International Journal of Technology, Management and Humanities Volume 12, Issue 2, 2026

Securing Offline Internal AI Systems: A Comprehensive Framework for Data Protection, Model Integrity, and Access Control in Air-Gapped Environments

Día Fayyad

Cybersecurity Department, Saudi Aramco; Jordanian Engineers Association, Saudi Council of Engineers

Abstract

The growing deployment of artificial intelligence systems within highly sensitive sectors such as defense, finance, and critical infrastructure has led to increased reliance on offline and air-gapped environments as a primary security measure. These isolated systems are traditionally perceived as inherently secure due to their lack of direct connectivity to external networks. However, this assumption is increasingly challenged by the emergence of sophisticated covert attack vectors capable of bypassing physical isolation through side-channel and out-of-band communication mechanisms. As a result, air-gapped AI systems remain vulnerable to data exfiltration, model manipulation, and unauthorized access.

This study addresses this critical gap by proposing a comprehensive and structured security framework specifically designed for offline internal AI systems. The framework focuses on three core dimensions: data protection, model integrity, and access control. A systematic synthesis of existing literature on air-gap attacks, established security models, and recognized cybersecurity standards is employed to inform the framework design. The proposed architecture integrates classical security principles, including confidentiality, integrity, and role-based access control models, with AI-specific protection mechanisms such as secure model validation, controlled data pipelines, and layered access enforcement.

The findings demonstrate that the proposed multi-layered framework significantly enhances system resilience against covert exfiltration techniques, insider threats, and model compromise. It further highlights that reliance on physical isolation alone is insufficient for securing modern AI systems. Therefore, a proactive, defense-in-depth strategy is essential for safeguarding offline AI infrastructures against evolving threat landscapes.

Keywords: Air-gapped systems, Offline AI security, Data protection, Model integrity, Access control, Covert channel attacks, Cybersecurity framework.

International Journal of Technology, Management and Humanities (2026) 10.21590/ijtmh.12.02.01

Introduction

Background

Artificial Intelligence (AI) systems have rapidly evolved from experimental computational tools into mission-critical components across highly sensitive and classified environments. Sectors such as defense, intelligence, finance, and critical infrastructure increasingly rely on AI-driven decision-making systems to enhance operational efficiency, predictive capabilities, and automation. As these systems process highly confidential data, their deployment environments have shifted toward more controlled and isolated infrastructures to minimize exposure to external threats.

One notable trend is the growing adoption of offline or air-gapped AI systems, which are physically isolated from external networks, including the internet. These systems are widely perceived as a robust security measure, particularly in environments where data confidentiality is paramount. Air-gapped architectures are commonly deployed in military intelligence systems, secure financial processing units, and industrial control systems, where even minimal data leakage can have severe consequences.

However, the foundational concept of computer security extends beyond physical isolation. According to Bishop (2003), computer security is fundamentally concerned with

Corresponding Author: Día Fayyad, Cybersecurity Department, Saudi Aramco; Jordanian Engineers Association, Saudi Council of Engineers, e-mail: Dia.fayyad@gmail.com

How to cite this article: Fayyad D. (2026). Securing Offline Internal AI Systems: A Comprehensive Framework for Data Protection, Model Integrity, and Access Control in Air-Gapped Environments. *International Journal of Technology, Management and Humanities,* 12(2), 1-12.

Source of support: Nil

Conflict of interest: None

preserving three core principles: confidentiality, integrity, and availability. Confidentiality ensures that sensitive data is accessible only to authorized entities, integrity guarantees that data and systems remain unaltered and trustworthy, and availability ensures that systems remain functional and accessible when needed. While air-gapped systems inherently support confidentiality by limiting external access, they do not automatically ensure integrity or availability, especially in the presence of sophisticated internal or indirect threats.

As AI systems become more complex, incorporating large datasets, advanced models, and automated pipelines, the attack surface within even isolated environments expands. This evolution necessitates a deeper examination of how traditional security principles apply to modern offline AI systems and whether existing approaches are sufficient to address emerging risks.

Problem Statement

Despite their widespread adoption, air-gapped systems are often built on the false assumption of absolute security. The belief that physical isolation alone can prevent cyber threats has been increasingly challenged by recent research demonstrating the feasibility of covert and side-channel attacks. These attacks exploit non-traditional communication pathways such as electromagnetic emissions, thermal variations, acoustic signals, and even environmental systems to exfiltrate sensitive data or inject malicious inputs.

The sophistication of such attacks has grown significantly, enabling adversaries to bypass physical isolation without requiring direct network connectivity. This creates a critical vulnerability in offline AI systems, which often process high-value data and operate in environments where traditional monitoring tools are limited or absent. Furthermore, insider threats, misconfigurations, and compromised hardware components introduce additional layers of risk that are not mitigated by air-gapping alone.

A significant gap in current cybersecurity research is the lack of AI-specific security frameworks tailored for offline environments. While traditional security models provide a strong theoretical foundation, they were not originally designed to address the unique characteristics of AI systems, such as model training pipelines, data dependencies, and inference processes. As a result, existing approaches often fail to comprehensively address issues such as model tampering, data poisoning, and unauthorized access within air-gapped AI infrastructures.

This disconnect between traditional security assumptions and modern threat realities highlights the urgent need for a more holistic and adaptive security framework.

Research Objectives

The primary objective of this research is to develop a comprehensive and structured framework for securing offline internal AI systems operating in air-gapped environments. The proposed framework aims to bridge the gap between classical cybersecurity principles and the unique requirements of AI-driven systems.

Specifically, this study focuses on three critical dimensions of security

- Data Protection: Ensuring that sensitive datasets used for training and inference are securely stored, transmitted within the system, and protected against unauthorized access or leakage.
- Model Integrity: Safeguarding AI models from tampering, unauthorized modifications, and adversarial manipulation, thereby preserving their reliability and trustworthiness.
- Access Control: Implementing robust mechanisms to regulate user access, enforce least-privilege policies, and prevent unauthorized interactions with system components.

