Adversarial Machine Learning Threats to Medical Device AI Controllers

A Threat Taxonomy and Defense Landscape Analysis

Venkata Sai Abhinav Piratla

Independent Researcher
New Haven, CT, USA

Abstract:
The integration of artificial intelligence into life-critical medical device controllers—including closed-loop insulin delivery systems and cardiac monitoring devices—introduces adversarial machine learning (AML) attack surfaces that conventional cybersecurity frameworks do not adequately address. Adversarial attacks targeting these systems carry direct patient safety implications, yet no comprehensive, medical-device-specific AML threat taxonomy exists in the current literature. This paper addresses that gap by presenting a structured taxonomy of 34 adversarial attacks spanning training-phase, inference-phase, privacy, and model integrity threat categories, with explicit analysis of applicability to resource-constrained medical hardware. Building on prior work in artificial pancreas security [1] and autoencoder-based anomaly detection [2], a catalog of 25 defense mechanisms is evaluated against the computational and memory constraints typical of implantable, wearable, and portable device classes. Identified threats are mapped to the MITRE ATLAS adversarial machine learning framework and aligned with FDA cybersecurity and AI lifecycle guidance, providing a regulatory reference for device manufacturers and premarket submission planning. Tiered defense recommendations are derived from device resource profiles, enabling practitioners to select contextually appropriate countermeasures. The taxonomy and defense landscape together constitute an actionable resource for engineering adversarially resilient AI-enabled medical devices.

Keywords: Adversarial machine learning, medical device security, threat taxonomy, artificial pancreas, MITRE ATLAS, evasion attacks, data poisoning.

I. INTRODUCTION

The adoption of artificial intelligence in medical device controllers has accelerated at a pace that substantially outstrips the development of corresponding security frameworks. The U.S. Food and Drug Administration (FDA) had authorized more than 1,000 AI/ML-enabled medical devices by mid-2025, with 168 such authorizations occurring in calendar year 2024 alone, of which approximately 75% were concentrated in radiology and diagnostic imaging [29]. Beyond static diagnostic applications, machine learning models now serve as active controllers in closed-loop physiological systems: reinforcement learning agents govern insulin dosing in artificial pancreas (AP) platforms [1], deep neural networks classify cardiac arrhythmias in real-time implantable monitors, and anomaly detection models form the basis of sensor-fusion safety layers in wearable devices [2]. These deployments represent a qualitative shift from passive decision-support tools to safety-critical feedback controllers whose outputs directly actuate physiological intervention.

This operational context introduces AML attack surfaces that are fundamentally distinct from those addressed by conventional cybersecurity frameworks. Where traditional medical device cybersecurity focuses on unauthorized access, data exfiltration, and denial-of-service disruption, adversarial attacks exploit the statistical properties of machine learning models to induce targeted misbehavior while evading conventional anomaly detection. A perturbation of only ±1 mg/dL applied to continuous glucose monitor (CGM) inputs—imperceptible to clinical monitoring thresholds—has been demonstrated to increase closed-loop dosing risk by 423% in pediatric subjects and extend hypoglycemic duration by 1,079% in adolescents [12]. A denial-of-service attack on a reinforcement learning AP controller has been shown to produce life-threatening glucose levels within 50 minutes of activation [12]. In cardiac monitoring, adversarial examples crafted using six standard attack methods have been demonstrated to reliably fool deep learning classifiers on clinically

validated ECG datasets. The clinical consequence of these attacks—wrong insulin doses, missed arrhythmias, false diagnostic outputs—is direct and potentially irreversible patient harm, a severity class that has no equivalent in conventional enterprise cybersecurity.

Despite the severity of this threat surface, no comprehensive adversarial machine learning threat taxonomy specific to medical device AI controllers exists in the current literature. General-purpose AML taxonomies, including the NIST AI 100-2e2025 framework [31] and the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) [33], provide indispensable vocabulary and technique catalogs; however, neither framework accounts for the resource constraints of implantable and wearable medical hardware, the clinical impact severity of misclassification, nor the regulatory obligations that govern medical device software modification.

This paper addresses that gap through four primary contributions. First, a hierarchical taxonomy of 34 adversarial attack types is presented, spanning training-phase data poisoning, inference-phase evasion, privacy extraction, and model integrity attacks, with explicit applicability ratings for six classes of medical device AI components. Second, the taxonomy is cross-referenced with MITRE ATLAS techniques and NIST AI 100-2e2025 terminology, providing a unified regulatory vocabulary for device manufacturers and premarket submission planning. Third, a catalog of 25 defense mechanisms is evaluated against the computational and memory constraints of four device resource tiers. Fourth, tiered defense selection recommendations are derived from device resource profiles, with explicit alignment to FDA cybersecurity guidance and the Predetermined Change Control Plan (PCCP) framework [28].

II. BACKGROUND

A. Adversarial Machine Learning Fundamentals

Adversarial machine learning encompasses a class of attacks that exploit the inherent statistical properties of trained models to induce erroneous outputs without violating observable data constraints. The foundational demonstration by Szegedy et al. [3] established that imperceptible perturbations to input vectors, discoverable through gradient-based optimization, could cause high-confidence misclassification in deep neural networks. Goodfellow et al. subsequently introduced the Fast Gradient Sign Method (FGSM), attributing adversarial vulnerability to the linearity of learned representations and establishing the perturbation budget ε as the canonical measure of attack magnitude [4]. The field has since bifurcated along two primary axes: the attacker's knowledge model (white-box access to model parameters versus black-box access limited to output queries) and the attack phase (training-time manipulation versus inference-time perturbation).

