Systems-Theoretic and Data-Driven Security Analysis in ML-enabled Medical Devices

Gargi Mitra[1][0000−0001−8011−4590], Mohammadreza Hallajiyan[1][0000−0003−1570−7351], Inji Kim[2][0009−0000−9211−8433], Athish Pranav Dharmalingam[3][0009−0000−7326−4662], Mohammed Elnawawy[1][0000−0002−4367−8060], Shahrear Iqbal[4][0000−0001−7819−5715], Karthik Pattabiraman[1][0000−0003−2380−3415], and Homa Alemzadeh[2][0000−0001−5279−842X]

[1] The University of British Columbia, Vancouver, British Columbia, Canada
{gargi, hallaj, mnawawy, karthikp}@ece.ubc.ca
[2] University of Virginia, Charlottesville, Virginia, USA
{ddh8jk, ha4d}@virginia.edu
[3] Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
cs21b011@smail.iitm.ac.in
[4] National Research Council Canada
shahrear.iqbal@nrc-cnrc.gc.ca

Abstract. The integration of AI/ML into medical devices is rapidly transforming healthcare by enhancing diagnostic and treatment facilities. However, this advancement also introduces serious cybersecurity risks due to the use of complex and often opaque models, extensive interconnectivity, interoperability with third-party peripheral devices, Internet connectivity, and vulnerabilities in the underlying technologies. These factors contribute to a broad attack surface and make threat prevention, detection, and mitigation challenging. Given the highly safety-critical nature of these devices, a cyberattack on these devices can cause the ML models to mispredict, thereby posing significant safety risks to patients. Therefore, ensuring the security of these devices from the time of design is essential. This paper underscores the urgency of addressing the cybersecurity challenges in ML-enabled medical devices at the pre-market phase. We begin by analyzing publicly available data on device recalls and adverse events, and known vulnerabilities, to understand the threat landscape of AI/ML-enabled medical devices and their repercussions on patient safety. Building on this analysis, we introduce a suite of tools and techniques designed by us to assist security analysts in conducting comprehensive premarket risk assessments. Our work aims to empower manufacturers to embed cybersecurity as a core design principle in AI/ML-enabled medical devices, thereby making them safe for patients.

Keywords: AI/ML-enabled medical devices · Security assessment · Safety assessment · System-theoretic security analysis · AI/ML security.

1 Introduction

Machine Learning (ML)-driven applications are becoming increasingly popular in the medical field. ML-enabled medical devices (software or software-driven hardware) assist physicians in critical activities such as remote patient monitoring, controlling surgical equipment, automatic drug administration, and preliminary/advanced disease diagnosis [110]. These tasks require high accuracy and reliability, and the loss of either of these can endanger patient safety. However, the use of ML in interconnected medical devices has expanded the threat surface of medical systems [22,4,68,80,64,20,81,124,71,24,133,89,78,59,77] making them more vulnerable to cyberattacks. If an adversary compromises such a device, it can force the ML engine to make incorrect predictions or decisions, which can have catastrophic consequences, such as wrong diagnoses and treatments, leading to severe health complications or even the death of the patient.

Detecting and mitigating cyberattacks in ML-based medical applications is significantly more challenging than in traditional systems for two reasons. First, these applications rely on large datasets and often employ complex, unexplainable models, making their behavior difficult to interpret even for developers. Second, they are highly interconnected with third-party devices that collect patient data for real-time predictions, which are subsequently transmitted to downstream systems or patients and physicians for clinical decision-making and treatment. This high degree of connectivity increases the attack surface, while the complexity of ML models complicates attack detection. Adversaries can exploit the vulnerabilities in the ML models and interface devices to poison training data [88], inject erroneous inputs during inference [49], or modify model parameters through compromised configuration files [126]. We are particularly interested in inference-time false data injection attacks, which are the easiest to execute and most difficult to detect. A recent study on the FDA adverse event reports involving ML-enabled medical devices indicated that over 80% of the reported events were related to data acquisition problems, leading to no data or erroneous data capture [76]. Several studies have also highlighted the vulnerability of data acquisition systems to adversarial examples [50,57,20]. The safety-critical nature of ML-enabled medical devices makes it crucial to identify and address these security vulnerabilities before deployment.

In this paper, we focus on pre-market security risk assessment, which is the process of identifying, assessing the severity of, and mitigating potential security risks in a given medical device before it is approved for market release. This process is crucial for ensuring patient safety, regulatory compliance, and cyber resilience, as well as reducing post-deployment threat mitigation costs. We inspected the publicly available device summaries [118] the manufacturers submitted to the FDA for pre-market approval to determine whether manufacturers conducted security risk assessments for ML-enabled devices. Our investigation reveals that for over 65% of these devices, the manufacturers either do not provide any information about the assessment method in their documentation, or employ inadequate assessment methods (see Figure 1). In fact, until 2014, no device summary mentioned any security risk assessment. Among the remaining devices,

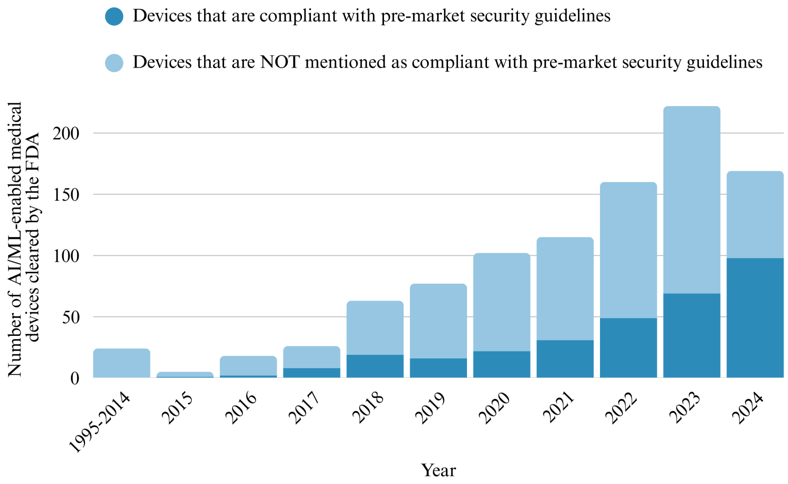


Fig. 1: Growing number of AI/ML-enabled medical devices and the rise of security-awareness among device manufacturers (Data as of April 2025)

a few use proprietary mechanisms that make it challenging to assess the adequacy of their approach, while others utilize existing risk assessment techniques. These techniques, as we discuss in the subsequent sections, are insufficient for securing interconnected ML-enabled medical systems. However, on a positive note, there is a growing security awareness among manufacturers, reflected in the increasing mention of security risk assessments in recent pre-market summaries.

Security practitioners and researchers have made significant efforts in assessing and ensuring the safety and security of medical devices by developing advanced methods for qualitative and quantitative risk assessment (e.g., fault tree analysis (FTA), failure mode and effect analysis (FMEA)) and formal assurance case reports [9,62,61]) and security analysis [14], model-based design and verification [16,95,7], closed-loop validation [63], encryption, and authentication [65]. However, less attention has been paid to the *end-to-end system security* of ML-enabled medical devices by considering the interactions of the ML-enabled device with other interconnected system components. Current security assessment methods primarily focus on algorithm, hardware, software, and firmware vulnerabilities, but they often overlook the *inherent vulnerabilities of the ML models* used in medical devices, how they can be exploited by first exploiting vulnerabilities in interconnected devices, and the potential impact of the ML mispredictions on *patient safety. To bridge this gap, it is imperative to perform a holistic system-theoretic analysis of ML-enabled medical systems.*

In this paper, we first present our experience with developing tools and techniques to automate the extraction of large-scale data on real-world security vulnerabilities and safety incidents for ML-enabled medical devices from public data and knowledge sources. Further, we devise techniques that use this data to enable security practitioners to perform system-theoretic analysis to identify potential threats, new attack paths, and their safety impacts. This will help medical device manufacturers anticipate post-deployment security risks early at design time, assess the severity of the risks, and implement risk prevention and mitigation strategies. For instance, a company developing an ML-enabled device that integrates with third-party commodity off-the-shelf cameras can use our techniques to identify known security vulnerabilities in compatible camera models, evaluate the likelihood of their exploitation, and assess potential risks to patient safety based on previously reported failures and adverse events of both ML-enabled and non-ML-enabled devices with the same functionality. Based on these insights, the company can either implement appropriate security measures and safety mechanisms or provide guidance to users to avoid connecting vulnerable cameras to the device. We demonstrate our tools and techniques on various ML-enabled medical devices, particularly blood glucose management systems (BGMS), as an example of safety-critical personalized devices with a broad and complex attack surface due to their high levels of connectivity and interoperability.

2 Background and Motivation

This section provides the technical background required to understand the subsequent sections and the motivation behind our research.

