Threat Modeling Report

Prepared by Isabelle Jaber

Date: September 18th, 2024

Table of Contents:

EXECUTIVE SUMMARY..3
THREAT MODELING OVERVIEW...4
THREAT MODEL...5
COMPENSATING CONTROLS.. 7
RECOMMENDATIONS..9
CONCLUSION... 10

EXECUTIVE SUMMARY

This report presents a detailed threat modeling analysis for VISION's healthcare imaging AI solution. The analysis focuses on AI-specific threats within the context of handling sensitive medical data. Key findings identify potential risks associated with AI model integrity, training data security, and compliance with data privacy regulations. The report uses the STRIDE framework to categorize and assess threats and provides compensating controls and recommendations to address these issues. Implementing these measures will enhance the security of the AI systems and protect patient data from various threats.

THREAT MODELING OVERVIEW

The STRIDE threat modeling framework has been applied to assess security risks related to AI components of VISION's healthcare imaging solution. STRIDE categorizes threats into six areas: Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. The application of STRIDE to this scenario emphasizes AI-specific threats and security measures.

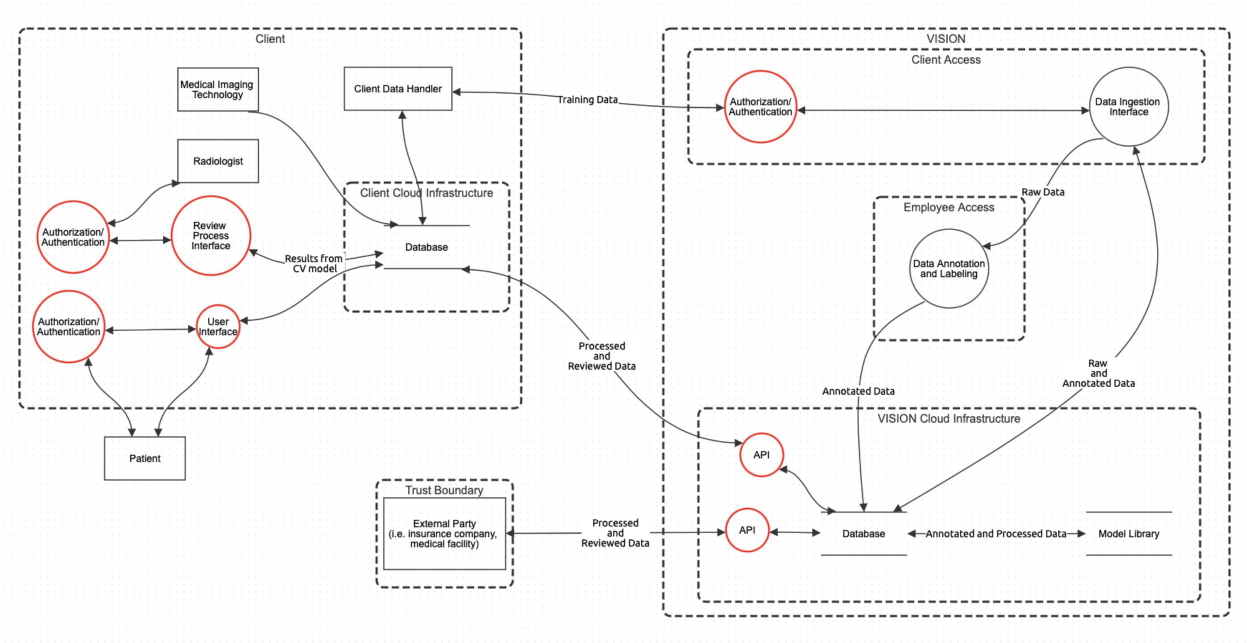
APPLICATION TO THE SCENARIO

1. **Spoofing:** Risks of unauthorized access to systems or impersonation of legitimate users within the AI model training and data processing environment.
2. **Tampering:** Potential for adversarial attacks or unauthorized modifications to AI models or data, impacting model accuracy and reliability.
3. **Repudiation:** Issues where users or systems deny performing actions related to AI data processing or model training, affecting accountability.
4. **Information Disclosure:** Risks of sensitive medical information being exposed through AI model outputs or data handling processes.
5. **Denial of Service:** Threats that could disrupt AI services or the availability of medical data and results.
6. **Elevation of Privilege:** Risks of unauthorized elevation of user privileges within the AI environment, leading to access beyond intended scopes.

APPLICATION DETAILS

- **AI Model Training:** Ensuring the security of the training data and the integrity of the models.
- **Data Processing and Storage:** Secure handling and storage of data processed by AI systems.
- **Result Communication:** Ensuring secure and accurate delivery of AI-generated results to clients and patients.

