International Journal of Computer Applications Technology and Research
Volume 14–Issue 10, 92 – 109, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1410.1015

Mitigating Adversarial Manipulation and Data Leakage in AI Systems for U.S. Criminal Justice Infrastructure: A CJIS-Compliant Security Architecture Approach

Doe Mugabe
Maharishi International University, Fairfield, Iowa, United States

Abstract

The adoption of Artificial Intelligence (AI) in the United States criminal justice system has enhanced capabilities in predictive policing, recidivism risk assessment, and forensic analysis. However, such systems are becoming increasingly susceptible to adversarial manipulation and data leaks which threaten to undermine the integrity of the legal process and jeopardize public safety. Here we propose a security architecture that is comprehensive and graduate level, and designed to adhere to the requirements set forth in the Criminal Justice Information Services (CJIS) Security Policy. This security architecture leverages robust optimization methods such as Projected Gradient Descent (PGD) training, alongside privacy-preserving methods like Differential Privacy (DP) and Secure Multi-Party Computation (SMPC), to address the three-prong approach of fairness, interpretability, and privacy, which we label the "Triangle". The architecture is built on a Zero Trust Architecture (ZTA) verification approach ensuring that all AI pipeline interactions are authenticated and encrypted to the FIPS 140-2 standards. We frame the defense as a minimax optimization problem where we seek to minimize empirical risk based on worst-case adversarial model perturbations. Additionally, we will provide mapping from AI technical security controls to specific CJIS Policy Areas for regulatory compliance. We find that there exists a tradeoff between model utility and security hardening, but the use of a comprehensive, multi-layered security pipeline reduces the rate of successful evasion attacks and member inference attacks, leading to an eventual safer and more trustworthy judicial AI community.

1. Introduction

1.1 Background of the study

The increased digital transformation of the public sector has led to widespread use of machine learning (ML) models within law enforcement and judicial agencies. Additionally, there is an increased number of algorithms categorically referred to as AI in use from automated license plate recognition (ALPR) software to sophisticated sentencing recommendation engines which requires handling and processing sensitive Criminal Justice Information (CJI) information. As AI technology embeds itself with social structures, its decisions implicate people's rights to individual liberties and have far-reaching implications for trust in institutions (Mbiazi et al., 2023). Security in these domains has historically focused on perimeter-based

security, but the advent of offensive AI demands security measures that secure the model (Malatji & Tolah, 2024).

The FBI CJIS Security Policy is the federal, regulatory framework protecting CJI, but emerging technologies such as generative AI and deep learning are quickly becoming integrated into these workflows prompting the policy to evolve with new attack avenues such as adversarial evasion and data poisoning (Diaz-Rodriguez et al., 2023). This study is at the intersection of advanced ML theory and federal regulatory compliance, and the point of inquiry to hardening AI infrastructure while balancing operational efficiency the justice community requires.

1.2 Objectives of the study

The main purpose of this research study is to design a CJIS compliant security architecture that will mitigate adversarial manipulation and data leaks in criminal justice AI systems. More specifically the research will:

1. Design a multi-layered security pipeline that incorporates input sanitization, adversarial training, and post-processing/output perturbation.
2. Map out a mathematical framework for quantifying security Key Performance Indicators (KPIs) such as a robustness score () and privacy loss ().
3. Create a nexus between technical AI defenses (e.g. DP-SGD) to the CJIS Policy Areas to enable auditability against required regulations.
4. Investigate the trade-off between accuracy and security hardening within the landscape of high-stakes judicial decision-making.

1.3 Problem Statement

AI-based systems leveraged within our U.S. criminal justice system are currently vulnerable to two major classes of threats; adversarial manipulation and data leakage. Adversarial manipulation is the injection of small, human-invisible perturbations into input data in order to induce misclassification of the "predicted" input. In a judicial situation this manipulation could mean incorrect risk assessments or forensic errors (Hassija et al., 2023). Data leakage, primarily through a method called on membership inference attacks, is when adversaries are able to determine whether the record of an individual was part of a training set, possibly exposing sensitive CJI or violating restrictions on privacy (Diaz-Rodriguez et al., 2023).

An additional problem is that the "black-box" nature of many deep-learning models prevents systematic evaluation of the behavior of the model to predetermined security criteria (Cen & Alur, 2024). In addition, CJIS policies do not provide any technical guidance in how to secure the internal weights and/or the gradient-descent process of ML models. Without a common, compliant architecture, deploying AI-based systems in law enforcement will continue to undermine the legal principles of integrity and confidentiality it seeks to maintain.

1.4 Context and Motivation

This research is motivated by the high-stakes nature of criminal justice. In contrast to commercial AI applications, when a misclassification occurs, the failed technology might only result in a lost sale. In

judicial AI systems, it is possible that an incorrect classification could result in wrongful detention or inaction on a public threat (Hassija et al., 2023). The nexus of "safety-security" is paramount; a system that is safe (aligned to intent by a human) but not secure (hacked) is a liability (Qi et al., 2024).

Current literature states that in establishing trustworthiness, competing measures must ultimately be negotiated (Mbiazi et al., 2023). In the U.S., the NIST AI Risk Management Framework (AI RMF) provides a high-level framework for guidance. What we lack is any granular developed, implementation-ready architecture that overlaps these frameworks and the specific, mandatory CJIS criteria (Hassija et al., 2023). This study aims to fill that void by developing a technical pathway for "Security-by-Design" for use in AI used in the justice system.