2.1 AI/ML-enabled Medical Devices

As of December 2024, the U.S. Food and Drug Administration (FDA) has authorized more than $1,016$ ML-enabled medical devices across 17 different medical disciplines (e.g., Cardiology, Ophthalmology, and Gastroenterology) [110]. These devices can be categorized into two types: Software as a Medical Device (SaMD) and Software in a Medical Device (SiMD). An SaMD is software that can be run on general-purpose computers (e.g., d-Nav for predicting insulin dose for diabetic patients [112]), whereas an SiMD is software that is sold bundled with hardware manufactured by the same company (e.g., GI Genius Intelligent Endoscopy Module [115]). Our analysis of the FDA data shows that while radiological imaging devices are the most common category of FDA-cleared ML-enabled devices (76.5%), safety-critical devices in clinical chemistry (e.g., BGMS), cardiovascular (e.g., arrhythmia diagnosis devices), and neurology (e.g., surgical procedures planning systems) have relatively higher numbers of reported adverse events (see Figure 5). Unlike radiological devices, most personalized cardiac monitors and BGMS are mobile-based devices used by patients in the absence of continuous medical supervision. Their compatibility with peripheral devices from multiple brands and various communication protocols creates a broad and complex attack surface, making them highly susceptible to false data injection attacks with potentially severe consequences. These factors make such

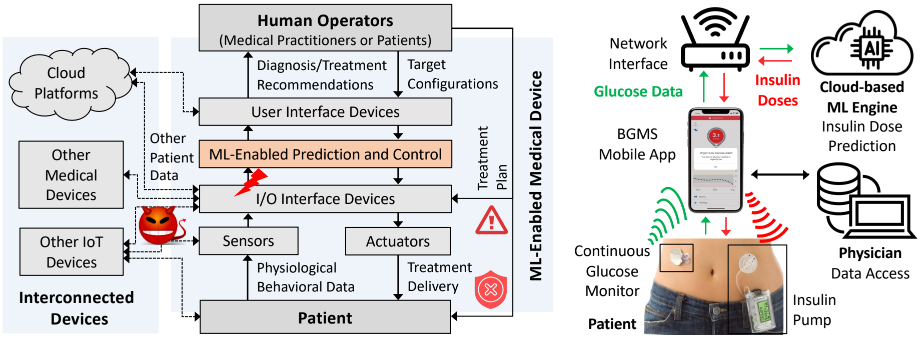


Fig. 2: Left: Typical System Control Structure of Interconnected ML-enabled Medical Cyber-Physical Systems. Right: Example ML-enabled Blood Glucose Management System (BGMS) Authorized by the U.S. FDA.

devices a compelling choice for our evaluation. Table 1 shows examples of ML-enabled BGMS, including d-Nav [112], WellDoc BlueStar [117], Dreamed Advisor Pro [114], Dario Blood Glucose Monitoring System [113], and the One Drop Blood Glucose [116] Monitoring System.

Interconnected medical devices. The ML models in these devices typically receive inputs from multiple sensory devices that collect various physiological data from a patient's body to predict their condition. Moreover, they can interface with third-party software, cloud platforms, and IoT devices, creating a highly interconnected system. For instance, as shown in Figure 2 (Right), an ML-based diabetes management app such as d-Nav can be installed on a mobile phone. It contains two user-interactive software elements - one for the patient and one for the physician. The system can receive glucose measurement data entered manually into the patient user software or automatically via the cloud from a linked blood glucose meter or continuous glucose monitor (CGM). Some backend components run locally on the phone, while others may be hosted either locally or in the cloud [112].

Interoperability in AI/ML-enabled medical devices. In recent years, there has been a growing trend toward enhancing *interoperability*, particularly in AI/ML-enabled medical systems. For example, to promote modular integration across BGMS from different manufacturers, the FDA has introduced a framework identifying three essential components in Automated Insulin Delivery (AID) systems, including Alternate Controller Enabled (ACE) pumps, interoperable CGMs (iCGMs), and interoperable glycemic controllers (iAGCs), that can reliably and securely communicate with digitally connected devices to send, receive, and execute drug delivery commands [35]. Motivated by a broader patient movement towards open and personalized configurations [106,94], several interoperable AID systems have gained FDA approval. Table 2 shows seven FDA-approved AID systems among which five incorporate officially designated interoperable components (ACE pump, iCGM, and iAGC). A recently approved iACG, Tidepool [106], supports a wide range of compatible CGMs and insulin pumps from different

Table 1: A study of different FDA-Approved ML-enabled medical devices and their security vulnerabilities that enable false data injection attacks †: SaMD, ‡: SiMD, *: Best-guessed ML algorithm, $^{\$}$YoA: Year of Approval, Ⓛ: Only locally exploitable vulnerability, Ⓡ: Remotely exploitable vulnerability

Device Name [110]	**YoA$^{\$}$**	**Device Function**	**ML Technique used**	**Known ML attacks**	**Possible third-party attack entry points {Known vulnerablity}**	**Potential impact of mis-prediction**
d-Nav System†	'19	Insulin dose prediction	Reinforcement learning*	[34]	Android vulnerabilities [31]	Wrong treatment (Fatal)
WellDoc BlueStar†	'19	Diabetes management	Light Gradient Boosting Machine*	[15]	Cloud Service API [29]	Wrong diagnosis
Dreamed Advisor Pro‡	'19	Diabetes management	Reinforcement learning*	[75]	Blood glucose meter [27]	Wrong treatment (Fatal)
Dario BGMS‡	'15	Diabetes management	k-means clustering*	[25]	Android vulnerabilities [31]	Wrong treatment (Fatal)
One Drop BGMS‡	'16	Diabetes management	Long short-term memory*	[104]	Bluetooth	Wrong treatment (Fatal)
Mammo-Screen‡	'24	Breast cancer detection	Deep learning	[74]	PACS server {[26]}Ⓡ	Wrong diagnosis
CardioLogs ECG Analysis Platform†	'17	Cardiac arrhythmia detection	Deep Neural Network (DNN)	[22]	Portable ECG Monitors - {[82]} Ⓛ, Cellular network, Bluetooth	Wrong treatment (Fatal)
GI Genius‡	'21	Gastro-intestinal lesion detection	Convolutional neural networks (CNN)*	[58]	Endoscope cameras - {[84]} Ⓡ, Intranet / Internet	Wrong diagnosis
NuVasive Pulse System‡	'18	Neurological monitoring	CNN*	[58]	Infra-red sensitive cameras - [125] Ⓛ, {[85]} Ⓡ, Internet	Mistake in surgery (Fatal)
Air Next‡	'20	Spirometer	CNN*	[58]	Bluetooth, Internet	Wrong diagnosis
BrainScope TBI‡	'19	Brain injury assessment	Regularized logistic regression model	[23]	Internet	Wrong treatment (Fatal)
IDx-DR v2.3†	'22	Diabetic Retinopathy Detection	CNN	[58]	This device uses the Topcon NW200 Fundus camera, which comes packaged with a PC running Windows 7 OS. The Windows 7 OS has known vulnerabilities - {[86]} Ⓡ, Internet	Wrong diagnosis (loss of vision)
Iris Intelligent Retinal Imaging System†	'15	Storage, management and display of retinal images	Deep Learning	[130]	Same as in the case of IDx-DR v2.3, Internet	Wrong diagnosis (loss of vision)
Paige Prostate†	'21	Cancer diagnosis	CNN + Recurrent neural networks	[53]	Medical scanners - {[83]} Ⓛ, Internet	Wrong diagnosis (Fatal)
Tissue of Origin Test Kit‡	'18	Malignant Tumor diagnosis	SVM	[77]	Internet	Wrong diagnosis (Fatal)

Table 2: Examples of FDA-approved Automated Insulin Delivery (AID) systems that support interoperability in connected components. Modified from [66].

AID System	**FDA Approval Date**	**Pump**	**AGC (Control Algorithm)**	**CGM**
Beta Bionics iLet Bionic Pancreas	05/19/2023	iLet	iLet Dosing Decision Software	Dexcom G6, Dexcom G7
Insulet Omnipod 5	08/26/2024	Omnipod 5 / DASH	SmartAdjust algorithm	Dexcom G6, Dexcom G7
Tandem Mobi	07/11/2023	Mobi	Control-IQ algorithm	Dexcom G6, Dexcom G7
Tandem t: slim X2	12/13/2019	t:slim X2	Control-IQ algorithm	Dexcom G6, Dexcom G7
Medtronic MiniMed 770G	09/01/2020	MiniMed 770G	SmartGuard technology	Guardian Sensor 3, FreeStyle Libre 2 Plus
Medtronic MiniMed 780G	04/21/2023	MiniMed 780G	SmartGuard technology	Guardian Sensor 3, Guardian Sensor 4
Twiist	04/02/2025	Deka insulin pump	Tidepool Loop algorithm	FreeStyle Libre 3 Plus

manufacturers (such as Medtronic, Tandem, Omnipod, and Dexcom) and is used in a newly FDA-approved AID, called Twiist. Although the current AID devices on the market are not ML-enabled, some of them adopt smart model-predictive control (MPC) algorithms (e.g., SmartAdjust in Omnipod 5 [60], Control-IQ by Tandem t) that are envisioned to use ML in the near future [97]. A similar trend towards growing interoperability in other ML-enabled diabetes management systems is also expected to happen.

This shift towards interconnectivity and interoperability underscores an urgent need for comprehensive system-level security analysis in ML-enabled medical devices. As interconnectivity increases, potential vulnerabilities such as data breaches, insecure interfaces, and compromised control integrity must be proactively addressed through secure-by-design architectures.