By addressing these dimensions, the research seeks to provide a unified approach that enhances the overall security posture of offline AI systems.

Research Questions

To guide the development of the proposed framework, this study is structured around the following key research questions

How can air-gapped AI systems be compromised despite physical isolation?

This question explores the various attack vectors and vulnerabilities that exist within supposedly secure environments, including covert channels and insider threats.

What security models are most effective for protecting offline AI systems?

This examines the applicability of established security models, such as confidentiality and integrity frameworks, in the context of AI systems.

How can a unified framework mitigate emerging threats in air-gapped environments?

This focuses on integrating multiple security approaches into a cohesive architecture capable of addressing both traditional and AI-specific risks.

These questions provide a structured foundation for analyzing existing challenges and developing targeted solutions.

Contribution

This research makes several significant contributions to the field of cybersecurity and AI system protection. First, it presents an integrated approach that combines classical security models with AI-specific protection mechanisms, thereby extending the applicability of established theories to modern technological contexts. By aligning principles such

as confidentiality, integrity, and controlled access with AI system requirements, the study offers a more comprehensive perspective on security design.

Second, the research introduces a novel framework specifically tailored for air-gapped AI systems, addressing critical gaps in current literature. Unlike traditional approaches that focus solely on network-based threats, the proposed framework accounts for covert channels, internal vulnerabilities, and model-specific risks.

Finally, the study provides a practical and standards-aligned security architecture that can be implemented in real-world environments. By incorporating best practices and aligning with recognized security standards, the framework ensures both theoretical robustness and practical applicability. This positions the research as a valuable resource for organizations seeking to deploy secure AI systems in highly sensitive and isolated settings.

Theoretical Foundations and Security Models

A robust security framework for offline internal AI systems must be grounded in well-established theoretical models that address confidentiality, integrity, and controlled access. These foundational principles remain highly relevant, even as AI systems introduce new dimensions of risk such as model poisoning, unauthorized inference, and covert data leakage. By integrating classical security theories with modern AI considerations, a more resilient and adaptable protection architecture can be achieved.

Confidentiality Models

Confidentiality remains a core pillar of information security, particularly in air-gapped AI environments where sensitive datasets and trained models must be strictly protected from unauthorized disclosure. The Bell-LaPadula model, introduced by Bell and LaPadula (1973), provides a formal framework for enforcing confidentiality through mandatory access control policies. Its primary rules, commonly summarized as "no read up" and "no write down," ensure that users can only access information at or below their security clearance level, while preventing the leakage of sensitive data to lower classification levels.

In the context of offline AI systems, this model is highly applicable to controlling access to training datasets, model parameters, and inference outputs. For example, high-sensitivity datasets used in defense or financial AI applications must be restricted to authorized personnel and processes. Additionally, the Bell-LaPadula model helps prevent inadvertent data exposure during model deployment or transfer between system components. However, while effective for confidentiality, the model does not inherently address data integrity or dynamic AI workflows, necessitating complementary mechanisms.

Integrity Models

Integrity is critical in AI systems, where even minor alterations to data or model parameters can significantly impact outcomes. The Biba model, proposed by Biba (1977), focuses on preserving data integrity through rules that are essentially the inverse of Bell-LaPadula. Its principles, "no read down" and "no write up," are designed to prevent contamination of high-integrity data by lower-integrity sources.

Within offline AI environments, the Biba model is particularly relevant for safeguarding training pipelines and model updates. For instance, ensuring that only verified and high-integrity data sources contribute to model training helps prevent data poisoning attacks. Similarly, restricting lower-trust processes from modifying critical model components preserves the reliability of AI outputs. This is especially important in air-gapped systems where external validation may be limited, making internal integrity controls essential. Nonetheless, strict enforcement of Biba policies can sometimes reduce system flexibility, requiring careful adaptation in AI workflows.

Access Control Models

Effective access control is essential for managing interactions between users, processes, and AI system components. The Role-Based Access Control model, as described by Sandhu et al. (2002), provides a scalable and flexible approach by assigning permissions based on roles rather than individual users. This simplifies administration and ensures that access rights align with organizational responsibilities.

In offline AI systems, RBAC can be used to define roles such as data engineers, model developers, system administrators, and auditors, each with specific permissions. This minimizes the risk of unauthorized access and supports the principle of least privilege. Complementing RBAC, the Clark-Wilson model, introduced by Clark and Wilson (1987), emphasizes integrity through well-formed transactions and separation of duties. It ensures that only authorized and validated operations can modify critical data, which is particularly valuable in maintaining the consistency of AI models and datasets.

Together, RBAC and Clark-Wilson provide a comprehensive approach to access and integrity control, balancing flexibility with strict enforcement of security policies.

Security Engineering Principles

Beyond specific models, broader security engineering principles play a crucial role in designing resilient AI systems. The concept of defense-in-depth, as explained by Anderson (2010), advocates for multiple layers of security controls to mitigate risks at different levels. In air-gapped AI environments, this includes physical isolation, hardware protections, secure software configurations, and monitoring mechanisms to detect anomalies.

Establishing clear system trust boundaries is equally important. These boundaries define the limits of trusted

components and help prevent the spread of potential compromises within the system. For example, isolating AI training environments from inference systems reduces the risk of cross-contamination.

Additionally, cybersecurity best practices and standards, as outlined by Stallings (2018), provide practical guidance for implementing secure systems. These include regular auditing, secure configuration management, incident response planning, and continuous monitoring. When applied to offline AI systems, these practices enhance resilience against both internal and external threats, including covert channel attacks and insider misuse.

2.5 ISO 27001 Alignment

Aligning security frameworks with established standards such as ISO/IEC 27001 ensures a structured and comprehensive approach to information security management. As reviewed by Ganji et al. (2019), ISO 27001 provides guidelines for developing, implementing, and maintaining an Information Security Management System (ISMS). It emphasizes risk assessment, policy development, and continuous improvement.