NIST AI 100-2e2025 provides the most current unified taxonomy, categorizing adversarial threats to predictive AI systems under three top-level objectives: availability (rendering the model non-functional), integrity (inducing incorrect outputs), and privacy (extracting information about training data or model parameters) [31]. This classification is complementary to the four-phase attack lifecycle formalized in MITRE ATLAS: Reconnaissance, ML Attack Staging, Defense Evasion, and Impact [33]. The ATLAS framework extends the traditional MITRE ATT&CK vocabulary with 15 AI-specific tactics and 66 techniques applicable across the AI development and deployment pipeline [33].

B. Medical Device AI Landscape

AI controllers in medical devices follow a common architectural pattern: raw physiological signals from sensors (glucose electrodes, ECG electrodes, accelerometers) are processed through preprocessing pipelines and fed to inference models, whose outputs drive actuators or clinical alerts. In closed-loop AP systems, CGM readings are consumed by model predictive control algorithms or reinforcement learning agents to compute insulin dosing commands delivered via infusion pumps [1]. In cardiac devices, ECG signals are classified by convolutional neural networks to detect arrhythmia and trigger defibrillation or pacing. Anomaly detection layers employing autoencoders or statistical thresholding provide secondary safety validation in both device classes [2].

The resource constraints of medical hardware impose severe limits on deployable defensive countermeasures. Implantable devices typically operate on ARM Cortex-M0/M3 microcontrollers with fewer than 256 KB of RAM; wearable CGM transmitters and insulin pumps occupy a moderate tier (Cortex-M4, 256 KB–1 MB RAM); smartphone-based AP controllers represent a light-constraint tier (>1 MB RAM); and cloud-based

training infrastructure operates without hardware constraints. This four-tier model governs the feasibility of any proposed defense mechanism.

C. Regulatory and Standards Context

The FDA's 2023 final guidance on cybersecurity in medical device premarket submissions established foundational requirements for threat modeling, Software Bill of Materials (SBOM), and security testing applicable to all networked medical devices [29]. Subsequent AI-specific guidance—the January 2025 draft on AI-enabled device software functions (AI-DSF) [30] and the August 2025 final guidance on Predetermined Change Control Plans (PCCPs) for AI-enabled devices [28]—introduced lifecycle management obligations that directly implicate adversarial robustness. The NIST AI Risk Management Framework (AI RMF 1.0) [32] provides complementary governance vocabulary, while NIST AI 100-2e2025 [31] offers the canonical attack and mitigation terminology adopted throughout this paper.

III. THREAT TAXONOMY

A. Taxonomy Methodology

The taxonomy presented herein was constructed through systematic synthesis of three primary sources: the NIST AI 100-2e2025 adversarial ML terminology standard [31], the MITRE ATLAS technique catalog (October 2025 release, 66 techniques across 15 tactics) [33], and a corpus of 30+ peer-reviewed studies on adversarial attacks in medical and safety-critical AI contexts [10,12,13,14]. Attack types were classified along four dimensions: attack phase (training vs. inference), attacker knowledge model (white-box, black-box, physical access), attack objective (availability, integrity, privacy), and medical device applicability (assessed against six device component classes). The resulting taxonomy comprises 34 attack types organized into four top-level categories, cross-referenced to ATLAS techniques and NIST AI 100-2e2025 codes where mappings exist. A hierarchical visualization of the complete taxonomy is provided in Figure 1.

B. Training-Phase Attacks

Training-phase attacks compromise the model before or during the training process, with effects that persist through deployment until retraining occurs with clean data. Label flipping is the foundational data poisoning method, in which an adversary modifies class labels to shift the decision boundary in a targeted direction [15]. Targeted poisoning forces a specific test-time misclassification—for example, causing an insulin dosing model to misclassify a high-glucose state as normoglycemic. Backdoor injection [16] embeds a trigger pattern in training data such that any input containing the trigger at inference time produces an attacker-chosen output while exhibiting normal behavior on clean inputs. Clean-label poisoning achieves similar effects without modifying labels. Supply chain attacks distribute trojaned pre-trained model weights prior to fine-tuning by the device manufacturer.

In federated learning deployments, Byzantine gradient attacks allow a minority of malicious participants to corrupt the global model by submitting adversarially crafted gradient updates [25]. Reward poisoning (ADV-034) poisons the training reward signal of a reinforcement learning AP controller to induce unsafe dosing policies that are undetectable through standard safety envelope testing.

C. Inference-Phase Attacks

Inference-phase attacks operate on deployed models by crafting adversarial inputs that cause misclassification or unsafe control actions without altering model parameters. The Fast Gradient Sign Method (FGSM) applies a single-step gradient perturbation in the direction of maximum loss increase [4]; Projected Gradient Descent (PGD) iterates this process and represents the strongest first-order attack [5]. The Carlini-Wagner (C&W) attack solves an optimization problem to find the minimum-norm adversarial perturbation and has empirically broken multiple proposed defenses [6]. AutoAttack provides a standardized ensemble evaluation benchmark [9]. Transfer-based black-box attacks craft adversarial examples on a surrogate model and exploit cross-architecture transferability [7]. Universal adversarial perturbations (UAPs) cause misclassification regardless of the specific input, presenting a persistent threat to deployed ECG classifiers and CGM predictors [8].

Physical and sensor-domain attacks are of specific relevance to medical devices because the attack surface extends to the physical transduction layer. Sensor spoofing involves electromagnetic or radiofrequency injection of false readings into a CGM or ECG sensor. Sensor denial-of-service disrupts the sensor's ability

to deliver readings, forcing the controller into an unsafe default state. Plausible ECG adversarial examples are perturbations constrained to remain within the physiologically admissible envelope of cardiac signals, making them indistinguishable from genuine rhythm variations while reliably fooling neural network classifiers.