2.2 Security Vulnerabilities in ML-enabled Medical Devices

The draft guidance containing recommendations for AI-enabled device software functions, published in 2025 by the U.S. Food and Drug Administration (FDA) agency, highlights a number of ML risks that are susceptible to cybersecurity threats [111]. These include data poisoning, model inversion/stealing, model evasion, data leakage, overfitting, model bias, and performance drift caused by adversaries. The highly interconnected nature of ML-based medical devices provides a multitude of attack vectors to adversaries. This is also evident from an increasing number of reported recalls, adverse events [9], and security vulnerabilities [67,128,92,1] and demonstrated attacks on medical devices across various clinical specialties [56,72,19,7]. A recent study [1] on over 966 medical devices from 117 vendors found 993 vulnerabilities across medical hardware, operating systems, and software applications. Further, 160 of these had publicly-available exploits that could allow the attackers to target patients and healthcare organizations. The majority of these vulnerabilities were found in health IT applications (741) and moderate-risk devices (292) such as medical imaging and monitoring/telemetry devices and infusion pumps.

2.3 Threat Model

In this work, we focus on false data injection attacks, a significant threat to interconnected medical devices. An adversary can force an ML engine to generate incorrect predictions or decisions by injecting carefully crafted malicious data through the data acquisition system during inference [22,77].

Preventing such attacks in ML-enabled medical devices is particularly challenging due to their interconnectivity with several other peripheral and sensor devices and networks. Figure 2 (Left) shows the various components of an ML-enabled medical system. Adversaries can exploit vulnerabilities in any of the third-party medical and Internet of Things (IoT) devices on the hospital network and/or interface and network devices to find their way into a target ML-enabled device, and inject malicious data into the ML engine even if the ML-enabled device is not compromised. Therefore, it is not enough to secure only the ML-enabled devices. A recent notification by the Federal Bureau of Investigation (FBI) indicated that about 53 percent of connected medical and IoT devices in hospitals have known critical vulnerabilities [92] that could enable such attacks. In our prior work [33], we manually analyzed 15 ML-enabled medical devices across various disciplines to examine their ML models, known vulnerabilities, and potential attack vectors in peripheral devices for false data injection during inference. Table 1 summarizes our findings, revealing that 11 of 15 devices were susceptible to false data injection attacks, with consequences ranging from vision loss to patient death.

2.4 Systems-Theoretic Safety and Security Analysis

Given the highly interconnected nature of ML-enabled medical devices, ensuring their security and safety requires a comprehensive, system-level approach that accounts for complex interactions between components. Modern system-theoretic approaches to safety and security of interconnected devices, such as STAMP (Systems-Theoretic Accident Model and Processes) [69], model accidents as complex processes resulting from safety and security constraint violations due to inadequate controls. Systems are represented as *hierarchical control structures*, with each level constraining the one below and communicating their conditions and behavior to the upper levels. System-Theoretic Process Analysis for Security (STPA-Sec) [131] and Causal Analysis using System Theory (CAST) [69], built upon STAMP, analyze hardware, software, physical systems, and human operators across control layers to pinpoint threat scenarios, security exploits, unsafe actions, and their causal factors. To assess ML-enabled device vulnerabilities, analysts must (i) model the device's control structure and (ii) identify technologies (e.g., protocols, software, OS, firmware) used in each component.

While several tools support STPA and STPA-Sec across various domains, we assess their suitability for ML-enabled medical systems based on three key

Table 3: Summary of State-of-the-Art STPA/STPA-Sec tools
(All these tools generate causal scenarios in semi-automated fashion)

Name	Focus	Application Domain
A-STPA, XSTAMPP	Safety	General Purpose
SafetyHAT	Safety	Transportation
WebSTAMP	Safety/Security	Healthcare, Transportation, Chemical Industry
SOT	Safety/Security	Aircraft Systems

features: (i) applicability to security attacks on ML systems, (ii) applicability to the medical domain, and (iii) automation of causal scenario generation. To this end, we evaluate four state-of-the-art STPA/STPA-Sec tools: **(1)** A-STPA [2] and its enhanced version, XSTAMPP [3] assist in linking unsafe control actions to safety hazards and provide graphical aids for control structure creation but require manual causal scenario identification; **(2)** SafetyHAT [18] is customized for the transportation sector. This tool offers a graphical interface, data management, and domain-specific guidewords but lacks automated causal scenario identification; **(3)** WebSTAMP [103] is a web application designed for STPA and STPA-Sec, that provides structured guidance for identifying hazardous control actions and causal scenarios. It has been applied to Glucose Monitoring and Insulin Pumping System, transportation applications[105], and chemical reactors[132]; and, **(4)** SOT [98] – this tool helps systems engineers conduct safety and security analyses by leveraging past knowledge to identify causal scenarios.

In summary, A-STPA, XSTAMPP, and SafetyHAT focus on safety risks from device failures, not malicious attacks. While WebSTAMP and SOT consider security concerns, they still rely on users' knowledge of vulnerabilities and require significant manual effort. Table 3 shows a summary of these tools.

Recent papers such as the survey by Qi et al. [99] explore the use of STPA in learning-enabled systems, and introduce DeepSTPA for analyzing ML lifecycle failures, which is beyond our scope. Other recent papers [100,91,90] explore the usability of Large Language Models (LLMs) in STPA, highlighting the need for human intervention in generating prompts and validating LLM responses. However, none of these studies focus on medical device security.

2.5 Medical Device Databases

The U.S. Food and Drug Administration (FDA) regulates medical devices sold in the US, and maintains several publicly available databases on premarket and postmarket data about cleared and approved medical devices, including device summary information, approval date, user instructions, and information on Premarket Approvals (PMA), Premarket Notifications (510[k]), Adverse Events, and Recalls. We analyze the following FDA databases to extract the information about medical device technologies and their reported safety and security flaws:

AI/ML-Enabled Medical Devices database [110] maintains the information about the FDA-authorized medical devices that incorporate AI/ML across medical disciplines. This data is not comprehensive and only contains releasable information about devices based on information provided in the summary descriptions of their marketing authorization document.

Premarket Notifications (510(k)s) database [118] contains the releasable records of premarket notifications submitted by medical device manufacturers for the devices introduced into commercial distribution for the first time or those reintroduced with significant changes. Each record includes device classification and approval information as well as summaries of device functionality and safety and effectiveness information for more recent submissions.

Recalls database [122] contains records of medical device recalls since November 01, 2002. A recall is a voluntary action that a manufacturer takes to correct or remove from the market any medical device that violates the FDA's laws. Each record in the database contains the information on a recalled device such as the product name, manufacturer name, number of devices on the market, recall class, FDA determined cause, and the human-written textual descriptions of manufacturer's reason for recall and recovery actions taken to correct the device or remove it from the market.

Manufacturer and User Facility Device Experience database [119] **(MAUDE)** is a collection of adverse events of medical devices that volunteers, user facilities, manufacturers, and distributors have reported to the FDA. Each adverse event report contains information such as device and manufacturer names, event type (e.g., Malfunction, Injury, or Death), event and report dates, and human-written event description and manufacturer narratives, which provide a short textual description of the incident, as well as any comments made or follow-up actions taken by the manufacturer to detect and address device problems.

2.6 Vulnerability Databases

We analyze the following publicly available vulnerability databases to identify common threats and security attacks targeting medical devices and peripheral devices:

ICS-CERT Alerts dataset [108] is developed and maintained by Industrial Control Systems Cyber Emergency Response Team and the United States Computer Emergency Readiness Team (US-CERT). US-CERT is responsible for analyzing and reducing cyber threats, vulnerabilities, disseminating cyber threat warning information, and coordinating incident response activities.

MITRE Common Vulnerability Enumeration (CVE) database [87]is a publicly accessible registry of known cybersecurity vulnerabilities, maintained by the MITRE corporation. It provides a comprehensive database of vulnerabilities, including those affecting peripheral medical and IoT devices used in medical systems. The data is contributed by software vendors, security researchers, penetration testers, as well as independent researchers.

Despite these publicly available databases, the information on ML-enabled medical devices and the peripheral devices they connect to is only available in a dispersed and unstructured format. It is particularly challenging to (i) extract relevant information on security vulnerabilities and safety impacts from the

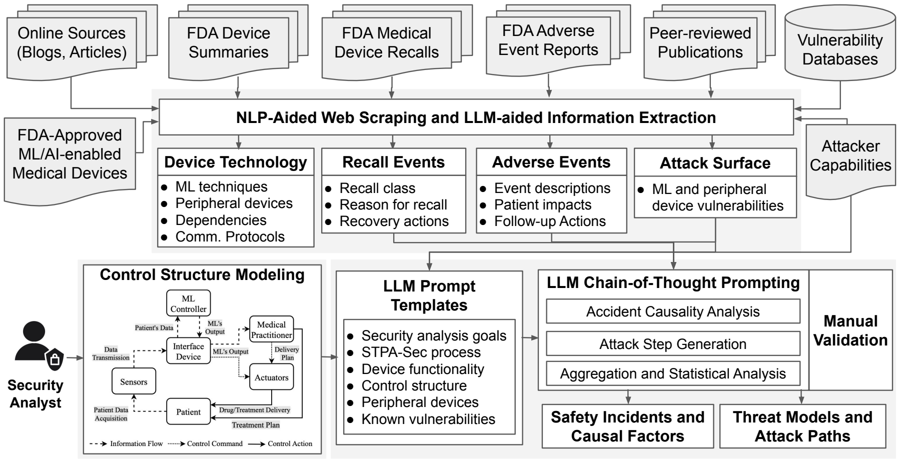


Fig. 3: Overall Approach for Systems-Theoretic and Data-Driven Analysis of Safety and Security of ML-enabled Medical Devices

dispersed data across millions of records in different databases and (2) analyze the free-form natural language text, written by the manufacturers and healthcare professionals, while understanding semantics and the contextual factors involved in the events. In the subsequent sections, we describe how some of the tools and techniques we developed alleviate the aforementioned challenges.