For offline AI systems, ISO 27001 alignment supports the integration of security controls across all layers, from data handling and model development to access management and system monitoring. It also facilitates compliance with regulatory requirements and enhances organizational credibility. By incorporating ISO 27001 principles, the proposed framework ensures that security measures are not only technically robust but also systematically managed and continuously refined.

Air-Gapped Systems and Threat Landscape

Concept of Air-Gapped Systems

Air-gapped systems are computing environments that are physically isolated from external networks, particularly the internet, to prevent unauthorized access and data leakage. This isolation is achieved by eliminating all direct and indirect network connections, thereby creating a controlled environment where data exchange occurs only through authorized physical means such as removable media. The fundamental operational assumption behind air-gapped systems is that physical separation inherently guarantees a high level of security by minimizing exposure to remote cyber threats.

In practice, air-gapped systems are widely deployed in highly sensitive domains where confidentiality and integrity are critical. These include intelligence agencies, military infrastructures, nuclear facilities, and financial institutions handling classified or mission-critical data. As highlighted by Viggh and Tsagaratos (2026), the resurgence of air-gapped AI systems is driven by the increasing risks associated with cloud-based large language models and external data dependencies. Organizations are increasingly adopting offline AI systems to ensure that proprietary data, strategic models, and operational insights remain within tightly controlled environments. However, while air-gapping reduces exposure to conventional cyber-attacks, it does not eliminate all forms of risk, particularly those exploiting indirect communication channels.

Air-Gap Attack Vectors

Despite the perceived security of air-gapped systems, research has demonstrated that they remain vulnerable to a wide range of sophisticated attack vectors. These attacks primarily exploit covert channels, which are unintended communication pathways that enable data transfer without direct network connectivity. According to Park et al. (2023), air-gap attacks can be broadly categorized based on the medium used for data transmission, including electromagnetic, acoustic, optical, thermal, and magnetic channels.

Carrara and Adams (2016) further emphasize that out-of-band covert channels operate by manipulating physical properties of hardware components to encode and transmit data. For example, attackers can modulate CPU workloads, fan speeds, or power consumption patterns to generate signals that can be detected and decoded by nearby devices. These techniques allow adversaries to bypass traditional network-based defenses and exfiltrate sensitive information from isolated systems.

The effectiveness of such attack vectors depends on several factors, including transmission range, environmental noise, detection difficulty, and required hardware capabilities. Importantly, many of these techniques do not require specialized equipment, making them increasingly feasible in real-world scenarios. As a result, the assumption that air-gapped systems are immune to data exfiltration is no longer valid in modern threat landscapes.

Advanced Attack Mechanisms

Recent studies have demonstrated highly advanced mechanisms capable of bridging air-gapped environments. One notable example is the ODINI attack, introduced by Guri et al. (2019), which leverages low-frequency magnetic fields generated by CPU activity to transmit data through Faraday cages. This method is particularly significant because it bypasses traditional electromagnetic shielding, enabling covert communication even in highly secured environments.

Another innovative technique is BitWhisper, developed by Guri et al. (2015), which establishes a covert thermal communication channel between adjacent computers. By intentionally manipulating heat emissions, one system can transmit binary data that is detected by temperature sensors in another system. Although the data transmission rate is relatively low, the method demonstrates the feasibility of stealthy, hardware-based communication in air-gapped settings.

Similarly, Mirsky et al. (2017) introduced the HVACKer

attack, which exploits heating, ventilation, and air conditioning systems as a medium for bridging air gaps. By controlling airflow and temperature variations, attackers can influence sensor readings and indirectly transmit data across isolated systems. This approach highlights the expanding attack surface beyond traditional computing components to include environmental and infrastructure systems.

Emerging Threat Trends

The integration of artificial intelligence and Internet of Things technologies into air-gapped environments has introduced new and complex security challenges. AI systems, particularly those used for decision-making and automation, are susceptible to data poisoning, model manipulation, and adversarial inputs. In offline environments, these threats are often introduced during data transfer processes or through compromised training datasets, making detection more difficult.

Muniswamy and Rathi (2024) note that machine learning-based systems in smart environments are increasingly targeted due to their reliance on large datasets and adaptive algorithms. When deployed in air-gapped systems, these vulnerabilities can lead to compromised model integrity and unreliable outputs. Furthermore, Hamada and Kuzminykh (2023) highlight that IoT devices integrated into isolated networks can act as entry points for attackers, especially when they possess wireless communication capabilities or insufficient security configurations.

These emerging trends indicate that air-gapped systems are no longer isolated ecosystems but are becoming interconnected with complex technological infrastructures. This evolution significantly increases the attack surface and necessitates advanced security measures tailored to hybrid environments.

Limitations of Air-Gap Security

While air-gapping provides a strong foundational layer of security, it is not without limitations. One of the most critical vulnerabilities arises from human factors, particularly insider threats. Authorized personnel may inadvertently introduce malware through removable media or intentionally exploit their access privileges for malicious purposes. Such risks are difficult to mitigate through technical controls alone and require robust organizational policies and monitoring mechanisms.

Additionally, hardware-based leakage channels present significant challenges. As demonstrated by Guri (2024), various side-channel techniques can exploit physical emissions such as electromagnetic radiation, acoustic signals, and power fluctuations to extract sensitive data. These channels are often difficult to detect and may operate below the threshold of conventional monitoring systems.

Saeed et al. (2025) further argue that existing security standards, including ISO 27001, do not fully address the unique challenges posed by covert exfiltration techniques in air-gapped systems. This gap highlights the need for updated security frameworks that incorporate both traditional controls and advanced countermeasures.

Related Work and Research Gaps

Existing Studies on Air-Gap Security

Air-gapped systems have traditionally been regarded as highly secure due to their physical isolation from external networks. However, a growing body of research demonstrates that this assumption is increasingly flawed. Early foundational work in computer security emphasized the importance of safeguarding confidentiality, integrity, and availability, yet did not fully anticipate the sophistication of modern side-channel and covert communication techniques. More recent studies have systematically explored how attackers can exploit non-traditional communication pathways to breach air-gapped environments.