D. Privacy Attacks

Privacy attacks extract sensitive information from model parameters or outputs without directly causing misclassification. Model extraction [19] replicates a proprietary model by submitting queries and training a surrogate on input-output pairs, enabling downstream adversarial example generation. Membership inference attacks [17] determine whether a specific data record was used in model training; in the medical context, this can reveal sensitive health conditions or clinical trial participation. Model inversion attacks [18] reconstruct approximate training inputs from model outputs. A membership inference study on partially synthetic electronic health records achieved 82% patient re-identification in data derived from a major academic medical center [20].

E. Model Integrity Attacks

Model integrity attacks compromise stored model artifacts. Weight tampering (AML.T0031) directly modifies serialized model weights on the device's storage medium, causing systematic misclassification across all inputs without a detectable trigger pattern. Energy-latency attacks (sponge examples) craft inputs that maximize neural network activation, exhausting computational resources of constrained medical devices and causing availability failures under load.

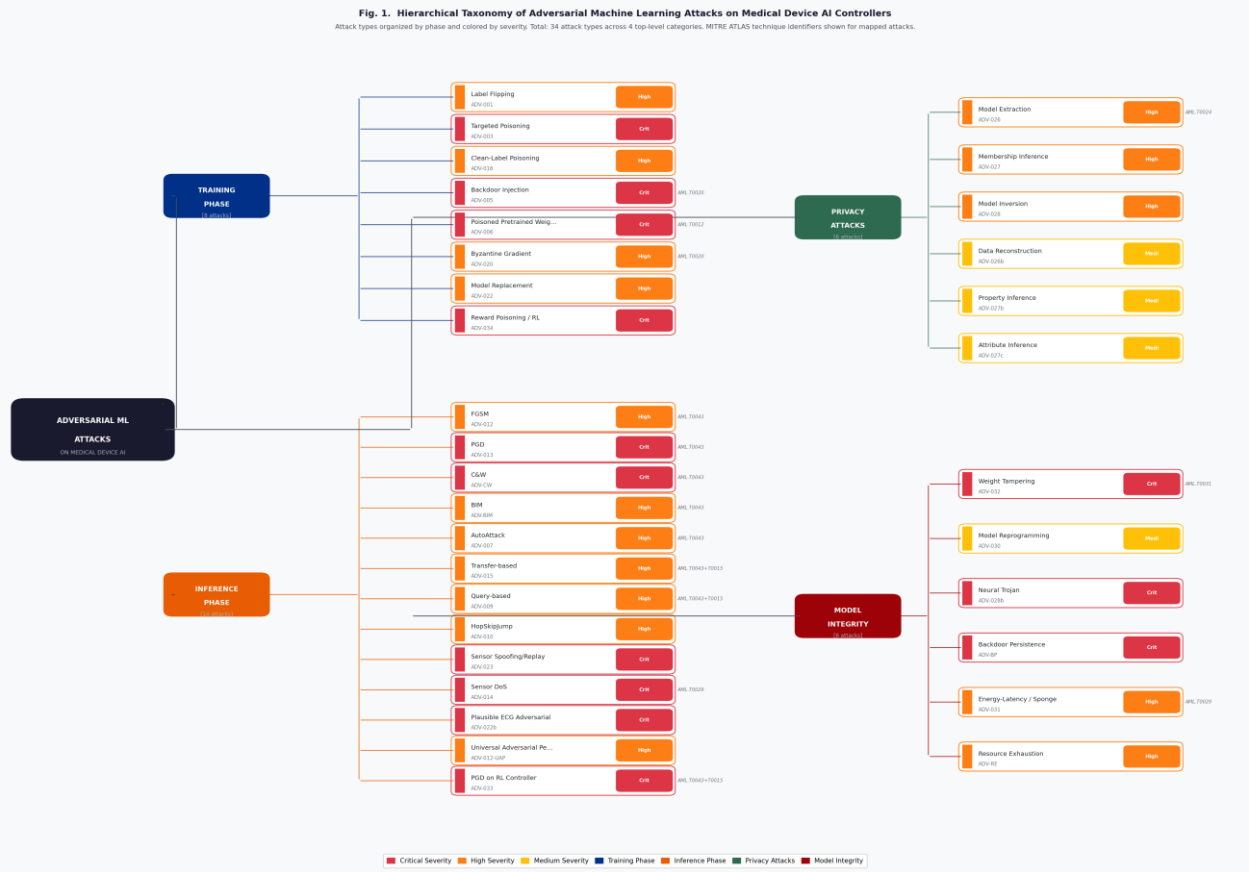

Figure 1. Hierarchical taxonomy of adversarial machine learning attacks on medical device AI controllers. Attack types are organized by phase and colored by severity (Critical: red, High: orange, Medium: yellow). MITRE ATLAS technique identifiers shown for mapped attacks. Total: 34 attack types across 4 categories.