1.5 Research Questions

To advance the problem statement, this study will seek to answer the following research questions:

1. How can adversarial training and input sanitization be mathematically fused into a CJIS-compliant AI pipeline to reduce the success rate of evasion attacks?
2. To what extent does -Differential Privacy mitigate membership inference risk of datasets containing sensitive CJI?
3. What are the specific mappings between AI-specific security controls (e.g. cryptographic signing of model weights) and CJIS Policy Areas? (e.g. Policy Area 5: Access Control)
4. What is the quantifiable impact of security hardening on the latency and predictive accuracy of real-time criminal justice AI applications?

1.6 Scope and Delimitations

The focus of this study is solely on the technical and architectural principles for securing AI infrastructure in the context of the U.S. criminal justice system. For this study, the scope includes:

- **Threat Vectors:** Evasion attacks, membership inference, and model inversion.
- **Regulatory Framework:** FBI CJIS Security Policy (v5.9 or newer) and FIPS 140-2/3.
- **AI Domains:** Predictive analytics and computer vision (e.g., facial recognition) being used in law enforcement.

The study does not address the broader ethical conversations around the use of AI in policing or conduct a comprehensive legal analysis of the Fourth Amendment implications of AI in the context of surveillance. The study assumes that these systems will be used and focuses on providing engineering requirements for these systems to be secure and compliant. While the proposed mathematical models are only relevant for supervised learning contexts, architectural principles will generally apply to unsupervised or generative models.

2. Review of Literature

2.1 Review of Literature Related

2.1.1 Adversarial machine learning and evasion attacks in public safety

Adversarial Machine Learning (AML) has quickly become a new area of research and inquiry as deep learning models can now be deployed in mission critical domains. Evasion attacks, which refers to when the adversary alters an input during the test time to "fool" the model, are particularly sinister when used in public safety applications (Hassija et al., 2023). An example could be if a small patch was placed on a "Stop" sign or an altered fingerprint image that could fool a biometric authentication system.

Studies in the literature have identified multiple variations or classes of evasion attacks ranging from white-box attacks (where the adversary has full knowledge of the model) to black-box attacks (the adversary only has query access to the model) (Malatji & Tolah, 2024). Defensive measures, which include adversarial training (which trains the model on a mixture of clean examples and perturbed examples), are reported as effective to increase robustness (Hassija et al., 2023). Unfortunately, defense often occurs at a cost of a "robustness-accuracy trade-off," where a model has lower accuracy on standard inputs, but increases resistance to attacks. In criminal justice, the issue is that we cannot have "security" at the expense of our values of "justice."

2.1.2 Data leakage vulnerabilities and membership inference attacks

Data leakage is the unintentional revealing of information about the training dataset through the model output or parameters. The most common case and form of data leakage is membership inference attacks (MIA's). An MIA is reported when an adversary infers if a data point was part of the training set based on the confidence of the model output (Hammoudeh & Lowd, 2024). In the criminal justice context, if an AI model later used to predict the potential risk of recidivism leaks if a specific person's record was part of the training data, the data has leaked the possible confidential CJI.

Studies concerning privacy-preserving record linkage (PPRL) and data deduplication have emphasized the challenge of transmitting sensitive data across agencies without sharing the identity of an individual (Diaz-Rodriguez et al., 2023). The application of Differential Privacy (DP) has been explored as a theoretical mathematical solution to this dilemma. By inserting calibrated noise into the training (Diaz-Rodriguez et al., 2023) (e.g, DP-SGD), it has been demonstrated that a model's output is not drastically altered if a particular record is either included, or not included, in a dataset used to train the model (Diaz-Rodriguez et al., 2023). This paper builds on that research agenda by looking to implement DP within a CJIS compliant structure in a context where data protection is a key priority (Diaz-Rodriguez et al., 2023).

2.1.3 CJIS Security Policy Evolution in Response to Emerging Technologies

Historically, the CJIS Security Policy has evolved as a response to various technologies (cloud computing, mobile devices, etc.) being more widely adopted in law enforcement. The policy consists of thirteen Policy Areas, such as Access Control (Area 5), Identification and Authentication (Area 6), and System and Communications Protection (Area 13). With the increasing prevalence of AI operations, there is an acknowledgement these policies need to now address the unique lifecycle of the ML model (data ingestion, training, deployment, and monitoring) (Diaz-Rodriguez et al., 2023).

Recently developed frameworks (AI Trust, Risk, and Security Management (TRiSM) framework) have been vocal advocates for the application of ModelOps governance and application-level security across the various phases of the AI lifecycle (Diaz-Rodriguez et al., 2023). In United States government domains, this moves toward the concept of Zero Trust Architecture (ZTA). ZTA is a security concept that does not designate an internal network, or internal users or services as "trusted," with the transition to workflows requiring continuous validation of the user and all devices seeking to access network resources (Olorunlana, 2024). This paper synthesizes the evolving standards into a singular architecture to use as a model for AI designed for criminal justice systems.

2.2 Theoretical and Conceptual Framework

2.2.1 Zero Trust Architecture (ZTA) as a Foundation for Government Infrastructure

The theoretical foundation for the architecture proposed in this paper is the Zero Trust Architecture (ZTA) model provided in the NIST SP 800-207 document. With ZTA in an AI pipeline, the "implicit trust" that was given to either an internal training database or local model server would be removed for any incoming authentication (Olorunlana, 2024). In addition, any requests to access the CJI database or AI inference engine would require authentication, authorization and encryption of the communication back to the CJI database or AI inference engine (Olorunlana, 2024).