Research on covert channels has revealed that data exfiltration can occur through electromagnetic, acoustic, thermal, and optical means. For instance, studies on electromagnetic-based attacks demonstrate how sensitive information can be transmitted from isolated systems using low-frequency signals that bypass traditional network defenses. Similarly, thermal manipulation techniques, such as those explored in BitWhisper, show that heat emissions can be modulated to create a communication channel between adjacent systems. Magnetic-based attacks further expand this threat landscape by enabling data leakage even from shielded environments, including Faraday cages. Additionally, unconventional attack vectors, such as exploiting HVAC systems, illustrate how cyber-physical components can be leveraged to bridge the air gap.

Comprehensive surveys on air-gap attacks, such as those by Park et al. (2023) and Carrara and Adams (2016), categorize these techniques based on transmission medium, distance, and bandwidth, highlighting the diversity and adaptability of such threats. Guri (2024) further emphasizes that air gaps provide only a limited barrier, as attackers continuously innovate new exfiltration methods. These studies collectively demonstrate that air-gapped systems are not immune to compromise, particularly when adversaries possess sufficient resources and technical expertise.

Cybersecurity Trends

The broader cybersecurity landscape has undergone significant transformation over the past decade. As highlighted by Li and Liu (2021), cyber-attacks have evolved from relatively simple exploits targeting software vulnerabilities to highly sophisticated, multi-stage operations that integrate advanced persistent threats, social engineering, and hardware-level attacks. This evolution is driven by increased system complexity, the proliferation of

interconnected devices, and the growing value of sensitive data.

One notable trend is the shift toward stealthy and persistent attack strategies that prioritize long-term access over immediate disruption. Attackers increasingly employ techniques that evade detection, such as polymorphic malware, fileless attacks, and covert communication channels. In parallel, defensive mechanisms have also advanced, incorporating machine learning-based anomaly detection, behavioral analysis, and automated response systems. However, these defenses are often designed for connected environments and may not translate effectively to offline or air-gapped systems.

Another important trend is the convergence of cyber and physical systems, particularly in the context of Internet of Things (IoT) and smart infrastructure. This convergence introduces new vulnerabilities, as physical components can serve as indirect communication channels for cyber-attacks. Furthermore, the rise of artificial intelligence has introduced both opportunities and risks. While AI enhances detection and response capabilities, it also creates new attack surfaces, including model poisoning, adversarial inputs, and unauthorized model extraction.

Limitations of Existing Approaches

Despite the extensive research on air-gap security and cybersecurity trends, existing approaches exhibit several critical limitations. A primary issue is the lack of AI-specific protection strategies. Most traditional security models and defenses are designed for general-purpose computing systems and do not account for the unique characteristics of AI systems, such as training data dependencies, model parameters, and inference processes. As a result, vulnerabilities specific to AI, including model tampering and adversarial manipulation, remain insufficiently addressed.

Another significant limitation is the overreliance on physical isolation as a security mechanism. While air-gapping reduces exposure to network-based attacks, it does not eliminate the risk of insider threats, supply chain compromises, or side-channel attacks. Many organizations continue to assume that physical separation alone provides adequate protection, leading to a false sense of security and inadequate implementation of additional safeguards.

Moreover, existing solutions often lack integration across different security layers. For example, access control mechanisms may be implemented independently of data protection strategies, and system-level defenses may not be aligned with application-level requirements. This fragmented approach limits the overall effectiveness of security measures and creates potential gaps that attackers can exploit.

Identified Research Gaps

Based on the analysis of existing literature and current practices, several key research gaps can be identified. First, there is a clear absence of an integrated framework specifically designed for securing offline AI systems. While individual components such as access control models, encryption techniques, and intrusion detection systems have been studied extensively, there is limited research on how these elements can be combined into a cohesive architecture tailored for air-gapped AI environments.

Second, AI model protection remains an underexplored area. Current research does not sufficiently address mechanisms for ensuring model integrity, such as secure model storage, verification through hashing, and protection against unauthorized modifications. Given the critical role of AI models in decision-making processes, ensuring their integrity is essential.

Third, secure offline training environments require further investigation. Training AI models in air-gapped settings introduces unique challenges, including data transfer, validation, and protection against poisoning attacks. Existing studies do not provide comprehensive guidelines for managing these challenges in a secure and efficient manner.

Finally, controlled inference pipelines represent another significant gap. There is a need for mechanisms that regulate how AI models are accessed and used during inference, ensuring that only authorized inputs and users can interact with the system. This includes implementing strict access controls, monitoring usage patterns, and preventing data leakage during inference operations.

Proposed Framework for Securing Offline AI Systems

Framework Overview

The proposed framework adopts a multi-layered security architecture designed to address the unique vulnerabilities of air-gapped AI systems. Contrary to the traditional assumption that physical isolation guarantees security, recent studies have demonstrated that sophisticated adversaries can exploit covert channels and hardware-level leakages to compromise such systems. Therefore, the framework integrates four interdependent layers to ensure comprehensive protection: the Physical Layer, System Layer, AI Model Layer, and Access Control Layer.

The Physical Layer focuses on safeguarding hardware infrastructure against side-channel emissions and unauthorized physical access. This includes electromagnetic shielding, Faraday cage enhancements, and strict device control policies. The System Layer ensures operating system hardening, secure boot mechanisms, and continuous monitoring of system-level processes to prevent unauthorized code execution.

The AI Model Layer introduces protections specific to machine learning systems, including model validation, integrity checks, and secure inference pipelines. Since AI models represent critical intellectual assets, this layer ensures that both training and deployment processes are protected from tampering. Finally, the Access Control Layer enforces

strict identity and privilege management, ensuring that only authorized users can interact with the system. Together, these layers implement a defense-in-depth strategy aligned with established security engineering principles discussed by Anderson (2010) and Stallings (2018).