Table I. Selected Adversarial Attack Taxonomy for Medical Device AI Controllers

ID	Attack Name	Category	Attacker Knowledge	Device Applic.	Severity	ATLAS Technique
ADV-001	**Label Flipping**	Training: Data Poisoning	Training data access	High	**High**	AML.T0020
ADV-003	**Targeted Poisoning**	Training: Data Poisoning	Training data access	High	**Critical**	AML.T0020
ADV-005	**Backdoor Injection**	Training: Data Poisoning	Training data access	High	**Critical**	AML.T0020
ADV-006	**Supply Chain / Poisoned Wts**	Training: Supply Chain	Pre-deploy access	High	**Critical**	AML.T0012
ADV-016	**Clean-Label Poisoning**	Training: Data Poisoning	Training data access	Medium	**High**	AML.T0020
ADV-020	**Byzantine Gradient (FL)**	Training: Federated	FL participant	High	**High**	AML.T0020
ADV-034	**Reward Poisoning (RL)**	Training: RL	Training access	High	**Critical**	AML.T0020
ADV-012	**FGSM Evasion**	Inference: White-box	Full model access	High	**High**	AML.T0043
ADV-013	**PGD Evasion**	Inference: White-box	Full model access	High	**Critical**	AML.T0043
ADV-CW	**C&W Evasion**	Inference: White-box	Full model access	High	**Critical**	AML.T0043
ADV-015	**Transfer-based Black-box**	Inference: Black-box	Surrogate model	High	**High**	AML.T0043+T0015
ADV-009	**Query-based Black-box**	Inference: Black-box	Inference API	Medium	**High**	AML.T0043+T0015
ADV-023	**Sensor Spoofing/Replay**	Inference: Physical	Physical access	High	**Critical**	—
ADV-014	**Sensor Denial-of-Service**	Inference: Physical	Physical/wireless	High	**Critical**	AML.T0029
ADV-022	**Plausible ECG Adversarial**	Inference: Time-series	White/black-box	High	**Critical**	AML.T0043
ADV-033	**PGD on RL Controller**	Inference: RL/Control	White-box	High	**Critical**	AML.T0043+T0015
ADV-026	**Model Extraction**	Privacy: Extraction	Inference API	Medium	**High**	AML.T0024
ADV-027	**Membership Inference**	Privacy: Membership	Inference API	Medium	**High**	Collection tactic
ADV-028	**Model Inversion**	Privacy: Inversion	Inference API	Medium	**High**	Collection tactic
ADV-032	**Weight Tampering**	Integrity: Manipulation	Physical/firmware	Medium	**Critical**	AML.T0031

Severity: Critical (direct life-threatening potential), High (significant patient safety risk), Medium (privacy/indirect). Device Applicability: proportion of six device component classes where the attack is plausibly executable. Color bands: blue = Training, orange = Inference, green = Privacy, red = Integrity.

IV. ATTACK SURFACE ANALYSIS

A. Medical Device AI Component Mapping

The attack surface of a medical device AI controller is distributed across six functional components: (1) the sensor layer (CGM electrochemical sensor, ECG electrodes), vulnerable to physical spoofing and electromagnetic interference; (2) the inference model (CGM prediction network, cardiac arrhythmia classifier), vulnerable to evasion and extraction; (3) the training pipeline and data storage, vulnerable to poisoning and supply chain compromise; (4) the actuator interface (insulin pump infusion commands, pacemaker stimulation parameters), the ultimate target of safety-impacting attacks; (5) the mobile application or gateway controller, presenting a software attack surface accessible to remote adversaries; and (6) the cloud backend, hosting training infrastructure, federated aggregation servers, and model repositories. Table II summarizes attack applicability across these components.

Table II. Attack Applicability by Medical Device AI Component

Component	Training Attacks	Inference Attacks	Privacy Attacks	Integrity Attacks	Key Notes
CGM Sensor	**Low**	**Critical**	**Low**	**Medium**	Spoofing/DoS demonstrated [12]
Insulin Dosing Model	**High**	**Critical**	**Medium**	**Critical**	AP RL attacks demonstrated [12]
Anomaly Detection	**High**	**Critical**	**Medium**	**High**	Evasion of autoencoder detectors [2]
Cardiac Classifier	**High**	**Critical**	**Medium**	**High**	Six attack methods validated on ECG
Mobile App / Gateway	**Medium**	**High**	**High**	**Medium**	Model extraction feasible
Cloud Backend	**Critical**	**High**	**Critical**	**High**	FL Byzantine; full pipeline exposure
Comm. Channel	**Low**	**High**	**Medium**	**Medium**	Replay/injection attacks

Ratings: Critical / High / Medium / Low based on published attack demonstrations and threat model analysis. Sources: [2] for anomaly detection, [12] for AP systems.

B. Trust Boundary Analysis

The medical device AI system operates across three distinct trust zones. The device trust zone encompasses firmware and on-device model weights, protected by secure boot and code signing in compliant devices but potentially exposed during firmware updates or physical access. The communication trust zone encompasses wireless and wired interfaces over which sensor data and software updates transit; replay and injection attacks operate at this boundary without requiring model access. The cloud trust zone encompasses training pipelines, model repositories, and federated learning servers, where training-phase attacks and supply chain compromises operate. An adversary breaching the cloud trust zone during training can implant backdoors that persist through the device trust zone after deployment, bypassing all inference-time defenses.

C. Representative Case Studies

Chang et al. demonstrated that a denial-of-service attack on the RL4BG AP controller produced blood glucose levels below the hypoglycemic threshold within 50 minutes, and that a PGD perturbation of ±1 mg/dL on CGM inputs increased dosing risk by 423% in children and prolonged hypoglycemic duration by 1,079% in adolescents using the UVa/Padova metabolic simulator [12]. Finlayson et al. demonstrated that standard adversarial attacks caused misclassification of diabetic retinopathy, pneumothorax, and melanoma classifiers with perturbations imperceptible to radiologists [10,11]. In cardiac monitoring, adversarial examples crafted with six attack methods achieved reliable misclassification on both the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database with physiologically plausible perturbations. A membership inference attack on partially synthetic electronic health records achieved 82% patient re-identification [20]. Medtronic MiniMed insulin pump systems (CVE-2019-10964) were recalled due to an unencrypted wireless interface enabling unauthorized remote insulin delivery commands, affecting approximately 4,000 U.S. patients.