3 Methods

This section presents our framework for performing holistic system-theoretic analysis of ML-enabled medical systems. The framework comprises a suite of Natural Language Processing (NLP) and LLM-aided tools and techniques to assist the systems and control-theoretic security analysis of ML-enabled medical devices. Specifically, we report our experience on the design and validation of tools for semi-automated device modeling and technology identification, information extraction, , and systems-theoretic accident causality analysis and attack step generation.

Figure 3 shows an overview of our framework, which consists of three main components.

The **first** component is a device modeling and technology identification technique. It helps security analysts model the interconnections and communications between the ML-enabled medical device and third-party peripherals as a control structure using a generic control structure template for ML-enabled medical devices. It also assists in identifying all technologies used in connected devices that could serve as potential attack entry points.

The **second** component is a set of NLP and LLM-aided web scraping and information extraction techniques to extract and cross-reference the information from publicly available databases. Given a natural language description of an

ML-enabled medical device, it extracts key details, including the ML technique used, connected peripherals, and device functionality. Using this information, it scrapes the web for information on relevant ML vulnerabilities that adversaries could exploit to induce misprediction. Additionally, it interfaces with public FDA medical device and vulnerability databases to identify vulnerabilities in similar medical devices and peripheral devices, as well as recalls and adverse events linked to their malfunctions and safety impacts.

The **third** component of this framework is an LLM-based tool that can assist security analysts in systems-theoretic and data-driven safety accident (CAST) and security (STPA-Sec) analysis. This tool integrates the knowledge of CAST and STPA-Sec processes with the extracted information on device technology, control structure, and vulnerabilities and encodes them as customized prompt templates that can guide LLMs in generating (i) a comprehensive list of safety issues and causal factors that could lead to patient harm and (ii) attack vectors that adversaries could exploit to deliberately trigger such safety events.

In the following subsections, we discuss these tools/techniques in detail. We also illustrate how each tool/technique contributes to the overall security assessment, by providing examples of their output when applied to ML-enabled devices, such as the BGMS in Figure 2.

3.1 Device Modeling and Technology Identification

To identify all possible attack vectors in a given ML-enabled medical system, a security analyst first needs to model the interconnections of the ML-enabled medical device with the peripheral devices, the data flow between various system components, and understand the technology used by various system components. To enable the systems-theoretic security analysis using STPA-Sec (in Section 3.3), we adopt the hierarchical system control structures from STAMP (see Section 2.4) for this purpose.

System Control Structure Modeling. In our previous work (SAM [54]), we partially automated the construction of the system control structure for ML-enabled medical devices by developing a template control structure that contains typical components and interconnections in an ML-enabled medical system. The security analysts could customize this generic control structure by adding/removing necessary components and interactions to match the description of the system under assessment. Once the control structure is built, the security analyst must manually identify the data flows among various system components from the device descriptions. We applied this technique on two ML-enabled medical devices – (1) d-Nav [112], a blood glucose monitoring system, and; (2) ABMD [109], a bone mineral density calculator. We built the control structure and inferred the data flow using the system description provided by the manufacturer, which we obtained from publicly available device summaries submitted to the FDA during the pre-approval process, as well as from information available on the manufacturers' websites. Figure 4 shows the control structure for d-Nav [112], as

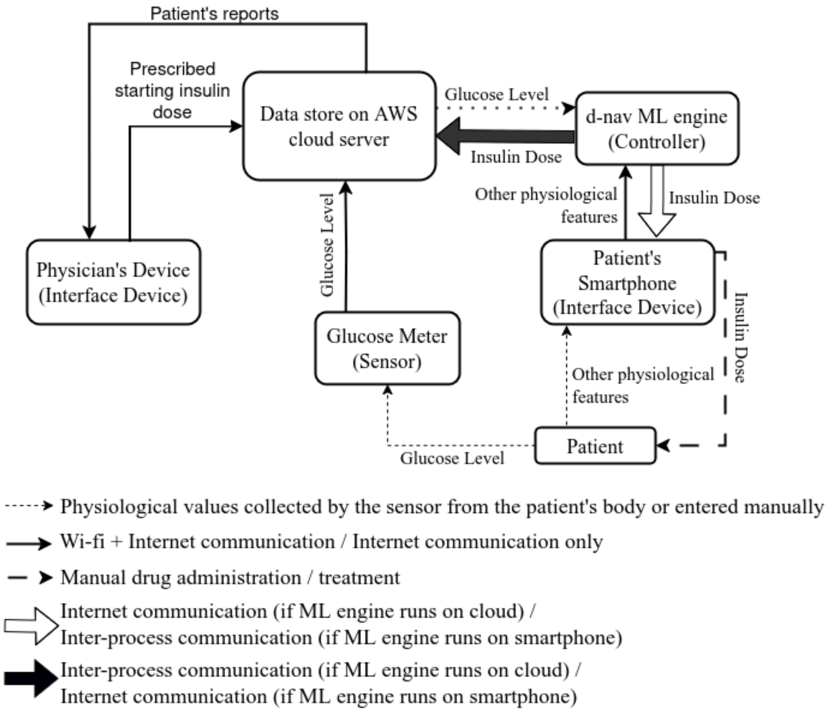

Patient's reports
Prescribed starting insulin dose
Data store on AWS cloud server
Glucose Level
d-nav ML engine (Controller)
Insulin Dose
Other physiological features
Insulin Dose
Glucose Level
Physician's Device (Interface Device)
Patient's Smartphone (Interface Device)
Insulin Dose
Glucose Meter (Sensor)
Other physiological features
Patient
Glucose Level
Physiological values collected by the sensor from the patient's body or entered manually
Wi-fi + Internet communication / Internet communication only
Manual drug administration / treatment
Internet communication (if ML engine runs on cloud) / Inter-process communication (if ML engine runs on smartphone)
Inter-process communication (if ML engine runs on cloud) / Internet communication (if ML engine runs on smartphone)

Fig. 4: Control Structure of the d-Nav System. Note that, while setting up the system, the ML engine can be configured to run either on the smartphone or on the cloud server.

generated by the technique proposed in SAM. Note that these documents do not follow a standardized format - the information is often dispersed across multiple sites, and the transparency varies across manufacturers. Hence, automating the information retrieval process remains a challenge.

Technology Identification. Once the security analyst builds the control structure, they must identify the technologies used across system components, such as the ML techniques, operating systems, firmware, and communication protocols used by the ML-enabled and connected peripheral devices, to assess the potential security vulnerabilities associated with each of them.

To identify the ML and peripheral device technologies, we have integrated two questionnaires [55] into our toolkit [54], which must be completed by the designers of the ML-enabled device. The first focuses on *compatibility conditions* for each peripheral device in the control structure and needs to be answered by the manufacturer of the ML-enabled device. For example, some blood glucose management systems use Bluetooth to transmit glucose readings from the glucose meter to the glucose management smartphone app, while others require a USB connection for data transfer. Following this, the security analyst must manually identify all commercial peripheral devices that meet the compatibility conditions specified by the ML-enabled device manufacturer. The second questionnaire helps analysts identify the *technologies used in the ML-enabled device, and each compatible peripheral device* by covering key technological and operational factors relevant to medical devices. These questions ensure a thorough assessment of potential attack entry points. The factors are categorized into four groups [54]: (i) Human Interaction – this includes data entry and supervision, data validation, authentication, and anomaly detection; (ii) Communication Protocol – this includes the exact protocol name, version, and whether it uses encryption; (iii) Electromagnetic Susceptibility – this includes whether the device is susceptible to electromagnetic radiation, and if so, what its repercussions would be, and if they have any known shielding or mitigation strategy in place; and, (iv) Dependencies on firmware, hardware, OS, and external libraries. This categorization is based on known attack vectors targeting ML-enabled medical devices [129].

Note that, for a security analyst working for a medical device manufacturing company, obtaining the aforementioned information from the device designers would be straightforward. However, for an analyst working independently or for a third-party company, the manufacturers might be unwilling to provide this information either due to reluctance to spend unnecessary time or effort (as might be the case for peripheral device technologies) or due to confidentiality concerns (as might be the case for ML technique details). In such cases, the analyst can infer compatibility conditions, such as communication links, input devices, and operating systems, from publicly available device descriptions on the FDA website [110] and publicly available information on each peripheral, such as product descriptions on the manufacturer's website. Following this, the analyst could also retrieve a fairly comprehensive list of compatible peripheral devices from third-party information repositories such as TidePool [106]. However, identifying the specific ML technologies used is far more challenging, as most manufacturers do not declare them on their website or do not disclose them publicly at all. To assist the analyst under such circumstances, we have developed NLP- and LLM-aided tools as described in the following subsection.