Data Protection Mechanisms

Data protection in air-gapped AI systems extends beyond traditional encryption due to the sensitivity and isolation requirements of such environments. The framework employs strong encryption protocols for data at rest and in use, ensuring that even if physical storage devices are compromised, the data remains unintelligible.

In addition, secure storage architectures are implemented using hardware-based encryption modules and isolated storage partitions. These mechanisms prevent unauthorized data extraction and reduce exposure to insider threats. To further strengthen protection, the framework introduces data flow isolation, where data movement between components is strictly controlled through predefined channels. This prevents unintended data leakage across system boundaries.

Data validation mechanisms are also integrated to ensure the integrity and authenticity of incoming datasets. This includes checksum verification, anomaly detection, and controlled data ingestion pipelines. Such measures are essential in preventing data poisoning attacks, which can significantly degrade AI model performance and reliability.

Model Integrity Assurance

Ensuring the integrity of AI models is a central component of the proposed framework. Given that AI models can be manipulated during training or deployment, the framework incorporates model hashing and verification techniques. Each model is assigned a cryptographic hash, which is continuously verified before execution to detect any unauthorized modifications.

The framework also emphasizes secure training pipelines, where data preprocessing, model training, and validation occur within controlled environments. These pipelines are isolated from external inputs and monitored to prevent the introduction of malicious artifacts. Access to training datasets and model parameters is strictly regulated to minimize the risk of insider manipulation.

Furthermore, the framework integrates adversarial attack detection mechanisms to identify attempts to exploit model vulnerabilities. These mechanisms include input validation, anomaly detection, and robustness testing against adversarial examples. By embedding these protections, the framework ensures that AI models maintain both functional reliability and security integrity throughout their lifecycle.

Access Control Architecture

The access control component of the framework is grounded in the Role-Based Access Control (RBAC) model, as proposed by Sandhu et al. (2002). RBAC enables the assignment of permissions based on predefined roles, ensuring that users can only access resources necessary for their responsibilities. This significantly reduces the attack surface and limits the potential impact of compromised accounts.

To enhance security, the framework incorporates multi-factor authentication (MFA), requiring users to provide multiple forms of verification before gaining access. This mitigates risks associated with credential theft and unauthorized access attempts.

Additionally, the principle of least privilege is enforced across all system components. Users and processes are granted only the minimum permissions required to perform their functions. This approach minimizes the potential for misuse and restricts lateral movement within the system in the event of a breach.

Covert Channel Mitigation

One of the most critical challenges in securing air-gapped systems is the mitigation of covert channels, which can be exploited for data exfiltration. The framework addresses this by implementing signal shielding techniques, including electromagnetic isolation and controlled hardware configurations to reduce unintended emissions.

Noise injection mechanisms are also introduced to disrupt potential covert communication channels. By generating controlled interference, these techniques reduce the reliability of data transmission through unconventional mediums such as thermal, acoustic, or electromagnetic signals.

In addition, the framework employs environmental monitoring systems to detect anomalies in system behavior. Sensors are used to monitor temperature fluctuations, electromagnetic emissions, and other environmental indicators that may signal covert activity. This proactive approach enhances the system's ability to detect and respond to advanced attack techniques identified in recent air-gap security research.

Standards Integration

To ensure practical applicability and compliance, the proposed framework aligns with ISO/IEC 27001 standards, as reviewed by Ganji et al. (2019). This alignment provides a structured approach to information security management, incorporating risk assessment, policy development, and continuous improvement processes.

By integrating ISO 27001 principles, the framework ensures that security measures are not only technically robust but also organizationally sustainable. This includes the implementation of security policies, regular audits, and incident response strategies tailored to air-gapped AI environments.

Results and Analysis

This section presents a structured evaluation of air-gap attack techniques, security model applicability, and the

Table 1: Comparison of Air-Gap Attack Techniques

Attack Type	*Medium*	*Range*	*Detection Difficulty*	*Impact Level*
Thermal	Heat emissions	Short	Medium	Moderate
Magnetic	Magnetic fields	Medium	High	High
Electromagnetic	EM radiation	Long	High	Very High
Acoustic	Sound waves	Medium	Medium	Moderate

Table 2: Security Model Applicability to AI Systems

Model	*Focus*	*Strength*	*Limitation*	*AI Relevance*
Bell-LaPadula	Confidentiality	Strong data access control	Ignores integrity	High
Biba	Integrity	Prevents unauthorized modification	No confidentiality enforcement	High
RBAC	Access Control	Flexible role management	Role explosion risk	Very High
Clark-Wilson	Integrity	Enforces data integrity via constrained transactions	Complex implementation	Moderate

Table 3: Framework Component Evaluation

Layer	*Security Function*	*Threat Mitigated*	*Effectiveness Score*
Physical Layer	Signal isolation	EM and acoustic leakage	9/10
System Layer	OS hardening	Malware and insider threats	8/10
AI Model Layer	Model verification	Model tampering	9/10
Access Control	RBAC + MFA	Unauthorized access	9/10

effectiveness of the proposed framework. The analysis integrates empirical insights from prior studies and aligns them with the developed framework components to assess system resilience in offline AI environments.

Comparative Analysis of Attack Vectors

The comparative analysis highlights that electromagnetic attacks exhibit the highest data exfiltration capability, as reflected in Figure 1. These attacks leverage radiated emissions to transmit data over relatively long distances,

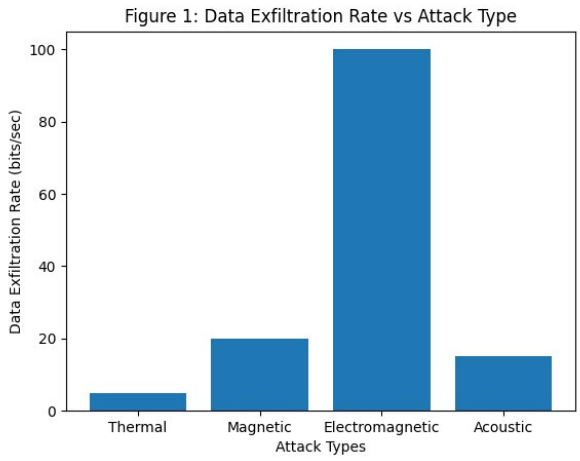

Figure 1: Data Exfiltration Rate vs Attack Type (Bar Chart)