V. DEFENSE LANDSCAPE

A. Defense Catalog Overview

Twenty-five defense mechanisms applicable to medical device AI controllers are cataloged herein, organized into four functional categories: training-phase defenses that reduce model vulnerability before deployment; inference-phase defenses that filter or detect adversarial inputs at runtime; privacy-preserving defenses that protect training data and model parameters from extraction; and integrity verification mechanisms that ensure deployed model artifacts have not been tampered with. Each defense is assessed against attack coverage, computational overhead, and device suitability. The defense-attack coverage matrix is visualized in Figure 2. A key structural finding is that no single defense mechanism provides coverage exceeding 42% of the identified attack categories. This result directly motivates the defense-in-depth strategy presented in Section VI, wherein multiple complementary mechanisms are layered according to device resource constraints.

B. Training-Phase Defenses

Adversarial training (DEF-001) in the PGD-AT formulation augments the training dataset with adversarially perturbed examples, forcing the model to learn decision boundaries robust to bounded perturbations [5]. It provides the strongest empirical evasion defense, reducing attack success by approximately 58% in medical AI evaluations [27], but is computationally feasible only during cloud-based training. Defensive distillation (DEF-002) trains a student model on the softened output probabilities of a teacher, reducing gradient magnitude and constraining the attack surface accessible to gradient-based adversaries [23]. Data sanitization and provenance tracking (DEF-009) counteract poisoning attacks by maintaining cryptographic audit trails of training data, enabling detection of label-flipping and backdoor injection through statistical distribution shift analysis. Neural Cleanse and activation clustering are post-training methods for detecting and removing backdoor triggers; both require significant computational resources and are feasible only in cloud or server environments.

C. Inference-Phase Defenses

Feature squeezing (DEF-003) applies input transformations—bit-depth reduction and spatial smoothing—to reduce the effective dimensionality of adversarial perturbations, providing lightweight detection feasible on devices with as little as 64 KB RAM [22]. Input denoising and signal preprocessing (DEF-004, DEF-018) apply domain-specific filters (Savitzky-Golay smoothing for CGM time-series, bandpass filtering for ECG signals) to attenuate high-frequency adversarial perturbations before inference; these are among the most resource-efficient defenses appropriate for all device tiers including severe-constraint implantables. Randomized smoothing (DEF-005) provides certified L2 robustness guarantees, delivering provable robustness bounds suitable for inclusion in FDA PCCP Impact Assessments [21]. The STRIP runtime trojan detection mechanism (DEF-007) identifies backdoor activation by measuring prediction entropy under strong input blending, requiring only a single additional forward pass per inference [26].

D. Anomaly Detection as a Unified Defense

Autoencoder-based input anomaly detection (DEF-012) monitors reconstruction error on incoming physiological signals to detect both adversarial perturbations and sensor faults in a unified framework. This

approach was demonstrated in the AP-GUARD system, where a convolutional autoencoder trained on normal CGM signal trajectories achieved a 205% improvement in anomaly detection rate compared to threshold-based baselines, with F1 > 0.75 across 3,456 experiments on simulated artificial pancreas data [2]. Runtime safety envelope enforcement (DEF-013) complements anomaly detection by bounding actuator outputs to physiologically safe ranges regardless of model output, providing an unconditional backstop implementable as a deterministic filter on any hardware tier.

E. Privacy and Integrity Defenses

Differential privacy via DP-SGD (DEF-010) provides formal privacy guarantees by injecting calibrated noise into training gradients [24]. A 74-study review established that a privacy budget of $\varepsilon \approx 10$ maintains acceptable diagnostic accuracy in most medical AI applications. Byzantine-robust aggregation (DEF-011), implemented via Krum, coordinate-wise median, or trimmed mean, mitigates Byzantine gradient attacks in federated learning [25]. Model integrity verification (DEF-015) employs cryptographic hash verification of stored model weights, providing a low-cost integrity check feasible on all device tiers.

Figure 2. Defense mechanism coverage matrix across adversarial attack categories. Green (✓) = effective, yellow (◐) = partial, red (–) = not effective. Coverage bars show the fraction of attack categories addressed by each defense. AP-GUARD [2] (DEF-012) and the Runtime Safety Envelope (DEF-013) achieve the highest individual coverage at 42%, confirming that defense-in-depth is required.