3.2 NLP and LLM-aided Web Scraping and Information Extraction

Once the security analyst builds the control structure and identifies the technologies used in the ML-enabled device and its connected components, they must proceed to identify known vulnerabilities in these technologies that might serve as an attack entry point. Today, information about security vulnerabilities, design flaws, and adverse events reported on medical devices is available on the Internet in an unstructured and dispersed manner. This makes it challenging to ensure the coverage of all relevant data during the security assessment process. Our set of NLP-aided web scraping and LLM-aided information extraction tools and techniques assists the system developers and security analysts in extracting and integrating data on all known vulnerabilities and safety issues relevant to the ML-enabled medical device under assessment. This information is also used by our subsequent tools for automated systems-theoretic safety and security analysis.

ML Technology and Vulnerability Identification. In our latest work [32], we proposed MedAIScout, a semi-automated NLP- and LLM-aided tool designed to retrieve information on known ML vulnerabilities relevant to ML-enabled medical devices. MedAIScout works in two steps:
(1) *ML technology identification:* Given a description of an ML-enabled medical device, MedAIScout uses NLP techniques to identify key terms related to the device's functionality, ML model type, and data characteristics. Often, the device manufacturers do not publicly disclose the exact ML technique used in their products. In case the security analyst (MedAIScout user) does not have access to a document containing the exact details (such as in the case of third-party analysts), MedAIScout can analyze available information and infer the most likely ML technique by referencing similar devices documented in existing literature. In this work, we sourced the device descriptions from the publicly accessible pre-market device summaries available on the FDA website [110,118] and peer-reviewed research articles indexed on Google Scholar.
(2) *ML vulnerability identification:* Next, MedAIScout constructs tailored search queries to retrieve peer-reviewed research articles on attacks targeting the device's ML model. MedAIScout uses local LLMs to differentiate between training-time and inference-time attacks and provides context and explanations for each retrieved article's relevance.

Throughout the device's lifecycle, security analysts can use MedAIScout to track emerging ML vulnerabilities. To the best of our knowledge, it is the first automated tool to retrieve known ML vulnerabilities specifically for medical applications. By applying MedAIScout to five FDA-approved ML-enabled medical devices, we found that MedAIScout successfully uncovered relevant vulnerabilities in four devices, thereby substantially assisting in security analysis. For example, when tested on the One Drop blood glucose monitoring system [116], MedAIScout retrieved a peer-reviewed research paper [107] describing an inference-time attack on a similar system. In this attack, an adversary manipulates blood glucose readings at mealtime by compromising the radio communication between the glucose meter and the controller, leading to incorrect insulin dose recommendations. The paper also proposes an appropriate attack detection technique.

Attack Surface Analysis. To capture a comprehensive attack surface for ML-enabled medical devices, we have developed tools for automated searching of public databases and identifying known vulnerabilities in the *peripheral* and *interconnected medical devices*. Comprehensive attack surface analysis is a prerequisite for systems-theoretic security analysis

In our recent work [54], we developed a method for capturing all the known vulnerabilities linked to each technology in every peripheral device used in a given ML-enabled medical device. This method uses the responses about the technological and operational factors used in the peripheral devices from the questionnaires (see §3.1) as search keywords to find known vulnerabilities in the MITRE Common Vulnerability Enumeration (CVE) database [87]. For instance, for the d-Nav BGMS [112], we found that vulnerabilities might exist in a compatible glucose meter [51], Wi-Fi communication between glucose meter and the cloud server [28], the communication between the cloud server for glucose meter and the ML controller [96], Wi-Fi communication between the interface device and the ML controller [28], and Android OS on the interface [31].

In another study [128], we examined cyberattacks targeting hospital networks and interconnected clinical environments. For this purpose, we used two publicly available vulnerability databases – the Common Vulnerabilities and Exposures (CVE) Database [87] and the Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) Alerts database [108]. To automate the collection of information on medical device-related vulnerabilities from ICS-CERT, we developed a tool for crawling the whole US-CERT website and extracting all vulnerability records that contain any medical-related keywords, including generic medical keywords and those describing the common categories and specialties of medical devices, as classified by the FDA Product Code Classification Database [121]. Using this tool, we extracted the vulnerability records reported from 1999 to 2018 that were potentially related to medical devices. We then manually parsed the HTML documents of a final set of 140 extracted records to extract information such as the corresponding CVE IDs, affected product names, and manufacturer or vendor names of products, as well as vulnerability details and backgrounds. Our analysis revealed that the most common vulnerabilities included improper credential management and authentication, weak access control, privilege escalation, and buffer and stack overflows. Furthermore, we found that 18 retrieved vulnerabilities had publicly available exploits. These vulnerabilities were widespread across various medical devices, including insulin pumps, from multiple manufacturers, thereby underscoring the need to consider them in the security analysis of interconnected ML-enabled devices.

Analysis of Recalls and Adverse Events in Medical Devices. In our early work [9,8,11], we developed a suite of NLP tools (called MedSafe [13,12]) for automated extraction, cross-referencing, and classification of records from two public FDA databases: the Medical Device Recalls [122] and the MAUDE (Adverse Event Reports) database [119]. We used these tools to identify all the recalls and adverse events caused by failures in computer-based medical devices, and categorized them by fault class, failure mode, device type, recovery action, and the number of recalled devices. This study was the first automated and large-scale analysis of FDA data on computer-based medical devices and highlighted the key causes of computer failures impacting patient safety. Our findings showed that while software failures continue to be the leading cause of medical device failures, hardware, battery, and I/O issues are also major contributors. Many recalled devices either lacked proper safety considerations during design or their safety mechanisms were inadequately implemented. Later, using these tools we extracted all the recalls and adverse events related to BGMS [135,134] and

Table 4: Examples of recalls of AI/ML-enabled medical devices due to software issues that could result in misdiagnosis or wrong treatment (Class II recalls). * indicates that the recalls are in *Open* state as of April 2025, meaning that not all the units have been corrected or removed yet.

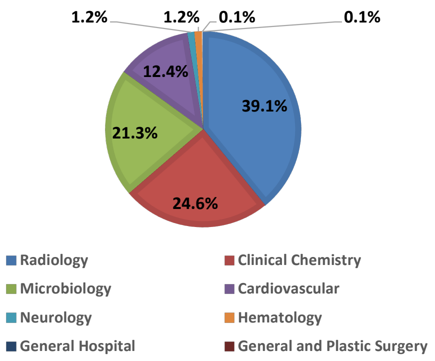
Device Name	**Approval Panel**	**Recall #**	**Reason for Recall**	**Action Summary**	**No. of units affected**
BodyGuardian Heart Remote Monitoring Kit	Cardiovascular	Z-2479-2020 [42]	The device data being collected and transferred to the monitoring center may not be accurate due to non-validated association between the phone software and the heart monitors, leading to inaccurate evaluation of the patients' condition.	The recalling firm contacted all patients and physicians that had potentially impacted devices. Patients that agreed were sent new devices to replace the affected one to finish their study.	8
Dario BGMS	Clinical chemistry	Z-0260-2020 [40]	The Dario Android App v4.3.0-4.3.2 may experience duplicate logging of a blood glucose level reading.	The firm released Android App v4.3.3. Users were informed about the issue via multiple push notifications and email, asking them to update to the new version.	126,271
Bioplex 2200 ANA Screen	Clinical toxicology	Z-1159-2008 [36]	False negative results due to reagent packs exhibiting low signal.	The firm contacted its consignees, informing them of the issue, recommended that they perform QC testing daily with each reagent pack, and updated the usage instructions.	8,804
Sight OLO CBC Test Kit	Hematology	Z-2173-2024* [46]	The kit shows a bias in the platelet count due to bacterial contamination, which can result in elevated counts with a bias, that results in the test kit performing outside of the device specification.	The manufacturer issued an urgent recall notice to their customers, asking them to discontinue the use of the affected test kits, return the unused kits, and dispose of the used ones.	7,450
UniCel DxH 600 Coulter Cellular Analysis System	Microbiology	Z-2158-2017 [37]	A possible data acquisition disruption may cause some unusual events, that may be incorrectly removed from analysis, which can result in erroneous diagnosis.	The manufacturers sent an Urgent Medical Device Recall letter to customers to inform them of the issue, impact, action, and resolution.	1,408
Incisive CT,(728143, 728144), Software v5.0.0.	Radiology	Z-0640-2024* [45]	Multiple software issues have the potential to lead to misdiagnosis due to image artifacts or incorrect image orientation labels, or need for a CT rescan.	The manufacturer communicated specific details regarding the issue to their customer, as well as advice on actions to be taken. They also promised to install a software upgrade.	828

Table 5: Examples of adverse events of AI/ML-enabled medical devices with data, interface device, and software related problems that could result in misdiagnosis or wrong treatment (Event Types: Malfunction).
* indicates several similar adverse events reported for the same device over 2018-2024.