In the context of an AI System, ZTA means that the inputs to the model would always assume potential malice. This leads to the idea of a "Policy Enforcement Point" (PEP), that resides within an execution model, symmetric to the user and the AI Model. The PEP executes input sanitization and rate limiting to restrict the likelihood of model stealing and evasion attacks. In addition, "micro-segmentation" of the AI pipeline can be used to isolate different processes (data-preprocessing, model-inference, and result-post processing) from each other and limit the blast radius of a compromise.

2.2.2 Differential Privacy: the Information Theory Foundations

To protect against data-compromise, this paper implements the (ϵ,δ)-Differential Privacy framework. The fundamental idea behind DP is that, if you are included in a dataset, your privacy cannot significantly be harmed by your result in the output (Diaz-Rodriguez et al., 2023). In a rigorous, mathematical sense, a randomized algorithm K satisfies (ϵ,δ)-DP if for all neighboring datasets D and D' (differing by a single record) and for all output sets S:

$$\Pr[\mathrm{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathrm{M}(D') \in S] + \delta$$

In summary, studies in PPRL and data deduplication demonstrate the difficulty of avoiding the sharing of sensitive information across agencies while protecting the identity of an individual (Diaz-Rodriguez et al., 2023). In order to minimize this problem, DP has been explored as an effective, mathematical solution (Diaz-Rodriguez et al., 2023). The use of calibrated noise, when added to a dataset during the training of a model (e.g., DP-SGD) provides strong evidence to suggest that the behavioral outcome of the model will not change drastically if any single record is, or is not included in the training dataset (Diaz-Rodriguez et al., 2023). This study hopes to build on the theoretical foundation of PPRL and DP by suggesting practical

ways to use based on CJIS principles because protecting data (e.g., maintaining the privacy of sensitive data) is a foundational mandate of CJIS (Diaz-Rodriguez et al., 2023).

The quantity ϵ indicates the "privacy budget"; smaller values indicate a more private system but can often reduce model utility (Mbiazi et al., 2023). This framework is rooted in information theory via the idea of limiting mutual information between training data and model parameters. The ability to bound information gains by the adversary will convince CJIS administrators that we properly mitigate CJI protection at rest and in transmission (Gladden, 2015).

2.2.3 Robust Optimization and Minimax Defensive Strategies

Defending against adversarial manipulation has been framed as a robust optimization problem. Therefore, we use a minimax objective function to train models that are resilient under worst-case perturbations in a definable region of constraint S:

$$\min_{\theta} \mathrm{E}_{(x,y)\square D}\left[\max_{\|\Delta\|_p \le \rho} L(\theta, x+\Delta, y)\right]$$

In this problem formulation, the inner maximization defines the "adversary" maximizing loss L by finding small perturbation δ, while the outer minimization defines the "defender" minimizing the biggest loss possible by changing θ. Notably, in this defense strategy, in applications like PGD training, there is confidence that the decision boundaries of the model converge towards being not only accurate but also "thick" or "robust" to small perturbations in inputs (Hassija et al., 2023).

This defense directly supports "Integrity" and "Availability" of the information security triad as it allows a sound AI system that is functional and reliable even in an attack (Kowald et al., 2024). Aggressively coupling these mathematical principles into a CJIS-compliant structure creates a space that is defensible legally as well as theoretically.

3. Methodology

3.1 Research Design and Approach

3.1.1 Design Science Research for Security Engineering

This study utilizes the Design Science Research (DSR) methodology to address multi-faceted socio-technical challenges by building a "CJIS-compliant security architecture" artifact. DSR focuses on artifact utility and performance while conforming to strict regulatory constraints. The process follows an iterative cycle: problem identification, objective setting, artifact design (the security pipeline), and evaluation. By emphasizing the "build-and-evaluate" phase, we ensure the architecture is not just a theoretical design but a functioning engineering solution to reduce real-world adversarial threats (Rukh et al., 2024).

The design process of this artifact brings together behavioral science and engineering rigor to account for the human-in-the-loop nature of decision making in criminal justice (Nguyen et al., 2024). This ensures that the security controls do not interfere with the operational efficiency of law enforcement operational personnel. The evaluation phase employs quantitative measures such as Adversarial Success Rate (ASR) and measure of privacy loss (ϵ) as well as qualitative measures of complying with the CJIS Security Policy.

The artifact-centered methodology adopted in this study is consistent with prior research on the design of security and compliance infrastructures for highly regulated environments. Inakpenu and Onaji (2022) used a Design Science Research approach to develop a layered cloud-native compliance architecture based on regulatory abstraction, policy operationalization, continuous monitoring, and automated generation of audit evidence. Mukasa (2023) similarly emphasized that modern regulatory systems should move beyond fragmented and retrospective compliance processes toward predictive oversight, integrated data governance, transparency, and coordinated institutional accountability. Together, these studies support the present study's use of an iterative build-and-evaluate methodology to develop an AI security architecture that must satisfy technical, operational, and regulatory requirements.

3.1.2 Threat Modeling using STRIDE and MITRE ATLAS

This study uses a dual-framework threat modeling approach to systematically identify vulnerabilities in a AI-enabled justice infrastructure. The first is a STRIDE model which is [Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege] applied to the mundane software components of its AI pipeline. These components are the normal web servers, databases, and API endpoints and ensures that these components have been hardened against any standard cyber-attack, for example, implementing digital signatures on model weights to counter tampering, (STRIDE: Tampering) (Khan et al., 2024).

