists, engineers, and other team members only have access to the ML models and datasets necessary for their role.

- **Use Role-Based Access (RBAC) for Model Deployment**: Restrict the ability to deploy or update models in production environments to limited, trusted personnel, and monitor these actions through a privileged access management system.
- **Model Access Logging**: Log all accesses and updates to models to detect privilege escalation attempts and other unauthorized actions.

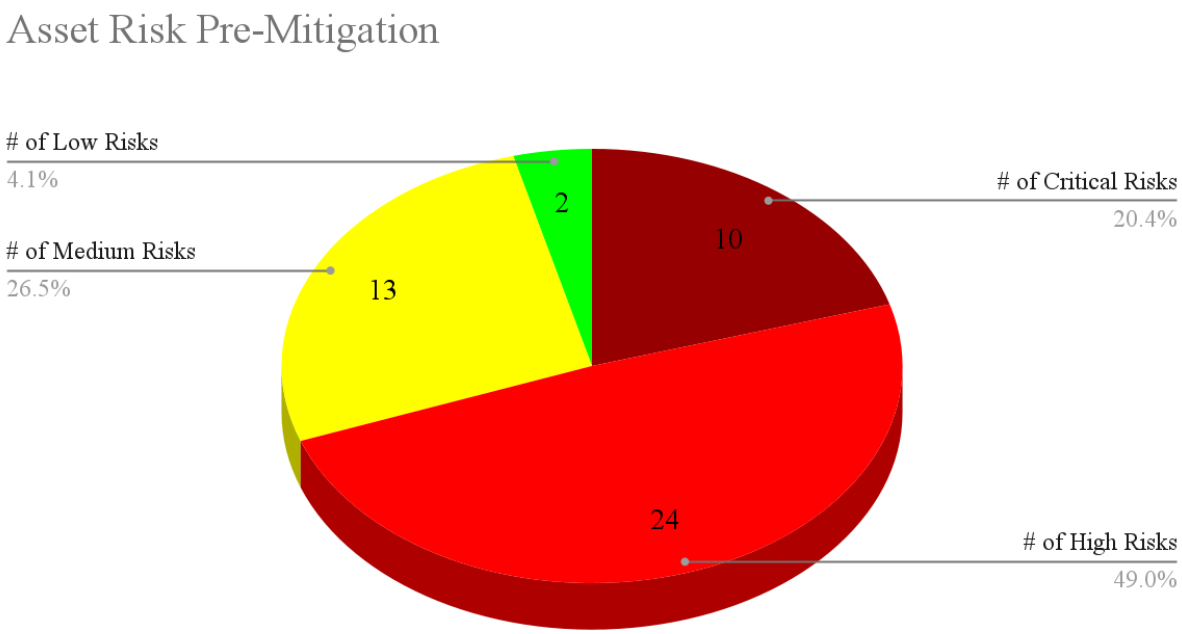
OUTDATED SOFTWARE AND UNPATCHED VULNERABILITIES

- **Automated Patch Management for ML Libraries**: Ensure that ML frameworks and libraries (e.g., TensorFlow, PyTorch) are part of an automated patch management system to keep them tested and updated with the latest security patches.
- **Monitor for Vulnerabilities in ML Software**: Continuously monitor for new security advisories related to machine learning platforms, libraries, and dependencies.

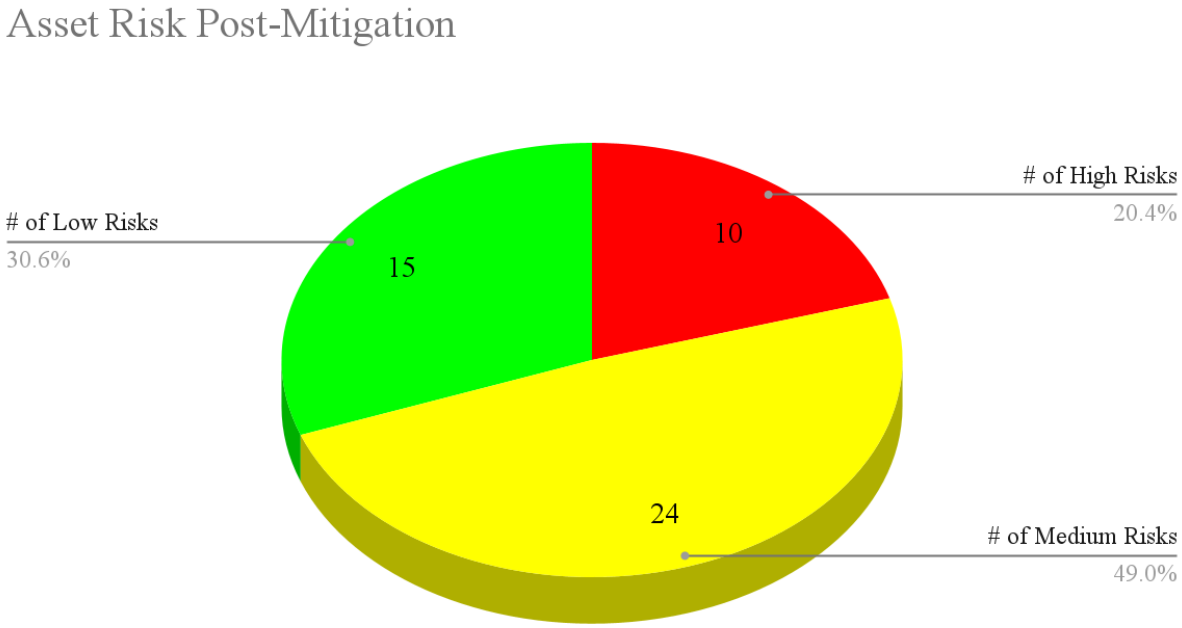
INSECURE CONFIGURATION REMEDIATION

- **Harden ML Infrastructure Configurations**: Ensure that cloud-based and on-premise ML infrastructures are configured according to security best practices (e.g., secure API configurations, and IAM policies for cloud services).
- **Container Security for ML Models**: If models are deployed in containers, ensure they are hardened and follow security best practices, such as minimizing the attack surface and ensuring secure image configurations.
- **Configuration Drift Management for ML Environments**: Implement tools that detect configuration drift in machine learning environments to ensure they remain aligned with security baselines over time.

Asset Risk Pre-Mitigation



Asset Risk Post-Mitigation



CONCLUSION

By addressing the identified high-impact risks, VISION can significantly improve its security posture, especially where it pertains to the privacy and security of sensitive healthcare data. The recommendations provided aim to reduce the likelihood and impact of threats such as data breaches, unauthorized access, and code injections. Implementing these controls will not only protect VISION's systems and client data but will also ensure compliance with relevant data protection laws such as HIPAA. Continuous monitoring and security assessments should be prioritized to adapt to evolving threats and maintain a strong defense.

RESOURCES

OWASP Machine Learning Security Top 10
https://owasp.org/www-project-machine-learning-security-top-10/
Odyssey: A Systems Approach to Machine Learning Security
https://www.mitre.org/news-insights/publication/odyssey-systems-approach-machine-learning-security
Microsoft and MITRE Create Tool to Help Security Teams Prepare for Attacks on Machine Learning Systems
https://www.mitre.org/news-insights/news-release/microsoft-and-mitre-create-tool-help-security-teams-prepare-attacks
Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal. Rahul Pankajakshan, Sumitra Biswal, Yuvaraj Govindarajulu, Gilad Gressel.
https://ar5iv.labs.arxiv.org/html/2403.13309v1
Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson.
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf
NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems
https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems
MITRE ATLAS
https://atlas.mitre.org/