Device Name	**Device Function**	**Approval Panel**	**Adverse Event # (Year)**	**Device Problem**	**Summary Event Description**
Zio AT ECG Monitoring System (ZEUS)	Arrhythmia detector and alarm	Cardio-vascular	8356453 [41] (2019)	Application Network Problem	False negative results (missed detection of asymptomatic arrhythmia) due to a BLE (bluethooth low energy) issue.
LINQ II Cardiac Monitor, Zelda AI ECG Classification System	Arrhythmia detector and alarm	Cardio-vascular	20916084 [48] (2024)	Program or Algorithm Execution Problem	An atrial fibrillation episode was adjudicated as false by the artificial intelligence (ai) algorithm.
Dario BGMS	Glucose Test System	Clinical chemistry	18904273 [44] (2022)*	Incorrect, Inadequate, or Imprecise Result or High Readings	Inconsistent and high blood glucose readings, different from other meters or hospital measurements.
HeartFlow FFRCT	Coronary Vascular Physiologic Simulation Software	Cardio-vascular	8269286 [38] (2018)*	False Negative Result	Potential false negative results in FFRCT (Fractional Flow Reserve derived from CT) analysis of coronary arteries due to image quality issues and anatomy uncertainty.
Clarius Ultrasound Scanner	Ultrasonic pulsed doppler imaging system	Radiology	20471171 [47] (2024)	Misconnection	A connectivity issue with ultrasound scanner during diagnostic evaluation in an emergency room, which could potentially lead to significant adverse outcomes.

surgical robots [11] and identified most common device malfunctions, examples of security vulnerabilities in different components and device interfaces (e.g., CGMs, insulin pumps, cameras), and the safety impact of device failures and potential harm to patients (e.g., hyperglycemia or injury). For example, we found a Class 1 recall (highest risk level) due to a potential security vulnerability related to the use of the remote controller accessories with the insulin pumps, which affected over 90,000 users on the market [39]. Another Class 2 recall, affecting over 64,000 insulin pumps, indicated the possibility of an unauthorized person connecting wirelessly to a nearby insulin pump to change settings and control insulin delivery due to potential cybersecurity vulnerabilities [43].

More recently, we have applied our techniques to extract and analyze the recalls and adverse events reported on ML-enabled devices and AID systems. Our analysis found over 1,460 adverse events reported for ML-enabled devices over 2015-2024, of which about 92% involved device malfunctions and 7.8% injuries. Although understanding the root causes of the reported events requires an in-depth investigation and consideration of all causal and contextual factors 3.3, these reports provide valuable insights on real problems encountered during the use of devices and how they impacted patient safety. Figure 5 shows the device



Reported Device Problem	Freq.
Poor Quality Image	421
Not Available (N/A)	295
False Negative Result	103
High Readings	73
Incorrect, Inadequate, or Imprecise Result or Readings	45
Computer Software Problem	42
Inaccurate Information	25
Application Program Problem: Parameter Calculation Error	17
Low Readings	7
Intermittent Program or Algorithm Execution	6
Program or Algorithm Execution Failure	4
Failure to Transmit Record	3
Low Test Results	3

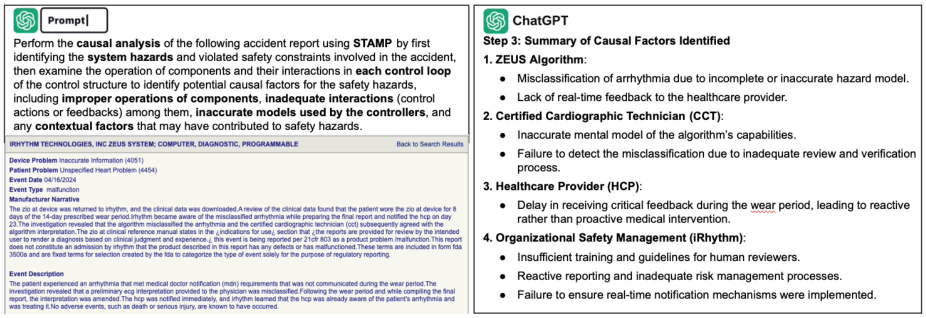
Fig. 5: Left: Adverse Events by Device Category (FDA Approval Panel), Right: Top Reported ML-enabled Device Problems (Data as of April 2025)

categories with the highest number of adverse event reports and the top device problems reported over the years. A major part of reported problems (about 60.7%) were related to poor quality and inaccurate inputs/readings and false negative results, which, even if not directly caused by an ML technology, could still impact the ML decision-making results and patient safety. Some examples of safety-critical recalls and adverse events across different device categories are shown in Tables 4 and 5.

In summary, this set of tools and techniques would help a security analyst cover known vulnerabilities in ML models, peripheral medical devices, as well as attack vectors in connected peripheral devices and communication channels, while designing a *secure* ML-enabled medical device. Additionally, it will also help security practitioners design efficient attack prevention and detection techniques.

3.3 Data-Driven Systems-Theoretic Safety and Security Analysis

To predict and proactively mitigate the occurrence of future attacks, it is crucial to not only consider the known vulnerabilities and exploits reported in existing data on past safety and security incidents, but also anticipate for the potential new attacks by considering a more comprehensive attack surface of unknown vulnerabilities or vulnerabilities in other connected devices and the potential attack steps and their safety impacts on patients. To do this, we adopt an LLM-aided and data-driven approach to the systems-theoretic security analysis (STPA-Sec) that incorporates the information extracted from public databases and results from CAST analysis on devices with similar functionality (e.g., a non-ML-enabled device predicate with the same functional specification and use cases as the ML-enabled device) to generate potential attacks steps and their impacts in ML-enabled devices.



Perform the **causal analysis** of the following accident report using **STAMP** by first identifying the **system hazards** and violated safety constraints involved in the accident, then examine the operation of components and their interactions in **each control loop** of the control structure to identify potential causal factors for the safety hazards, including **improper operations of components**, **inadequate interactions** (control actions or feedbacks) among them, **inaccurate models used by the controllers**, and any **contextual factors** that may have contributed to safety hazards.

IRHYTHM TECHNOLOGIES, INC ZEUS SYSTEM; COMPUTER, DIAGNOSTIC, PROGRAMMABLE

Device Problem Inaccurate Information (4051)
Patient Problem Unspecified Heart Problem (4454)
Event Date 04/16/2024
Event Type malfunction
Manufacturer Narrative

The zio at device was returned to irhythm, and the clinical data was downloaded.A review of the clinical data found that the patient wore the zio at device for 8 days of the 14-day prescribed wear period.Irhythm became aware of the misclassified arrhythmia while preparing the final report and notified the hcp on day 23.The investigation revealed that the algorithm misclassified the arrhythmia and the certified cardiographic technician (cct) subsequently agreed with the algorithm interpretation.The zio at clinical reference manual states in the ¿indications for use¿ section that ¿the reports are provided for review by the intended user to render a diagnosis based on clinical judgment and experience.¿ this event is being reported per 21cfr 803 as a product problem /malfunction.This report does not constitute an admission by irhythm that the product described in this report has any defects or has malfunctioned.These terms are included in form fda 3500a and are fixed terms for selection created by the fda to categorize the type of event solely for the purpose of regulatory reporting.

Event Description

The patient experienced an arrhythmia that met medical doctor notification (mdn) requirements that was not communicated during the wear period.The investigation revealed that a preliminary ecg interpretation provided to the physician was misclassified.Following the wear period and while compiling the final report, the interpretation was amended.The hcp was notified immediately, and irhythm learned that the hcp was already aware of the patient's arrhythmia and was treating it.No adverse events, such as death or serious injury, are known to have occurred.

Step 3: Summary of Causal Factors Identified

1. ZEUS Algorithm:
- Misclassification of arrhythmia due to incomplete or inaccurate hazard model.
- Lack of real-time feedback to the healthcare provider.

2. Certified Cardiographic Technician (CCT):
- Inaccurate mental model of the algorithm's capabilities.
- Failure to detect the misclassification due to inadequate review and verification process.

3. Healthcare Provider (HCP):
- Delay in receiving critical feedback during the wear period, leading to reactive rather than proactive medical intervention.

4. Organizational Safety Management (iRhythm):
- Insufficient training and guidelines for human reviewers.
- Reactive reporting and inadequate risk management processes.
- Failure to ensure real-time notification mechanisms were implemented.

Fig. 6: Example Accident Causality Analysis using STAMP for an FDA-Authorized ML-enabled Medical Device

Systems-Theoretic Accident Causality (CAST) Analysis. Analysis of real-world safety incidents, including medical device recalls [122] and adverse events [119] can provide valuable insights into how device flaws and security vulnerabilities could lead to system hazards and negatively impact patients and caregivers. However, these incidents are mainly reported by the device users and manufacturers in free-form natural language text, and their analysis requires a semantic understanding of the underlying causal factors. Several previous studies [17,10,93,79,21,70] have shown the advantage of Causal Analysis using System Theory (CAST) [69] in identifying causal and contextual factors contributing to medical adverse events. However, these papers solely focus on the manual causal analysis of single incidents and do not consider security-related hazards and safety-critical vulnerabilities. Such an approach cannot provide a comprehensive understanding of all potential causal factors, including vulnerabilities in IoT and peripheral devices, nor can it yield statistically significant measures of their importance. Additionally, it is not easily scalable to thousands of unstructured adverse event reports on a single device due to the significant human effort required. Therefore, techniques and tools for automated semantic analysis of these reports are needed to extract both safety and security-related causal factors, and summarize key information for CAST analysis.