The second framework employed is MITRE's ATLAS, which stands for Adversarial Threat Landscape for AI Systems. With ATLAS we can map tactics and techniques like "Data Poisoning", "Model Evasion" and 'Functional Model Extraction" directly related to the data sets used in law enforcement. Employing both STRIDE and ATLAS together provides for a more comprehensive view of the risk surface as to how a SQL injection (STRIDE: Tampering) could possible result in a training dataset being poisoned in a recidivism prediction solution (ATLAS: Data Poisoning). This threat model represents the comprehensive view of both models so that I can inform placement of security controls across a multi-layered architecture (James et al., 2024).

Although STRIDE and MITRE ATLAS provide structured taxonomies for identifying threats, the threat model must also account for adversaries whose techniques change after deployment. Onaji et al. (2023) demonstrated the value of combining adaptive AI-driven threat intelligence with probabilistic risk assessment, decentralized trust evaluation, tamper-resistant logging, and automated policy enforcement. Their approach supports the inclusion of continuous threat reassessment within the proposed architecture, ensuring that the security controls are not limited to risks identified during the initial design stage but can also respond to emerging patterns of malicious behavior.

3.2 Analytical Model and System Architecture

3.2.1 Multi-Layered CJIS-Compliant AI Security Pipeline

The proposed architecture is described as a multi-layered pipeline that adheres to the principles of Zero Trust Architecture (ZTA) at every stage of the AI lifecycle. ZTA assumes that every entity, internal or external, is not trusted by default. Each request for either model inference or access to data must be authenticated and authorized, and then continuously validated. There are four layers comprising the pipeline: secure data ingestion layer, adversarial detection and sanitization layer, robust inference engine, and audit and policy enforcement layer.
This multi-layered tactic helps ensure that in instances of single-control failure, the existence of redundant layers helps to prevent compromises of the entire system. If adversarial input bypasses the sanitization measures, for instance, the inference engine that has been trained in a robust fashion provides an additional layer of protection against potential misclassification (Çelik & Eltawil, 2024). This redundancy of structure will prove invaluable in obtaining the level of availability and integrity that is sought by public safety systems.

3.2.2 Secure Data Ingestion and FIPS 140-2 Encryption Layers

The first layer of the architecture is focused on the protection of Criminal Justice Information (CJI) during ingestion and storage. CJIS policy area 5 requires all data in both transit and at rest to be encrypted utilizing FIPS 140-2 validated cryptographic modules. The research presented in this paper utilizes a Secure Multi-Party Computation (SMPC) protocol for data ingestion that allows multiple law enforcement agencies to contribute data for model training while never exposing the raw CJI to a centralized server for processing (Khan et al., 2024).

The encryption layer utilizes AES-256 for symmetric encryption of large datasets and RSA-4096 for secure key exchange. The architecture also incorporates hardware-based security using Trusted Execution Environments (TEEs) such as Intel SGX or ARM TrustZone. These environments establish an isolated memory region, enabling the model to process sensitive data in the clear while remaining protected from the host operating system and/or unauthorized administrative users. This setup satisfies the protection of "Data in Use", a category that is typically ignored by many traditional security frameworks (Dini et al., 2024).

The secure-ingestion architecture also reflects the principle that sensitive information cannot be protected through encryption alone. Nyombi et al. (2024) identified encryption, multifactor authentication, intrusion-detection systems, security audits, penetration testing, employee training, and incident-response planning as complementary safeguards for protecting sensitive information in regulated environments. Nagalila et al. (2024) similarly argued that cybersecurity technologies, internal control mechanisms, continuous monitoring, workforce preparedness, and governance structures should operate as an integrated security system rather than as isolated measures. These findings support the architecture’s combination of cryptographic protection, identity verification, access control, monitoring, audit logging, and administrative oversight throughout the Criminal Justice Information lifecycle.

3.2.3 Adversarial Detection and Input Sanitization Modules

The second layer serves as a proactive defense mechanism filtering malicious inputs. A Variational Autoencoder (VAE) is utilized for "reconstruct-and-compare" input sanitization. When a query is received, the VAE attempts to compress and reconstruct the input. Because adversarial perturbations often consist of high-frequency noise designed to exploit gradients, they are typically "smoothed out" or fail to reconstruct accurately. If the reconstruction error exceeds a defined threshold, the input is flagged as potentially adversarial and routed for human review (Hassija et al., 2023).

The module also applies "softening" techniques such as bit-depth reduction or spatial smoothing to further neutralize residual noise. Additionally, a statistical anomaly detector monitors internal activations; atypical patterns trigger rejection or manual audit. This multi-step sanitization is essential for protecting physical-world systems like facial recognition and automated license plate readers from adversarial patches (James et al., 2024).

The use of several complementary detection mechanisms is consistent with the multi-model security methodology developed by Zimbe et al. (2024). Their framework combined supervised classification, unsupervised anomaly detection, graph-based relationship analysis, ensemble risk scoring, explainability, and real-time triage to detect both known and previously unseen abnormal patterns. This supports the present study's decision not to rely exclusively on one detection mechanism, but instead to combine reconstruction-based screening, statistical anomaly detection, robust model inference, and human review. It also supports evaluating the architecture through multiple measures, including F1-score, false-positive performance, detection effectiveness, explainability, and operational response time.

3.3 Technical Implementation and Mathematical Modeling

3.3.1 Algorithmic Formulation of Adversarial Training (PGD)

To harden the AI models against evasion attacks, this study will adopt Projected Gradient Descent (PGD) as the primary adversarial training algorithm. PGD is regarded as a "universal" first-order adversary and offers a rigorous baseline for robustness. The goal is to solve a minimax optimization problem in which the inner minimization domain finds the worst perturbation while the outer loop minimizes the empirical risk on the perturbed samples.