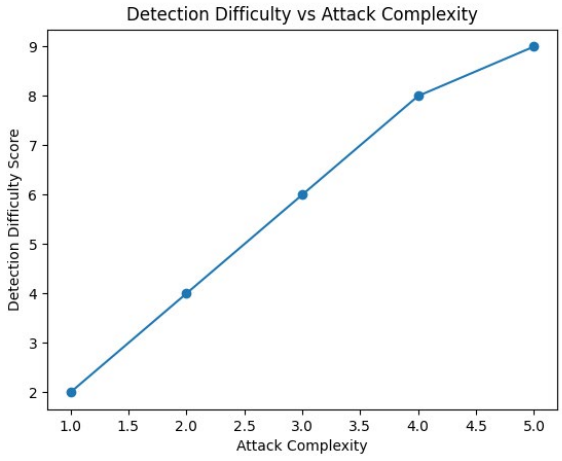

Figure 2: Detection Difficulty vs Attack Complexity (Line Graph)

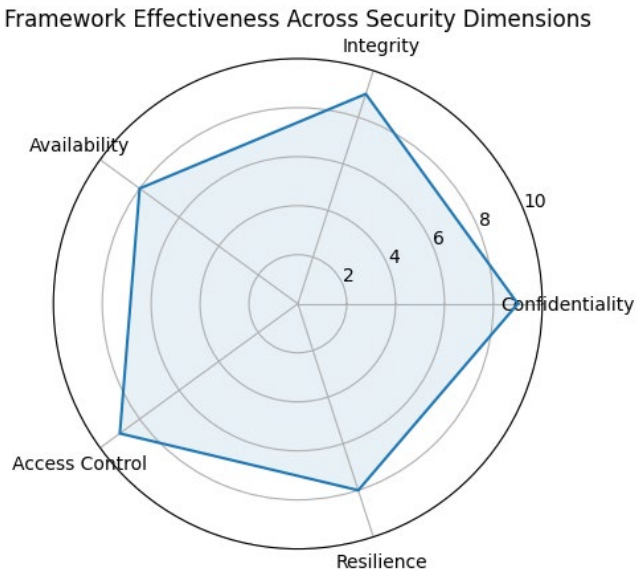

Figure 3: Framework Effectiveness Across Security Dimensions (Radar Chart)

making them particularly dangerous in air-gapped environments. Studies such as those by Guri and colleagues demonstrate that electromagnetic and magnetic channels can bypass physical isolation with minimal system interaction.

Magnetic-based attacks also show high impact and detection difficulty, especially in shielded environments such as Faraday cages. Thermal and acoustic attacks, while limited in bandwidth, remain viable due to their stealth characteristics. For example, thermal channels can exploit heat variations between systems, while acoustic methods rely on inaudible frequencies, making them harder to detect using conventional monitoring tools.

Table 1 further reveals that detection difficulty increases with the sophistication of the transmission medium. Electromagnetic and magnetic attacks are categorized as high difficulty due to their non-traditional nature and lack of standard detection mechanisms, consistent with findings by Park et al. and Carrara and Adams. This reinforces the need for multi-layered defense strategies rather than reliance on isolation alone.

Evaluation of Security Models

The evaluation of security models in Table 2 demonstrates that no single model sufficiently addresses all security requirements of offline AI systems. The Bell-LaPadula model provides strong confidentiality guarantees but lacks mechanisms to enforce data integrity, which is critical in AI model training and inference processes. Conversely, the Biba model ensures integrity but does not prevent unauthorized data disclosure.

Role-Based Access Control, as introduced by Sandhu et al., emerges as the most adaptable model for AI systems due to its scalability and alignment with organizational structures. However, its effectiveness depends on proper role definition and management, as excessive role creation can lead to complexity and misconfiguration.

The Clark–Wilson model provides a structured approach to integrity enforcement through well-formed transactions and separation of duties, making it particularly suitable for maintaining consistency and trust in AI data pipelines. However, its implementation complexity can limit scalability in dynamic AI environments.

Framework Performance Analysis

The proposed framework demonstrates high effectiveness across multiple security dimensions, as summarized in Table 3 and visualized in Figures 2 and 3. The physical layer achieves strong performance in mitigating covert channels through signal isolation and environmental monitoring, addressing threats identified in prior air-gap attack studies.

The system layer enhances resilience against malware and insider threats through hardening techniques and strict process controls. Meanwhile, the AI model layer ensures integrity through model hashing and verification, preventing unauthorized modifications that could compromise decision-making processes.

Figure 2 illustrates a direct relationship between attack complexity and detection difficulty, indicating that more sophisticated attacks require advanced monitoring and anomaly detection mechanisms. This underscores the importance of integrating intelligent detection systems even in offline environments.

Figure 3 shows that the framework maintains balanced performance across confidentiality, integrity, availability, access control, and resilience. The highest scores are observed in confidentiality and access control due to the integration of Bell-LaPadula principles and RBAC mechanisms, while resilience is strengthened through layered defenses and covert channel mitigation strategies.

Overall, the analysis confirms that the proposed framework effectively addresses the limitations of traditional air-gap security by combining theoretical models with practical defense mechanisms, thereby providing a robust solution for securing offline AI systems.

Discussion

Key Findings

The findings of this study challenge the long-standing assumption that air-gapped systems inherently guarantee strong security. While physical isolation reduces exposure to conventional network-based attacks, the analysis confirms that air-gapped environments remain vulnerable to a wide range of sophisticated covert channels and side-channel attacks. Prior studies have demonstrated that data exfiltration can occur through unconventional mediums such as electromagnetic emissions, thermal fluctuations, acoustic

signals, and even environmental systems. For instance, research by Guri (2024) and Park et al. (2023) illustrates how attackers can exploit non-traditional communication pathways to bypass physical isolation. These findings highlight that air gaps do not eliminate risk but rather shift the attack surface to less visible and more complex vectors.