To facilitate an aggregated CAST analysis, we leverage our NLP techniques for automated classification, summarization, and cross-referencing of large-scale FDA data on recalls and adverse events [9,8,11,135,33]. This information can assist in systems-theoretic analysis of several similar adverse events reported on the same medical device or devices with the same functional specification using CAST to identify the distribution of causal factors and potentially inadequate safety mechanisms in both system design and operational practices [5,6]. Given a natural language description of an adverse event and the control structure model for a medical device, we first map different sections of the text into different control loops in the system control structure. Then, for each control loop, the set of violated safety constraints is identified. These steps are done through device and medical entity and relation extraction from the text and semantic analysis of causal factors using rule-based parts of speech analysis [5]. Finally, the similar causal factors and safety violations across multiple adverse event reports of the same device can be identified and aggregated for statistical analysis. In [11,6] we performed such an aggregated causality analysis of over 10,000 adverse event reports on tele-operated surgical robots. This analysis identified the most critical causal factors for safety incidents in different models of the same device over a period of 14 years. To further reduce the manual cost of this analysis method, we have recently explored an LLM-aided technique based on customized prompt templates and chain-of-thought prompting [52,127] to decompose the tasks of entity and relation extraction and semantic analysis of causal factors into sub-tasks that can be performed using LLMs and be later manually validated by security analysts. Figure 6 shows an example of the key causal factors extracted by this LLM-aided technique from an adverse event report [120] for an FDA-authorized ML-enabled cardiac event detection software [123]. The insights on the causes and patient impacts of past incidents can be used for analyzing and specifying the safety impact of the device vulnerabilities.

Systems-Theoretic Security (STPA-Sec) Analysis. In this final step, we analyze consolidated data on the ML model, its functionality, peripheral technologies, and associated safety and security risks to identify how an adversary could inject false data during inference. We developed STPA-Sec for ML-enabled Medical Devices (SAM), a technique for conducting STPA-Sec on AI/ML-enabled medical devices [54]. SAM first assesses the attack surface by identifying all potential attack entry points (Section 3.2). Thereafter, it performs STPA-Sec analysis to determine the attack steps. This information would help the ML-enabled device manufacturer design appropriate security measures or devise advisories for the users.

In the attack step generation step, SAM performs an LLM-aided STPA-Sec analysis to generate the attack steps for a given hazard and an exploitable peripheral device vulnerability. To overcome a human security analyst's limited cross-domain knowledge, we leverage LLMs to automatically identify causal scenarios based on the latest vulnerabilities in the system's underlying technologies. A key challenge in using LLMs is the design of effective prompts to generate optimal task-specific responses. For SAM, the ideal response outlines detailed attack steps exploiting a peripheral vulnerability to inject false data during inference on a given ML technique. To achieve this, we developed the following prompt.

> "Act as a security engineer who has the task of identifying the steps that an adversary follows to cause a security breach in an ML-enabled medical system. <*Description of an ML-enabled medical system*>. <*Definition of security breach*>. You are given a system description, an ML attack, a targeted input peripheral component, and a known vulnerability in the input component. Give a list of steps to show how an adversary can

exploit the vulnerability to mislead the ML-enabled component and how that affects the action of the output device on the patient.
System Description: <*The SAM user manually writes this description by inspecting information disclosed by the manufacturer.*>
Data flow: <*This can be derived from the control structure constructed using the Control structure builder in §3.1.*>
ML attack: <*The ML attack identified in §3.2*>
Targeted input peripheral component: <*One of the peripheral input devices in the control structure built in §3.1*>
Targeted technology: <*One of the underlying technologies in the input device, as identified by the technology identifier (§3.1)*>
Known vulnerability: <*Description of the known vulnerability in the targeted technology, as retrieved from the CVE database during attack surface analysis*>"

We observed that explicitly assigning the LLM the role of a security analyst before giving it additional information improves the readability and relevance of the generated results - this is in line with other work in this area [102,73,101]. Similarly, mentioning the data flow provides clarity to the LLM regarding the sequence of data transmission between different components in the system.

By running this prompt for each vulnerable point in the system and each vulnerability uncovered at that point, SAM, regardless of the existence of safety/security margins, generates a comprehensive set of steps an adversary might take to compromise the security of an ML-enabled medical device. Device manufacturers or security analysts can then disregard those that have already been mitigated and develop design recommendations for the remaining ones.

For d-Nav, we selected hypoglycemia as the hazard, and *injecting excess insulin* as the control action that causes it. We consider an adversary who conducts a model inversion attack (identified by MedAIScout [32] in a previous step described in §3.2) on the ML engine to infer sensitive details about a targeted patient, followed by false data injection. This attack would make the ML engine mispredict the insulin dose. To execute this attack, the adversary injects false glucose readings into the Wi-Fi channel that transmits the patient's glucose readings from the glucose meter to the ML engine running in the cloud server. We assume that the patient uses a Wi-Fi router with an unpatched known vulnerability, *CVE-2023-35836* [30], that the adversary exploits for injecting the malicious glucose readings. SAM outputs a list of nine steps for this attack, which are summarized in Table 6. By following these steps, an adversary could inject false data into the BGMS to make it miscalculate the insulin dose.

4 Future Directions and Conclusion

Based on the capabilities of our tools and techniques demonstrated in this paper, and the insights obtained from the experimental results, we would like to expand

Table 6: STPA-Sec output produced by SAM for the attack scenario on d-Nav BGMS, described in $3.3

Step #	**Step name**	**Description**
1	Reconnaissance	Identifying the targeted patient's Wi-fi network and its router vulnerabilities
2	Exploitation	Exploitation of router vulnerability to infiltrate the target's network
3	Wi-fi network infiltration	Compromising the connection between glucose meter and cloud server
4	Data interception	Interpreting the data in transit
5	Data tampering	Manipulating the data in transit with a value that would make the ML model mispredict a future blood glucose value
6	Model inversion attack	Compute the manipulated value such that the patient becomes hypoglycemic
7	ML Controller manipulation	Expected reaction of the ML model: Misprediction of patient's future blood glucose level
8	Output device manipulation	Expected reaction of the insulin dose calculator: Computing an insulin dose higher than that required by the patient and sent to the insulin pump or displayed on the d-Nav app
9	Insulin pump misadministration	Expected end result: Wrong insulin dose administered to the patient, either manually or by an automated insulin pump

the scope of our research in the following directions, with a high-level goal of ensuring the security of ML-enabled medical devices by design and efficient post-market security surveillance.

1. Early prediction of vulnerabilities based on existing events - The domain of ML-enabled medical devices has become increasingly competitive, with manufacturers developing ML-enabled devices that offer similar core functionalities as existing non-ML-enabled devices, but with enhanced performance and features such as greater interoperability. As a result, newer devices may inherit existing vulnerabilities in connected devices or similar or more severe vulnerabilities than their predecessor devices. To address this, we plan to develop an LLM-aided technique that analyzes the design of an ML-enabled medical device and, based on known vulnerability data, predicts potential security risks specific to the new device, even without performing STPA-Sec on it. Furthermore, we would expand the scope of our tools and techniques to cover other types of ML-specific attacks in addition to false data injection attacks.
2. Real-time post-market security risk assessment - Our tools and techniques can be extended to support near real-time post-market security surveillance by

continuously monitoring large-scale vulnerability databases and performing on-demand risk assessments whenever new vulnerabilities are reported.
3. Designing efficient defense techniques - The output of our STPA-Sec technique can be leveraged to identify the most efficient defense technique in terms of reliability, patient convenience, and cost of implementation.

This paper presents a suite of tools and techniques developed for holistic security risk assessment of ML-enabled medical devices, with a focus on false data injection attacks. We demonstrated the effectiveness of these tools and techniques across multiple ML-enabled blood glucose management systems. The novelty of our tools and techniques are in (i) identifying attack vectors that require exploiting vulnerabilities in third-party connected components to practically execute known attacks on the ML models and (ii) anticipating for the potential safety impacts of such attacks based on the analysis of past incidents on similar devices. This helps security analysts (working for the device manufacturers) assess the feasibility and impact of such attacks more accurately. In the future, we aim to extend our tools to support additional types of ML-specific attacks and facilitate post-market security risk assessments.

Acknowledgments. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the National Research Council of Canada's (NRC) Digital Health and Geospatial Analytics Program, and the U.S. National Science Foundation (CNS-2146295).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 2023 State of Cybersecurity for Medical Devices and Healthcare Systems. Tech. rep., Health-ISAC (August 2023), "https://info.finitestate.io/2023-state-of-cybersecurity-for-medical-devices-and-healthcare-systems"
2. Abdulkhaleq, A., Wagner, S.: Open tool support for system-theoretic process analysis. Universitätsbibliothek der Universität Stuttgart (2014)
3. Abdulkhaleq, A., Wagner, S.: XSTAMPP: an eXtensible STAMP platform as tool support for safety engineering (2015)
4. Albattah, A., Rassam, M.A.: Detection of Adversarial Attacks against the Hybrid Convolutional Long Short-Term Memory Deep Learning Technique for Healthcare Monitoring Applications. Applied Sciences **13**(11), 6807 (2023)
5. Alemzadeh, H.: Data-driven resiliency assessment of medical cyber-physical systems. Ph.D. thesis, University of Illinois at Urbana-Champaign (2016)
6. Alemzadeh, H., Chen, D., Lewis, A., Kalbarczyk, Z., Raman, J., Leveson, N., Iyer, R.: Systems-theoretic safety assessment of robotic telesurgical systems. In: Koornneef, F., van Gulijk, C. (eds.) Computer Safety, Reliability, and Security. pp. 213–227. Springer International Publishing, Cham (2015)
7. Alemzadeh, H., Chen, D., Li, X., Kesavadas, T., Kalbarczyk, Z.T., Iyer, R.K.: Targeted attacks on teleoperated surgical robots: Dynamic model-based detection and mitigation. In: IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). pp. 395–406. IEEE (2016)