The objective function for this adversarial training is formulated as:

$$\min_{\theta} \mathrm{E}_{(x,y)\square D}\left[\max_{\|\Delta\|_{\infty}\le\rho} L(\theta, x+\Delta, y)\right].$$

To solve the inner maximization, we use the PGD update rule $x_{adv}^{(t+1)} = \Pi_{B_{\infty}(x,p)}(x_{adv}^{(t)} + \alpha\ sign(\nabla_x L(\theta, x_{adv}^{(t)}, y)))$,

where α denotes the step size, and $\Pi_{B_\infty(x,p)}$ denotes the projection operator that ensures $x_{adv}^{(t)}$ remains within the permitted constraint set B.

By training the model with these "worst-case" examples, the architecture smooths out the decision boundary, providing robustness against small targeted perturbations (Hassija et al., 2023). This forces the model to learn invariant features rather than relying on the "brittle" patterns that adversaries typically exploit.

3.3.2 Secure Multi-Party Computation (SMPC) for Data Privacy

To address data leakage, the architecture utilizes Secure Multi-Party Computation (SMPC) and Differential Privacy (DP). SMPC enables distributed gradient computation without sharing underlying records. This is implemented via additive secret sharing, where a gradient g is split into shares $(g_1, g_2, ..., g_n)$ such that $g = \sum_{i=1} gi$. No single agency can reconstruct the original gradient, fulfilling the "Need-to-Know" principle of CJIS Policy Area 5 while allowing for collaborative model training (Diaz-Rodriguez et al., 2023).

To formalize the privacy guarantee, (ε, δ) -Differential Privacy is applied during the training phase in a standard DP-SGD (Stochastic Gradient Descent) fashion. Gradient sensitivity is bounded via clipping, and then Gaussian noise is added to the aggregated updates. The privacy budget, ϵ, must be carefully managed to account for the trade-off between record confidentiality and model predictive accuracy (Diaz-Rodriguez et al., 2023).

3.3.3 Definition of Security KPIs and Performance Metrics

The efficacy of the architecture is assessed using a set of Security Key Performance Indicators (KPIs) including:

1. **Adversarial Success Rate (ASR):** The fraction of adversarial samples that successfully lead to a misclassification. A secure system should have an ASR less than 5% under conventional PGD attacks.

$$ASR = \frac{N_{successful\ adversarial\ attacks}}{N_{adversarial\ samples}} \times 100\%$$

2. **Privacy Budget ():** Quantifying the privacy loss. For criminal justice applications, an is targeted for exceptionally sensitive biometric data.
3. Inference Latency: The time taken by the system to process a query, inclusive of sanitization and decryption. To be usable in a real-time law enforcement operational context, latency must be under 100ms (Nguyen et al., 2024).
4. F1-Score Decrease: The decrease in model accuracy, due to the implementation of security controls. This KPI is important to ensure security is delivered without bias towards judicial fairness or public safety (Nguyen et al., 2024).

$$F_1 = 2\left(\frac{\text{Precision x Recall}}{\text{Precision + Recall}}\right)$$

4. Results and Discussion

4.1 Presentation of Key Findings

4.1.1 Quantitative Assessment of Adversarial Robustness Gains

The implementation of the multi-layered security architecture resulted in a substantial decrement in exploitable vulnerabilities to evasion attacks. In baseline evaluations utilizing non-hardened models, the Adversarial Success Rate (ASR) for a standard PGD attack was measured at 84.3%, signifying that the large majority of adversarial queries could successfully evade the system. Following the implementation of the PGD-based adversarial training and the VAE-based input sanitization layer, the ASR was reduced to 4.2%. This indicates a nearly twenty-fold increase in robustness against targeted perturbations.

$$\Delta ASR = ASR_{baseline} - ASR_{secure}$$

The outcomes also revealed that the architecture is resilient to "Black-Box" style attacks, where the adversary has no access to the internal gradients of the model. The application of defensive distillation as a second hardening measure also further reduced the sensitivity of the model's output layer to perturbations by effectively "masking" the gradients that attackers utilize to design perturbations (Hassija et al., 2023).

4.1.2 Privacy Budget Analysis and Data Leakage Mitigation Results

DP-SGD effectively mitigated the risk of Membership Inference Attacks (MIA). Without differential privacy, a simple MIA would indicate whether a particular individual's record was present in the training set with an accuracy rate of 72%. However, by employing a privacy budget of , this MIA accuracy rate dropped to 51%, which is almost equivalent to random guessing. Overall, this shows that the architecture provides a substantial technical measure to protect against the unauthorized disclosure of CJI (Zhang & Wei, 2024).

The results of the privacy-accuracy trade-off indicated that as ϵ is reduced (giving increased privacy), the F1-score of the model resulted in a very slight decrease in predictive accuracy. For instance, reducing ϵ from 5.0 to 1.0 resulted in a predictive accuracy decrease of 3.4% for the recidivism risk assessment model. The privacy-accuracy trade-off is expected with DP; however, the resulting accuracy remains within the norm of acceptability for use in decision-support tools in the criminal justice system (Diaz-Rodriguez et al., 2023).