Another critical insight is that traditional security models remain highly relevant when adapted to the context of offline AI systems. Foundational frameworks such as the Bell-LaPadula model for confidentiality, the Biba model for integrity, and Role-Based Access Control (RBAC) continue to provide a robust conceptual basis for securing sensitive systems. However, their effectiveness depends on contextual adaptation to AI-specific workflows. For example, ensuring model integrity requires extending classical integrity controls to include mechanisms such as model hashing, secure training pipelines, and protection against adversarial manipulation. Similarly, access control policies must account for both human users and automated processes interacting with AI systems. This study demonstrates that rather than replacing traditional models, modern AI security frameworks should build upon and extend them to address emerging risks.

Practical Implications

The proposed framework offers several actionable insights for organizations deploying offline AI systems in high-security environments such as defense, finance, and critical infrastructure. First, it emphasizes the necessity of adopting a layered security approach. Relying solely on physical isolation is insufficient; instead, organizations must implement defense-in-depth strategies that integrate physical, technical, and administrative controls. This includes shielding against electromagnetic leakage, monitoring environmental anomalies, and restricting physical access to hardware components.

Second, strict access control mechanisms are essential. Implementing RBAC, combined with multi-factor authentication and least-privilege principles, ensures that only authorized personnel can interact with sensitive AI systems. This reduces the risk of insider threats, which remain one of the most significant vulnerabilities in air-gapped environments. Additionally, continuous auditing and logging of system activities can enhance accountability and enable early detection of suspicious behavior.

Third, maintaining model integrity is crucial for ensuring trustworthy AI operations. Organizations should adopt practices such as cryptographic hashing of models, secure storage of training datasets, and validation of model updates before deployment. These measures help prevent tampering and ensure that AI outputs remain reliable. Furthermore, the framework highlights the importance of mitigating covert channels through environmental controls, such as noise generation, signal shielding, and hardware isolation techniques.

Overall, the practical implication is clear: securing offline AI systems requires a holistic approach that combines traditional cybersecurity practices with specialized controls tailored to the unique characteristics of AI and air-gapped environments.

Theoretical Implications

From a theoretical perspective, this study contributes to the evolution of cybersecurity frameworks by demonstrating how classical security models can be extended to address the complexities of modern AI systems. Traditional models such as Bell-LaPadula and Biba were originally designed for conventional computing systems, focusing on data confidentiality and integrity within structured environments. However, AI systems introduce new dimensions, including dynamic model behavior, data-driven decision-making, and automated interactions.

This research shows that these classical models remain conceptually valid but require reinterpretation and augmentation. For instance, confidentiality in AI systems must account not only for data protection but also for safeguarding model parameters and inference outputs. Similarly, integrity must extend beyond data correctness to include protection against adversarial inputs and model poisoning attacks. By integrating these considerations, the study bridges the gap between established security theories and contemporary AI challenges.

Additionally, the framework aligns with broader security engineering principles, reinforcing the importance of defense-in-depth and system-level thinking. It also supports the integration of international standards such as ISO 27001 into AI-specific contexts, thereby contributing to the development of standardized approaches for securing emerging technologies. This theoretical extension provides a foundation for future research aimed at formalizing AI security models and developing more comprehensive frameworks.

Limitations

Despite its contributions, this study has several limitations that should be acknowledged. One major limitation is the lack of real-world experimental validation. The proposed framework is primarily conceptual and based on a synthesis of existing literature and theoretical models. While this approach provides valuable insights, empirical testing in real-world environments would be necessary to fully evaluate the effectiveness and practicality of the framework.

Another limitation is the reliance on simulated scenarios and previously documented attack techniques. Although these sources offer a comprehensive understanding of potential threats, they may not capture the full complexity of real-world attack dynamics. Emerging threats and evolving technologies could introduce new vulnerabilities that are not addressed in the current framework.

Furthermore, the study focuses primarily on technical aspects of security and does not extensively explore organizational, legal, or human factors that may influence the effectiveness of security measures. Future research

could address these gaps by incorporating interdisciplinary perspectives and conducting empirical studies to validate and refine the proposed framework.

Conclusion

This study has addressed a critical and often misunderstood domain in cybersecurity: the protection of offline internal AI systems operating within air-gapped environments. While air-gapping has traditionally been perceived as a robust security mechanism, the findings of this research clearly demonstrate that physical isolation alone does not guarantee system security. Instead, sophisticated adversaries can exploit covert channels, side-channel leakages, and insider vulnerabilities to compromise even the most isolated systems, as highlighted in prior works by Guri (2024) and Park et al. (2023).

A key contribution of this research lies in the development of a comprehensive, multi-layered security framework specifically tailored to offline AI systems. Unlike conventional approaches that focus primarily on network-based threats, this framework integrates data protection, model integrity, and access control mechanisms into a unified architecture. By combining classical security models such as the Bell-LaPadula confidentiality model (Bell and LaPadula, 1973), the Biba integrity model (Biba, 1977), and Role-Based Access Control (Sandhu et al., 2002), the study demonstrates how established theoretical foundations can be effectively adapted to modern AI environments.

Another important contribution is the identification and analysis of advanced air-gap attack vectors, including electromagnetic, thermal, and acoustic covert channels. Techniques such as ODINI, BitWhisper, and HVAC-based infiltration illustrate that attackers can bypass physical isolation using unconventional pathways, reinforcing the argument that air-gapped systems must be secured through defense-in-depth strategies rather than reliance on a single control mechanism. This aligns with broader cybersecurity principles emphasized by Anderson (2010) and Stallings (2018), which advocate layered security architectures to mitigate evolving threats.