Table III. Defense Mechanism Catalog for Medical Device AI Controllers

ID	Defense Name	Attack Coverage	Compute	Min. Tier	Primary Attacks Countered
DEF-001	Adversarial Training (PGD-AT)	Evasion: High	High	Cloud	FGSM, PGD, C&W
DEF-002	Defensive Distillation	Evasion: Low-Med	Medium	Light	FGSM, PGD (partial)
DEF-003	Feature Squeezing	Evasion: Medium	Low	Severe+	FGSM, PGD, Transfer
DEF-004	Input Denoising	Evasion: Low-Med	Low	Severe+	FGSM, Sensor spoof (partial)
DEF-005	Randomized Smoothing	Evasion: Certified	Med-High	Light/Cloud	FGSM, PGD, C&W (certified L2)
DEF-006	Ensemble Methods	Evasion/Trans: Med	Medium	Light	Transfer, PGD (partial)
DEF-007	STRIP Backdoor Detection	Backdoor: Medium	Low	Moderate+	Backdoor, Neural Trojan
DEF-008	Neural Cleanse	Backdoor: Medium	High	Cloud	Backdoor, Weight tampering
DEF-009	Data Sanitization / Provenance	Poisoning: High	Medium	Training	Label flip, Targeted poison, Backdoor
DEF-010	Differential Privacy (DP-SGD)	Privacy: High	Medium	Cloud	MIA, Model inversion, Extraction
DEF-011	Byzantine-Robust Aggregation	FL Poisoning: High	Medium	Cloud FL	Byzantine gradient, Model replacement
DEF-012	**Autoencoder Anomaly Det. (AP-GUARD)**	Multi: 42%	Low-Med	**Moderate+**	FGSM, Sensor DoS, Spoofing, PGD
DEF-013	**Runtime Safety Envelope**	Multi: 42%	Very Low	**Severe+**	PGD, Sensor DoS, Spoofing, RL attacks
DEF-014	Output Statistical Monitor	Multi: Low-Med	Very Low	Severe+	FGSM, Availability attacks
DEF-015	Model Integrity Hash	Integrity: High	Very Low	Severe+	Weight tampering
DEF-016	Query Rate Limiting	Privacy: Med	Very Low	All	Model extraction, MIA (partial)
DEF-017	Activation Clustering	Backdoor: Medium	High	Cloud	Backdoor, Model inversion
DEF-018	Signal Preprocessing (ECG/CGM)	Evasion: Low-Med	Low	Severe+	ECG adversarial, Sensor spoof
DEF-019	Adversarial Distillation (CardioDefense)	Evasion: Medium	Medium	Light/Mod	ECG adversarial, PGD
DEF-020	Conformal Prediction	Evasion: Low-Med	Low	Light	FGSM, PGD (uncertainty bounds)

ID	Defense Name	Attack Coverage	Compute	Min. Tier	Primary Attacks Countered
DEF-021	Multi-modal Sensor Fusion	Physical: Med	Low-Med	Moderate+	Sensor spoofing, DoS
DEF-022	Manifold-Aware TinyML	Evasion: Medium	Low	Severe/Mod	FGSM, PGD (embedded-optimized)
DEF-023	Federated Secure Aggregation	FL Poisoning: Med	Medium	Cloud FL	Byzantine, Model replacement
DEF-024	Data Augmentation (Domain-specific)	Evasion: Low	Low	Training	FGSM, PGD (partial)
DEF-025	Hybrid Defense (AT + Preprocessing)	Evasion: High	Medium	Light+	FGSM, PGD, Transfer, Sensor spoof

Coverage: High (>50% reduction), Medium (25–50%), Low (<25%). DEF-012 and DEF-013 (highlighted in blue) achieve maximum 42% attack-category coverage. Defenses accessible to Severe-constraint tier (≤256 KB RAM) are listed as "Severe+".

VI. DEFENSE SELECTION AND REGULATORY ALIGNMENT

A. Resource-Constrained Defense Selection Framework

The defense selection framework maps each hardware resource tier to a prioritized defense stack. The framework employs a tiered inclusion model: each higher tier inherits all defenses from lower tiers and adds mechanisms that become computationally feasible with additional resources. Figure 3 provides a quadrant visualization of this framework across the axes of computational overhead and attack coverage breadth, and Table IV summarizes the recommended stacks per tier.