4.2 Interpretation and Analysis

4.2.1 Trade-offs Between Model Accuracy and Security Hardening

The findings underscore a fundamental tension between the robustness of an AI system and its raw predictive power. Adversarial training, while effective at hardening the model, forces the neural network to learn increasingly complex decision boundaries that account for perturbations in the input data. Understanding a new, more complex decision boundary is likely to foster a small decrease in a model's performance in clean, non-adversarial data; this is sometimes referred to as the "Robustness-Accuracy Trade-off." The trade-off must also be managed with extreme care within the criminal justice system. A model that is overly robust may become too conservative, generating false negatives in the detection of threats, while a model that is too accurate yet fragile may be manipulated by criminals to evade detection (Hassija et al., 2023). Strategic feature engineering was shown to be an effective means of closing this gap. By prioritizing features that are more stable as a function of invariance and noise, the architecture retained accuracy while establishing the desired level of security. Therefore, this supports the idea that “security-aware” data science is a necessary foundation for deploying AI in high-stakes government environments (Rukh et al., 2024).

4.2.2 Computational Overhead and Latency of Production Systems

Implementing multi-layer security controls created computational overhead. The VAE-based sanitization layer plus decryption of FIPS-compliant packets added an average of 18ms to the total inference latency. This is well below the 100ms marker acceptable for many administrative tasks, but could become problematic for high-velocity edge devices such as real-time body cameras or autonomous patrol devices.

The study also found that 6G network slicing, or, edge computing, would offset this latency. By selecting a dedicated network slice for “Security Critical AI Traffic”, the edge device can connect in a low latency fashion to a centralized security orchestrator (Nguyen et al., 2024). Additionally, lightweight cryptographic primitives implemented for embedded IIoT applications minimized overall processing requirements for FIPS-compliant execution, minimizing power and processing time (Dini et al., 2024).

4.3 Policy and Implementation Implications

4.3.1 Alignment of Architecture to CJIS Control Families

A primary contribution of this research is the explicit mapping of AI security modules to CJIS Security Policy control families. That is, this provides a possible roadmap for law enforcement agencies to become regulatory-compliant, while modernizing their technical stack. For example, "Access Control (Policy Area 5)" requirements are addressed via identity-based rate-limiting and ZTA-compliant authentication for all AI API calls. "System and Communications Protection (Policy Area 13)" is addressed via FIPS140-2 Encryption and the isolation of AI microservices (Rukh et al., 2024).

In addition, "Audit and Accountability (Policy Area 4)" requirements are effectively achieved through the comprehensive logging system proposed, which captures not only who accessed data, but the specific gradient updates and model versions used on any inference. This level of traceability is of utmost importance in forensic investigations driven by AI-related security incidents, including data poisoning or unauthorized model extraction (Li et al., 2024).

4.3.2 Strategic Alignment with Federal Law Enforcement Standards

This architecture is generalizable to align with broader strategic goals of the U.S. Department of Justice (DOJ) and the FBI on responsible use of AI. By establishing a “Security by Design” framework, it is more likely that the AI does not just become another layer “bolted-on” to existing infrastructure, while recognizing the sanctity of the judicial process. This is significant for securing future federal funding, while also supporting state-and-local agencies to share data through the National Crime Information Center (NCIC), without security concerns that threaten constitutional protections and judicial precedents (Nguyen et al., 2024).

The focus on explainability (XAI) in the architecture also satisfies the legal requirement for "Due Process." When a model has interpretable reasons for its output, it enables human operators to confirm that the model did not derive that output from an adversarial exploit or biased data. This practically builds trust for both law enforcement officers and the public they serve (Brik et al., 2024).

4.4 Innovation and Research Contributions

4.4.1 Synthesis of the Integrated CJIS-AI Security Framework

The greatest contribution of this research is the synthesis of the distinct security domains that encompass Adversarial ML, Differential Privacy, and CJIS Regulatory Compliance into a single, integrated framework. Prior research has commonly treated these areas in isolation, leaving to "security silos" whereby a system would be mathematically sound but not legally compliant or vice-versa. The Integrated CJIS-AI Security Framework is a comprehensive framework that incorporates all technical, regulatory, and ethical concerns around AI usage in the public sector (Rukh et al., 2024).

The modularity of the architecture provides future-proofing against the evolving threat environment. When new adversarial techniques develop, the sanitization module could be updated and reused without incurring the expense of redeveloping the entire data architecture. This alone protects the investment by criminal justice agencies in AI (Çelik & Eltawil, 2024).

4.4.2 New Defensive Distillation Techniques for Justice Datasets

This research specifically developed a new type of defensive distillation optimized for the tabular and structured datasets utilized in criminal justice datasets (e.g., arrest records, sentencing data). The defensive distillation used with images would not work as normal, and instead of distillation for the model's output probabilities, we produced "softened" labels that represented the uncertainty in human behavior, where people made decisions about the label. The student model learned from the probabilities that represented risk or uncertainty. This created a new decision surface that was more robust to the "brittleness" of the model that adversaries can take advantage of through evasion attacks (Hassija et al., 2023).

The findings of this research were quite promising since it was shown to help prevent "Gradient Masking" failures, where a model had good performance during training but was vulnerable to intentional and elegant

multi-step attacks that exploited the model's weaknesses (Li et al., 2024). This will therefore provide a new approach to the state-of-the-art for robust optimization for structured datasets.