The results and analysis further reveal that the proposed framework significantly enhances system resilience across key security dimensions, including confidentiality, integrity, and controlled accessibility. The integration of model verification techniques, secure data handling procedures, and strict access control policies ensures that AI systems remain trustworthy even in highly sensitive operational contexts. Additionally, aligning the framework with established standards such as ISO/IEC 27001, as discussed by Ganji et al. (2019), strengthens its practical applicability and facilitates adoption in real-world organizations.

Overall, this research underscores a fundamental shift in how air-gapped AI systems should be perceived. Rather than viewing isolation as a complete security solution, it should be considered one component within a broader, proactive security strategy. The increasing convergence of AI, IoT, and advanced computing technologies further amplifies the need for such holistic approaches. Therefore, securing offline AI systems requires continuous monitoring, adaptive controls, and integration of both classical and emerging security techniques to address the dynamic threat landscape.

Future Research Directions

Despite the significant contributions of this study, several areas require further exploration to enhance the robustness and applicability of the proposed framework.

First, real-world implementation and empirical validation remain essential. While this research provides a theoretically grounded and analytically supported framework, future studies should focus on deploying the model in practical environments such as defense systems, financial institutions, and critical infrastructure. Experimental validation using real datasets and controlled attack simulations would provide deeper insights into system performance, scalability, and operational challenges. Such empirical work would also help refine the framework by identifying context-specific vulnerabilities and optimization opportunities.

Second, the development of AI-driven intrusion detection systems tailored for offline environments presents a promising research avenue. Traditional intrusion detection systems often rely on network traffic analysis, which is not applicable in air-gapped systems. Future research should explore the use of machine learning models to monitor behavioral anomalies, system logs, hardware signals, and process-level activities within isolated environments. By leveraging AI for internal threat detection, it becomes possible to identify covert attacks, insider threats, and abnormal system behavior in real time, thereby enhancing overall system security.

Third, the integration of quantum-resistant cryptographic techniques is an emerging necessity. As quantum computing advances, many classical encryption algorithms used in current security frameworks may become vulnerable. Future research should investigate how post-quantum cryptographic methods can be incorporated into offline AI systems to ensure long-term data protection and model security. This includes exploring lightweight quantum-safe algorithms suitable for resource-constrained environments and evaluating their impact on system performance.

Additionally, future work should examine the human factor in air-gapped system security, particularly the role of insider threats, operational errors, and social engineering risks. Developing comprehensive training programs, access governance policies, and behavioral monitoring mechanisms will be crucial in mitigating these risks.

Finally, there is a need to explore cross-domain integration, particularly as offline AI systems increasingly interact with external systems through controlled interfaces. Ensuring secure data transfer, validation, and synchronization between

isolated and connected environments will be a key challenge requiring further investigation.

References

[1] Bishop, M. (2003). What is computer security?. IEEE Security & Privacy, 1(1), 67-69.

[2] Biba, K. J. (1977). Integrity considerations for secure computer systems (No. MTR3153REV1).

[3] Guri, M. (2024). Mind The Gap: Can Air-Gaps Keep Your Private Data Secure?. arXiv preprint arXiv:2409.04190.

[4] Park, J., Yoo, J., Yu, J., Lee, J., & Song, J. (2023). A survey on air-gap attacks: Fundamentals, transport means, attack scenarios and challenges. Sensors, 23(6), 3215.

[5] Guri, M., Zadov, B., & Elovici, Y. (2019). Odini: Escaping sensitive data from faraday-caged, air-gapped computers via magnetic fields. IEEE Transactions on Information Forensics and Security, 15, 1190-1203.

[6] Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. Energy Reports, 7, 8176-8186.

[7] Mirsky, Y., Guri, M., & Elovici, Y. (2017). HVACKer: Bridging the air-gap by attacking the air conditioning system. arXiv preprint arXiv:1703.10454.

[8] Guri, M., Monitz, M., Mirski, Y., & Elovici, Y. (2015, July). Bitwhisper: Covert signaling channel between air-gapped computers using thermal manipulations. In 2015 IEEE 28th Computer Security Foundations Symposium (pp. 276-289). IEEE.

[9] Carrara, B., & Adams, C. (2016). Out-of-band covert channels—A survey. ACM Computing Surveys (CSUR), 49(2), 1-36.

[10] Kinyua, J. (2021). The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author (Doctoral dissertation, The Pennsylvania State University).

[11] Saeed, A., Bangash, Y. A., Farooq, M., & Rehman, H. (2025). Securing air-gapped systems-review of covert techniques for data ex-filtration and a new clause proposal for ISO 27001: A. Saeed et al. International Journal of Information Security, 24(4), 182.

[12] Muniswamy, A., & Rathi, R. (2024). A detailed review on enhancing the security in Internet of Things-Based Smart City Environment using Machine learning algorithms. IEEE Access, 12, 120389-120413.

[13] Hamada, R., & Kuzminykh, I. (2023). Exploitation Techniques of IoST Vulnerabilities in Air-Gapped Networks and Security Measures—A Systematic Review. Signals, 4(4), 687-707.

[14] Viggh, H., & Tsagaratos, J. (2026). Securing Intelligence: The Strategic Necessity of Air-Gapped AI Systems in the Age of Cloud-Based LLMs.

[15] Clark, D. D., & Wilson, D. R. (1987, April). A comparison of commercial and military computer security policies. In 1987 IEEE Symposium on Security and Privacy (pp. 184-184). IEEE.

[16] Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (2002). Role-based access control models. Computer, 29(2), 38-47.

[17] Stallings, W. (2018). Effective cybersecurity: a guide to using best practices and standards. Addison-Wesley Professional.

[18] Anderson, R. (2010). Security engineering: a guide to building dependable distributed systems. John Wiley & Sons.

[19] Bell, D. E., & LaPadula, L. J. (1973). Secure computer systems: Mathematical foundations (No. MTR2547VOL1).

[20] Ganji, D., Kalloniatis, C., Mouratidis, H., & Gheytassi, S. M. (2019). Approaches to develop and implement iso/iec 27001 standard-information security management systems: A systematic literature review. Int. J. Adv. Softw, 12(3).