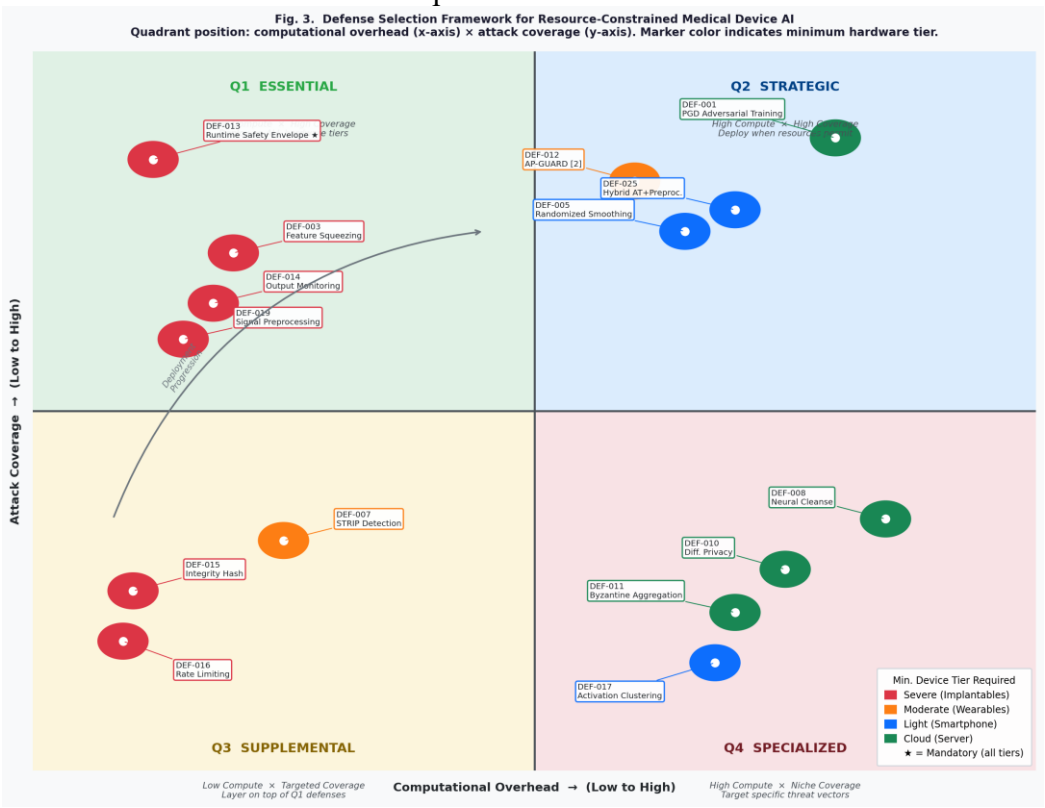

Figure 3. Defense selection quadrant for resource-constrained medical device AI. Q1 (Essential): low compute, high coverage — mandatory for all tiers. Q2 (Strategic): high compute, high coverage — deploy when resources permit. Q3 (Supplemental): low compute, targeted coverage. Q4 (Specialized): high compute, niche coverage. Marker color indicates minimum required device tier.

Table IV. Tiered Defense Recommendations by Device Resource Profile

Device Tier	Example Devices	Mandatory Defenses	Recommended Additions	Priority
Tier 1: Severe (≤256 KB RAM)	Implantable CGM sensors, pacemakers, implantable pump controllers	DEF-013 (Safety Envelope), DEF-014 (Output Monitoring), DEF-015 (Integrity Hash), DEF-018 (Signal Preprocessing)	DEF-003 (Feature Squeezing), DEF-016 (Rate Limiting), DEF-022 (TinyML Manifold)	Deploy immediately
Tier 2: Moderate (256 KB–1 MB)	Wearable CGM (Dexcom G7), insulin pumps (Medtronic 780G), wearable cardiac monitors	All Tier 1 + DEF-003, DEF-007 (STRIP), DEF-012 (AP-GUARD [2]), DEF-021 (Multi-modal Fusion)	DEF-004 (Input Denoising)	Short-term
Tier 3: Light (>1 MB RAM)	Smartphone AP controllers (Loop, AndroidAPS), bedside gateways, connected infusion pumps	All Tier 2 + DEF-005 (Randomized Smoothing), DEF-006 (Ensemble), DEF-019 (CardioDefense), DEF-025 (Hybrid AT+Pre)	DEF-002 (Distillation, student model)	Medium-term
Tier 4: Cloud (No constraint)	Training infra, FL servers, model repositories, analytics backends	All Tier 3 + DEF-001 (PGD-AT), DEF-009 (Data Sanitization), DEF-010 (DP-SGD, $\varepsilon \approx 10$), DEF-011 (Byzantine Agg.), DEF-008 (Neural Cleanse), DEF-017 (Activation Clustering)	DEF-023 (Secure FL Aggregation)	Comprehensive

Each tier inherits all defenses from lower tiers. Color bands correspond to hardware tiers: blue = Severe, green = Moderate, yellow = Light, orange = Cloud.

B. FDA Regulatory Alignment

The taxonomy and defense framework are aligned to three FDA guidance documents. The 2023 final cybersecurity guidance requires a threat model, SBOM, and security testing as conditions of marketing authorization [29]. Adversarial ML threats constitute a distinct threat class within the required threat model that is not addressed by conventional penetration testing and SBOM analysis alone; the taxonomy in Section III provides the threat enumeration vocabulary for this documentation. The January 2025 AI-DSF draft guidance [30] introduces robustness testing obligations for the AI/ML lifecycle management documentation, including post-market surveillance for performance degradation attributable to adversarial manipulation. The August 2025 PCCP final guidance [28] requires that each permissible model modification be accompanied by a Modification Protocol and Impact Assessment. Adversarial robustness metrics—robust accuracy against FGSM and PGD at specified ε, and reconstruction error thresholds for anomaly detection—constitute measurable, pre-specifiable performance bounds suitable for inclusion in PCCP Impact Assessments.

C. MITRE ATLAS Alignment for Premarket Documentation

MITRE ATLAS technique identifiers provide a standardized vocabulary for documenting adversarial ML threats in premarket cybersecurity submissions [33], analogous to the role of MITRE ATT&CK in traditional cybersecurity threat modeling. The 13 ATLAS techniques mapped in Table I can be incorporated directly into

the threat model section of a 510(k), De Novo, or PMA submission to satisfy the FDA's requirement for identification of cybersecurity risk sources. The NIST AI 100-2e2025 taxonomy [31] provides complementary terminology for the mitigation documentation component, enabling complete bidirectional traceability from identified attack type to implemented defense to regulatory requirement.

VII. DISCUSSION

A. Research Gaps and Open Problems

No standardized adversarial robustness benchmark exists for medical time-series data. The adversarial ML literature has converged on well-defined benchmark datasets for image classification—MNIST, CIFAR-10, ImageNet—against which robustness claims can be validated; no equivalent exists for CGM glucose trajectories, ECG waveforms, or photoplethysmography signals. Without standardized benchmarks, defense effectiveness claims in the medical time-series domain remain difficult to compare or reproduce. The development of adversarially annotated CGM and ECG datasets structured around the NIST AI 100-2e2025 attack taxonomy [31] represents a high-priority community contribution.

Defense mechanisms designed for server-grade GPU hardware have not been empirically validated on the ARM Cortex-M and embedded Linux platforms that constitute the actual deployment environment for medical device AI controllers. Resource feasibility ratings in this paper are derived from analytical assessment of computational complexity; empirical measurement of inference latency, memory footprint, and battery consumption on representative hardware remains outstanding. Such benchmarks are a prerequisite for any claim of deployment readiness.