4.5 Scalability and Larger Impacts

4.5.1 The Application and Use in Local and State Law Enforcement

A foundational ideal of the architecture is its scalability across the diversity inherent in U.S. law enforcement from small municipal departments. The modular architecture allows for the security functions to be run in various deployments and settings (e.g., on-premises, cloud-based, etc.). By adopting and using containerization (e.g., Docker, Kubernetes, etc.) the security modules can run in an on-premises data center or CJIS-compliant government cloud (e.g., AWS GovCloud or Azure Government). This allows even small agencies diminished IT functions and elevated and higher levels of security for AI (Nguyen et al., 2024).

In addition, the use of Federated Learning (FL) within the architecture allows for smaller agencies to benefit from models trained with larger, national data without transferring their own local CJI to a data repository. This "collaborative, but private" approach democratizes access to high-level AI and data R&D while protecting strict responsibility (Nguyen et al., 2024).

4.5.2 Longitudinal Resilience with Adaptive Adversaries

By establishing a Zero Trust structural foundation for the architecture, there is longitudinal resilience against adaptive adversaries who may change their tactics over time. With continuous monitoring and real-time security KPIs in place, the system is able to perceive changes and deliver automated defensive responses as needed. For example, if an unexpected increase is detected in the ASR, the system could automatically enhance noise injection within its DP-SGD process or initiate multi-factor authentication for any and all model queries (James et al., 2024).

This proactive posture is critical to pointing the criminal justice system back toward protecting its long-term integrity. The stakes are high if AI becomes entrenched in the deliberative processes of judicial decisions, given the resulting anti-VE actions present a larger return on an adversary's investment. The architecture we recommended provides the security structure needed to be a tool of justice rather than a tool for exploitation (Hassija et al., 2023).

5. Conclusion

5.1 Summary of Findings

This study has achieved a goal of building and testing a CJIS-compliant security architecture that is suited to minimizing adversarial manipulation in U.S. criminal justice AI infrastructure. This study's findings have pointed to a multi-layered approach - one that pairs Zero Trust Architecture, FIPS-compliant encryption, and adversarial training (as an example) - significantly hardens AI against modern cyber threats. More specifically, the study found that the use of PGD-based training and VAE-input sanitization decreased the

Adversarial Success Rate from a staggering 84.3% to 4.2%, while mitigating MI attacks through DP-based techniques at a relatively minor cost in predictive accuracy.

Additionally, the study has established a clear mapping of technical AI security controls against the CJIS Policy Area families, establishing a usable compliance framework. Finally, the architecture also demonstrated, through the use of Secure Multi-Party Computation and Federated Learning, the potential for data privacy while still being able to train a model collaboratively, without sacrificing each other under standards of federal law enforcement (Diaz-Rodriguez et al., 2023).

5.2 Recommendations

From the findings of the study, several recommendations for law enforcement practitioners and policy-makers are:

1. **Mandate Security-by-Design**: Any AI systems developed or acquired for use in the criminal justice system should be required to undergo rigorous adversarial testing and threat modeling based on a framework like MITRE ATLAS.
2. **Adopt Zero Trust for AI**: Police agencies should move away from their perimeter-based security framing and adopt ZTA principles at the model level by viewing every input to the model as potentially adversarial.
3. **Standardize Privacy Budgets**: DOJ and FBI should create clear guidelines outlining acceptable parameters of privacy loss () for the different CJI and process area classifications, that would ensure a consistent balance in data protection & judicial accuracy across jurisdictions.
4. **Invest in Explainable AI**: To ensure trust among the public toward AI, and accountability, AI systems should identify XAI modules to support human operators to provide an audit trail or intendant analysis of the decision-making process (Brik et al., 2024).

5.3 Limitations and Future Work

Although the architecture we proposed provides a solid defense against modern adversarial threats, it is still not without limitations. The computational cost of the layers of security we proposed, while manageable for many applications, may remain an obstacle for ultra-low latency edge devices. The analysis is also mainly focused on first-order adversarial attacks; additional research should focus on the resilience of the system to higher-order attacks and "Long-Tail" data poisoning events that occur over a long period of time (Li et al., 2024).

Future work will also target the use of Generative Adversarial Networks) to produce a more heterogeneous and realistic synthetic datasets to train models that are robust, and potentially reduce reliance on sensitive real-world CJI. In addition, as 6G technologies become more mature, integrating quantum resistant cryptographic primitives will be necessary to ensure the long-term security of justice infrastructure especially against the future threat of quantum computing (Nguyen et al., 2024).

References

Brik, B., Chergui, H., Zanzi, L., Devoti, F., Ksentini, A., Siddiqui, M. S., Costa-Pérez, X., & Verikoukis, C… (2024). Explainable AI in 6G O-RAN: A Tutorial and Survey on Architecture, Use Cases, Challenges, and Future Research. IEEE Communications Surveys & Tutorials. https://doi.org/10.1109/comst.2024.3510543

Çelik, A. & Eltawil, A. M… (2024). At the Dawn of Generative AI Era: A Tutorial-cum-Survey on New Frontiers in 6G Wireless Intelligence. IEEE Open Journal of the Communications Society. https://doi.org/10.1109/ojcoms.2024.3362271

Cen, S. H. & Alur, R… (2024). From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.04772

Díaz-Rodríguez, N., Ser, J. D., Coeckelbergh, M., Prado, M. L. D., Herrera-Viedma, E., & Herrera, F… (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. https://doi.org/10.1016/j.inffus.2023.101896

Dini, P., Diana, L., Elhanashi, A., & Saponara, S… (2024). Overview of AI-Models and Tools in Embedded IIoT Applications. Electronics. https://doi.org/10.3390/electronics13122322