THREAT MODEL


Client
Medical Imaging Technology
Client Data Handler
Training Data
VISION
Client Access
Authorization/ Authentication
Data Ingestion Interface
Radiologist
Client Cloud Infrastructure
Authorization/ Authentication
Review Process Interface
Database
Results from CV model
Raw Data
Employee Access
Data Annotation and Labeling
Authorization/ Authentication
User Interface
Processed and Reviewed Data
Raw and Annotated Data
Annotated Data
VISION Cloud Infrastructure
Patient
API
Trust Boundary
External Party (i.e. insurance company, medical facility)
Processed and Reviewed Data
API
Database
Annotated and Processed Data
Model Library

This threat model represents the actors, processes, data flows, and trust boundaries involved in VISION's business with their client in the healthcare industry. Starting with the Client side of the model there is the Client Data Handler. This is the actor who interacted with the Data Ingestion Interface application (process) provided by VISION. They then need access to the Client's Cloud Infrastructure (trust boundary), to access their database. They then send these medical images to VISION through the interface and VISION's model is trained on them. Next, is the Medical Imaging Technology required to capture these raw images and send them to the database in the Client's cloud. Next, there is a Radiologist (or other specialist), who will go through the Authentication/Authorization process to gain access to the Review Process Interface where they can access the database where raw and processed data is stored, in order to review the results generated by VISION's Computer Vision (CV) model. There is also an Authentication/Authorization process and user interface by which patients can gain access to their medical data stored in their database.

Continuing to VISION's side of the threat model, there is a trust boundary around the resources that only clients are allowed to access. These resources include Authentication/Authorization process and Data Ingestion Interface application (process) required for the Client Data Handler(s) to submit data for the model to train on. The data submitted through this interface is then sent to another trust boundary that only an employee can access where they conduct Data Annotation and Labeling activities. Through this interface they can also view raw and annotated data previously submitted. Next there is another trust boundary for VISION's Cloud Infrastructure that contains VISION's database for Client data, a database for their Model Library, and API's that communicate data to other applications outside of VISION's network. VISION's cloud resources are contained within a trust boundary, because it is necessary to apply the Principle of Least Privilege, and make sure that only the employees who need to have access to them do. Both of these databases send data back and forth. The database sends annotated data to either train the model, or for the model to process and the model library sends back the processed data. One of the API's is responsible for sending *select* data to External Parties (such as, insurance companies, or other medical facilities) and the other is for sending processed and reviewed data back and forth between VISION's database and the Client's database. All of the data flows represented in this threat model are secured using encryption protocols such as HTTPS or Transport Layer Security (version 1.3).

COMPENSATING CONTROLS

SPOOFING

- **Authentication and Authorization:** Implement strong multi-factor authentication (MFA) and role-based access controls (RBAC) for all systems involved in AI model training and data processing.
- **Identity Management:** Use identity and access management (IAM) tools to ensure that only authorized individuals have access to sensitive data and AI systems.

TAMPERING

- **Model Integrity Protection:** Apply cryptographic techniques to ensure the integrity of AI models and prevent unauthorized modifications. Implement monitoring for changes to model parameters and data.
- **Adversarial Training:** Incorporate adversarial training techniques to improve model robustness against malicious inputs designed to deceive the AI.

REPUDIATION

- **Comprehensive Audit Trails:** Establish detailed and immutable logging for all AI-related activities, including data processing, model training, and result generation. Regularly review logs to detect and investigate potential repudiation.
- **Accountability Mechanisms:** Implement mechanisms to attribute actions to specific users or systems to prevent and address repudiation issues.

INFORMATION DISCLOSURE

- **Data Encryption:** Encrypt sensitive data both in transit and at rest using strong encryption standards. Ensure that data handled by AI systems is protected to prevent unauthorized access or exposure.
- **Access Controls:** Implement strict access controls to limit who can view or handle sensitive data. Use data masking or anonymization techniques where appropriate.

DENIAL OF SERVICE (DoS)

- **Redundancy and Failover:** Design resilient infrastructure with redundancy and failover capabilities to maintain AI service availability during disruptions.
- **Rate Limiting:** Implement rate limiting and traffic management to protect AI systems from denial of service attacks and ensure stability under high loads.

ELEVATION OF PRIVILEDGE

- **Least Privilege Access:** Enforce the principle of least privilege to limit user access to only the resources and actions necessary for their roles. Regularly review and update permissions to prevent unauthorized privilege escalation.
- **Privilege Management:** Use privilege management tools to monitor and control elevated access permissions within AI systems.

RECOMMENDATIONS

Strengthen Data Security: Implement end-to-end encryption and robust access controls for all data used in AI model training and processing. Ensure that sensitive medical data is protected against unauthorized access.

Enhance AI Model Security: Use cryptographic validation to protect the integrity of AI models and incorporate adversarial training to bolster model resilience against attacks.

Improve Authentication Measures: Adopt MFA and IAM tools to secure access to AI systems and data, ensuring that only authorized personnel can perform sensitive actions.

Implement Comprehensive Logging: Enable detailed logging and monitoring of all AI-related activities to detect and address any potential repudiation or unauthorized actions.

Ensure Compliance: Regularly review and update security practices to comply with relevant regulations such as HIPAA. Conduct periodic audits to maintain adherence to privacy laws and security standards.

Develop Incident Response Plans: Create and maintain incident response plans tailored to AI-specific threats, including procedures for handling model attacks and data breaches.

CONCLUSION

The threat modeling analysis of VISION's healthcare imaging AI solution underscores the importance of addressing AI-specific security risks. By implementing the recommended controls and strategies, VISION can enhance the security of its AI systems and protect sensitive medical data from various threats. These measures will help ensure compliance with regulatory requirements and build trust with clients, while also maintaining the integrity and reliability of the AI solutions provided.