Adversarial attacks targeting reinforcement learning controllers in medical devices are systematically understudied. The Chang et al. results on the RL4BG artificial pancreas controller represent the only published systematic study of adversarial attacks on an RL-based medical control system [12]; the broader class of RL-based drug delivery, rehabilitation, and cardiac stimulation controllers remains untested. The reward poisoning attack (ADV-034) and state-observation manipulation attack (ADV-033) cataloged in Section III lack medical device-specific empirical validation.

B. Limitations

The taxonomy is constructed from published literature and publicly documented attack demonstrations; undisclosed vulnerabilities in commercial medical device systems may not be represented. Clinical impact severity ratings are derived from simulation studies and may not fully capture the clinical heterogeneity of patient populations, device configurations, and care contexts. Defense feasibility ratings represent analytical upper bounds; actual deployability on a specific device platform requires hardware-level validation. The regulatory alignment discussion reflects FDA guidance documents available through March 2026; subsequent guidance revisions may alter specific documentation requirements.

VIII. CONCLUSION

This paper has presented the first comprehensive adversarial machine learning threat taxonomy and defense landscape analysis specifically designed for medical device AI controllers. A structured taxonomy of 34 adversarial attack types spanning training-phase data poisoning, inference-phase evasion, privacy extraction, and model integrity manipulation was developed and cross-referenced with MITRE ATLAS techniques and NIST AI 100-2e2025 terminology. Attack applicability was systematically analyzed across six medical device component classes, supported by documented case studies demonstrating clinically significant outcomes including life-threatening glucose dysregulation [12] and high-confidence medical image misclassification [10]. A catalog of 25 defense mechanisms was evaluated against four hardware resource constraint tiers, establishing that no single defense exceeds 42% attack coverage and that defense-in-depth strategies are necessary across all deployment contexts. Tiered defense recommendations were aligned to FDA cybersecurity guidance [29], the AI-DSF lifecycle management framework [30], and the PCCP documentation structure [28], providing device manufacturers with actionable premarket submission guidance. Building on prior work in artificial pancreas security [1] and autoencoder-based anomaly detection [2], the taxonomy and defense landscape together constitute a practitioner-oriented resource for engineering adversarially resilient AI-enabled medical devices. The medical device community is encouraged to adopt

adversarial robustness testing as a mandatory component of AI/ML development pipelines and regulatory submissions, commensurate with the direct patient safety consequences of adversarial failures.

REFERENCES:

[1] V. S. A. Piratla, "Security of AI-Enabled Artificial Pancreas Systems: A Systematic Review," in Proc. IEEE Int. Conf. Scalable Computing and Communications (ICSC), 2025.

[2] V. S. A. Piratla, "AP-GUARD: Autoencoder-Based Anomaly Detection for Artificial Pancreas Systems," in Proc. IEEE Int. Conf. Communication and Networking in Computer Networks (CICN), 2025.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. ICLR, 2014, arXiv:1312.6199.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. ICLR, 2015, arXiv:1412.6572.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018, arXiv:1706.06083.

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Security and Privacy, 2017, pp. 39–57.

[7] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Proc. IEEE EuroS&P, 2016, pp. 372–387.

[8] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in Proc. IEEE CVPR, 2017, pp. 1765–1773.

[9] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in Proc. ICML, 2020, pp. 2206–2216.

[10] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019.

[11] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," arXiv:1804.05296, 2019.

[12] P. Chang, V. Krish, and A. Rahmati, "Security analysis of RL-based artificial pancreas systems," in Proc. ACM CCS Workshop HealthSec '24, Oct. 2024, doi:10.1145/3689942.3694740.

[13] M.-J. Tsai, P.-Y. Lin, and M.-E. Lee, "Adversarial attacks on medical image classification," Cancers, vol. 15, no. 17, p. 4228, Aug. 2023.

[14] A. I. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "Adversarial attacks to machine learning-based smart healthcare systems," in Proc. IEEE GLOBECOM, 2020.

[15] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in Proc. ICML, 2012, pp. 1467–1474.

[16] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv:1708.06733, 2017.

[17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Security and Privacy, 2017, pp. 3–18.

[18] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. ACM CCS, 2015, pp. 1322–1333.

[19] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in Proc. USENIX Security Symp., 2016, pp. 601–618.

[20] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data—anonymisation groundhog day," in Proc. USENIX Security Symp., 2022, pp. 1451–1468.

[21] J. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proc. ICML, 2019, pp. 1310–1320.

[22] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in Proc. NDSS, 2018.

[23] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proc. IEEE Symp. Security and Privacy, 2016, pp. 582–597.

[24] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM CCS, 2016, pp. 308–318.

[25] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. NeurIPS, 2017, pp. 119–129.
[26] J. Gao, B. Wang, Z. Lin, W. Xu, and V. Shmatikov, "STRIP: A defence against trojan attacks on deep neural networks," in Proc. ACSAC, 2019.
[27] A. Gerhart and B. Iyangar, "Adversarially-aware architecture design for robust medical AI systems," arXiv:2510.23622, 2025.
[28] U.S. Food and Drug Administration, "Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions," FDA Final Guidance, Aug. 18, 2025.
[29] U.S. Food and Drug Administration, "Cybersecurity in medical devices: Quality system considerations and content of premarket submissions," FDA Final Guidance, Sep. 2023.
[30] U.S. Food and Drug Administration, "Artificial intelligence-enabled device software functions: Lifecycle management and marketing submissions," FDA Draft Guidance, Jan. 7, 2025.
[31] National Institute of Standards and Technology, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," NIST AI 100-2e2025, Mar. 2025.
[32] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
[33] MITRE Corporation, "ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems," atlas.mitre.org, Oct. 2025 update.