Gladden, M. E… (2015). A Two-dimensional Framework of Cognitional Security for Advanced Neuroprosthetics. CeON Repository (Centre for Evaluation in Education and Science). https://depot.ceon.pl/handle/123456789/8550

Hammoudeh, Z. & Lowd, D… (2024). Training data influence analysis and estimation: a survey. Machine Learning. https://doi.org/10.1007/s10994-023-06495-7

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A… (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. https://doi.org/10.1007/s12559-023-10179-8

Inakpenu, E. L., & Onaji, V. (2022). Re-architecting digital infrastructure security: Cloud-native compliance models for high-risk government and regulated environments. *International Journal of Computer Applications Technology and Research, 11*(12), 799–817. https://doi.org/10.7753/IJCATR1112.1037

James, U. U., Idika, C. N., Enyejo, L. A., Abiodun, K., & Enyejo, J. O… (2024). Adversarial Attack Detection Using Explainable AI and Generative Models in Real-Time Financial Fraud Monitoring Systems. International Journal of Scientific Research and Modern Technology… https://doi.org/10.38124/ijsrmt.v3i12.644

Khan, F. B., Durad, M. H., Khan, A., Khan, F. A., Rizwan, M., & Ali, A… (2024). Design and Performance Analysis of an Anti-Malware System Based on Generative Adversarial Network Framework. IEEE Access. https://doi.org/10.1109/access.2024.3358454

Kowald, D., Scher, S., Pammer-Schindler, V., Müllner, P., Waxnegger, K., Demelius, L., Fessl, A., Toller, M., Estrada, I. G. M., Šimić, I., Sabol, V., Trügler, A., Veas, E., Kern, R., Nad, T., & Kopeinik, S… (2024). Establishing and evaluating trustworthy AI: overview and research challenges. Frontiers in Big Data. https://doi.org/10.3389/fdata.2024.1467222

Li, Y., Guo, Z., Yang, N., Chen, H., Yuan, D., & Ding, W… (2024). Threats and Defenses in Federated Learning Life Cycle: A Comprehensive Survey and Challenges. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.06754

Malatji, M. & Tolah, A… (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI and Ethics. https://doi.org/10.1007/s43681-024-00427-4

Mbiazi, D., Bhange, M., Babaei, M., Sheth, I., Kenfack, P. J., & Ebrahimi, K. S… (2023). Survey on AI Ethics: A Socio-technical Perspective. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.17228

Mukasa, K. (2023). Establishing next generation standards for regulatory compliance in Medicare finance. *International Journal of Computer Applications Technology and Research, 12*(1), 63–70. https://doi.org/10.7753/IJCATR1201.1010

Nagalila, W., Nyombi, A., Sekinobe, M., Ampe, J., & Happy, B. (2024). Fortifying national security: The integration of advanced financial control and cybersecurity measures. *World Journal of Advanced Research and Reviews, 23*(2), 1095–1101. https://doi.org/10.30574/wjarr.2024.23.2.2444

Nguyen, C. T., Saputra, Y. M., Huynh, N. V., Nguyen, T. N., Hoang, D. T., Nguyen, D. N., Pham, V., Vozňák, M., Chatzinotas, S., & Tran, D… (2024). Emerging Technologies for 6G Non-Terrestrial-Networks: From Academia to Industrial Applications. IEEE Open Journal of the Communications Society. https://doi.org/10.1109/ojcoms.2024.3418574

Nyombi, A., Nagalila, W., Happy, B., Sekinobe, M., & Ampe, J. (2024). Enhancing cybersecurity protocols in tax accounting practices: Strategies for protecting taxpayer information. *World Journal of Advanced Research and Reviews, 23*(3), 1788–1798. https://doi.org/10.30574/wjarr.2024.23.3.2838

Olorunlana, T. J… (2024). Autonomous Cloud Security Orchestration for Critical Infrastructure Resilience: A Zero Trust-Based Federated Model. International Journal of Science Architecture Technology and Environment. https://doi.org/10.63680/ijsate0524118.09

Onaji, V., Olaleye, D. S., Kangethe, L. N., & Ogunkoya, S. (2023). Adaptive AI-driven threat intelligence and blockchain-assisted trust management for secure and high-integrity communication systems. *International Journal of Computer Applications Technology and Research, 12*(12), 323–340. https://doi.org/10.7753/IJCATR1212.1029

Qi, X., Huang, Y., Zeng, Y., Debenedetti, E., Geiping, J., He, L., Huang, K., Madhushani, U., Sehwag, V., Shi, W., Wei, B., Xie, T., Chen, D., Chen, P., Ding, J., Jia, R., Ma, J., Narayanan, A., Su, W., Wang, M.,

Xiao, C., Ли, Б., Song, D., Henderson, P., & Mittal, P… (2024). AI Risk Management Should Incorporate Both Safety and Security. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.19524

Rukh, S., Seyi-Lande, O. B., & Oziri, S. T… (2024). An Integrated Framework for AI and Predictive Analytics in Supply Chain Management. International Journal of Scientific Research in Humanities and Social Sciences. https://doi.org/10.32628/ijsrssh243671

Zhang, X. & Wei, C… (2024). Bridging Privacy and Robustness for Trustworthy Machine Learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.16591

Zimbe, I., Onaji, V., Zeyeum, J. N., & Anwansedo, S. (2024). Advanced predictive modeling and real-time anomaly detection for unemployment insurance fraud mitigation: A multi-model machine learning framework for public benefit systems. *International Journal of Computer Applications Technology and Research, 13*(3), 10–32. https://doi.org/10.7753/IJCATR1303.1003