8. Alemzadeh, H., Hoagland, R., Kalbarczyk, Z., Iyer, R.K.: Automated classification of computer-based medical device recalls: An application of natural language processing and statistical learning. In: IEEE International Symposium on Computer-Based Medical Systems. pp. 553–554. IEEE (2014)
9. Alemzadeh, H., Iyer, R.K., Kalbarczyk, Z., Raman, J.: Analysis of safety-critical computer failures in medical devices. IEEE Security & Privacy **11**(4), 14–26 (2013). https://doi.org/10.1109/MSP.2013.49
10. Alemzadeh, H., Raman, J., Leveson, N., Iyer, R.K.: Safety implications of robotic surgery: A study of 13 years of FDA data on da Vinci surgical systems. Coordinated Science Laboratory Report no. UILU-ENG-13-2208 (2013)
11. Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., Iyer, R.K.: Adverse events in robotic surgery: a retrospective study of 14 years of FDA data. PloS one **11**(4), e0151470 (2016)
12. Alemzadeh, Homa: MedSafe MAUDE, URL: https://github.com/homa-alem/MedSafe_MAUDE, Last accessed: Apr 18, 2025
13. Alemzadeh, Homa: MedSafe Recalls, URL: https://github.com/homa-alem/MedSafe_Backend/, Last accessed: Apr 18, 2025
14. Almohri, H., Cheng, L., Yao, D., Alemzadeh, H.: On threat modeling and mitigation of medical cyber-physical systems. In: IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 114–119. IEEE (2017)
15. Amich, A., Eshete, B.: Explanation-guided diagnosis of machine learning evasion attacks. In: Proceedings of Security and Privacy in Communication Networks: EAI International Conference, Part I 17. pp. 207–228. Springer (2021)
16. Arney, D., Pajic, M., Goldman, J.M., Lee, I., Mangharam, R., Sokolsky, O.: Toward patient safety in closed-loop medical device systems. In: Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems. pp. 139–148 (2010)
17. Balgos, V.H.: A systems theoretic application to design for the safety of medical diagnostic devices. Ph.D. thesis, Massachusetts Institute of Technology (2012)
18. Becker, C., Van Eikema Hommes, Q.: Transportation systems safety hazard analysis tool (SafetyHAT) user guide (version 1.0). Tech. rep. (2014)
19. Bonaci, T., Yan, J., Herron, J., Kohno, T., Chizeck, H.J.: Experimental analysis of denial-of-service attacks on teleoperated robotic systems. In: Proceedings of the ACM/IEEE International Conference on Cyber-physical Systems. pp. 11–20 (2015)
20. Bortsova, G., González-Gonzalo, C., Wetstein, S.C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J.P., Veta, M., et al.: Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Medical Image Analysis **73**, 102141 (2021)
21. Canham, A.: Examining the application of STAMP in the analysis of patient safety incidents. Ph.D. thesis, Loughborough University (2018)
22. Chen, H., Huang, C., Huang, Q., Zhang, Q., Wang, W.: Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 3446–3453 (2020)
23. Chen, X., Meng, L., Xu, Y., Wu, D.: Adversarial artifact detection in EEG-based brain–computer interfaces. Journal of Neural Engineering **21**(5), 056043 (2024)
24. Chen, Y., Yan, J., Jiang, M., Zhang, T., Zhao, Z., Zhao, W., Zheng, J., Yao, D., Zhang, R., Kendrick, K.M., et al.: Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification. IEEE Transactions on Neural Networks and Learning Systems (2022)

25. Chhabra, A., Roy, A., Mohapatra, P.: Suspicion-free adversarial attacks on clustering algorithms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3625–3632 (2020)
26. CVE: CVE-2017-14008 (2017), URL: https://www.cve.org/CVERecord?id=CVE-2017-14008, Last accessed: Apr 18, 2025
27. CVE: CVE-2019-10964 (2019), URL: https://www.cve.org/CVERecord?id=CVE-2019-10964, Last accessed: Apr 18, 2025
28. CVE: CVE-2020-26145 (2020), URL: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE-2020-26145, Last accessed: Apr 18, 2025
29. CVE: CVE-2020-8933 (2020), URL: https://www.cve.org/CVERecord?id=CVE-2020-8933, Last accessed: Apr 18, 2025
30. CVE: CVE-2023-3583 (2024), URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2023-35836, Last accessed: Apr 18, 2025
31. CVE: CVE-2024-43093 (2024), URL: https://www.cve.org/CVERecord?id=CVE-2024-43093, Last accessed: Apr 18, 2025
32. Dharmalingam, A.P., Mitra, G.: MedAIScout: Automated Retrieval of Known Machine Learning Vulnerabilities in Medical Applications. In: Red Teaming GenAI: What Can We Learn from Adversaries? (2024)
33. Elnawawy, M., Hallajiyan, M., Mitra, G., Iqbal, S., Pattabiraman, K.: Systematically Assessing the Security Risks of AI/ML-enabled Connected Healthcare Systems. In: IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 97–108 (2024). https://doi.org/10.1109/CHASE60773.2024.00019
34. Elshazly, A.A., Elgarhy, I., Eltoukhy, A.T., Mahmoud, M., Eberle, W., Alsabaan, M., Alshawi, T.: False data injection attacks on reinforcement learning-based charging coordination in smart grids and a countermeasure. Applied Sciences (2076-3417) **14**(23) (2024)
35. FDA: Product Classification: Interoperable Automated Glycemic Controller, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpcd/classification.cfm?id=714, Last accessed: Apr 18, 2025
36. FDA: Class 2 Device Recall BioPlex 2200 ANA Screen on the BioPlex 2200 MultiAnalyte Detection System (2008), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=66385, Last accessed: Apr 18, 2025
37. FDA: Class 2 Device Recall UniCel DxH 600/800 Coulter Cellular Analysis System (2017), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=155030, Last accessed: Apr 18, 2025
38. FDA: MAUDE Adverse Event Report: HEARTFLOW, INC. FFRCT) (2018), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=8269286, Last accessed: Apr 18, 2025
39. FDA: Class 1 Device Recall MMT500 (2019), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm?id=170857, Last accessed: Apr 18, 2025
40. FDA: Class 2 Device Recall Dario Blood Glucose Monitoring System (2019), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=176305, Last accessed: Apr 18, 2025
41. FDA: MAUDE Adverse Event Report: IRHYTHM TECHNOLOGIES, INC ZIO AT SYSTEM (2019), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=8356453, Last accessed: Apr 18, 2025
42. FDA: Class 2 Device Recall BodyGuardian Heart Remote Monitoring Kit (2020), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfres/res.cfm?id=180336, Last accessed: Apr 18, 2025

43. FDA: Class 2 Device Recall MiniMed Insulin Pump (2020), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=175194, Last accessed: Apr 18, 2025
44. FDA: MAUDE Adverse Event Report: LABSTYLE INNOVATIONS LTD. DARIO BLOOD GLUCOSE MONITORING SYSTEM; GLUCOMETER) (2022), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=18904273, Last accessed: Apr 18, 2025
45. FDA: Class 2 Device Recall Incisive CT (2024), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=204881, Last accessed: Apr 18, 2025
46. FDA: Class 2 Device Recall Sight OLO (2024), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm?id=207976, Last accessed: Apr 18, 2025
47. FDA: MAUDE Adverse Event Report: CLARIUS MOBILE HEALTH CORP. CLARIUS ULTRASOUND SCANNER; DIAGNOSTIC ULTRASOUND SYSTEM AND ACCESSORIES (2024), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=20471171, Last accessed: Apr 18, 2025
48. FDA: MAUDE Adverse Event Report: MEDTRONIC, INC. ACCURHYTHM ZA410 (AF); RECORDER, EVENT, IMPLANTABLE CARDIAC, (WITH ARRHYTHMIA DETECTION) (2024), URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=20916084, Last accessed: Apr 18, 2025
49. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
50. Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296 (2018)
51. Garbelini, M.E., Wang, C., Chattopadhyay, S., Sumei, S., Kurniawan, E.: {SweynTooth}: Unleashing mayhem over bluetooth low energy. In: USENIX Annual Technical Conference (USENIX ATC). pp. 911–925 (2020)
52. Ge, X., Williams, R.D., Stankovic, J.A., Alemzadeh, H.: DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction. arXiv preprint arXiv:2310.07059 (2023)
53. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., et al.: Adversarial attacks and adversarial robustness in computational pathology. Nature communications **13**(1), 5711 (2022)
54. Hallajiyan, M., Dharmalingam, A.P., Mitra, G., Alemzadeh, H., Iqbal, S., Pattabiraman, K.: SAM: Foreseeing Inference-Time False Data Injection Attacks on ML-enabled Medical Devices. In: Workshop on Cybersecurity in HealthCare (HealthSec). pp. 77–84 (August 2024), co-held with ACM CCS'24
55. Hallajiyan, M., Dharmalingam, A.P., Mitra, G., Alemzadeh, H., Iqbal, S., Pattabiraman, K.: Sam questionnaires for collecting information on peripheral device technologies (2024)
56. Halperin, D., Heydt-Benjamin, T.S., Ransford, B., Clark, S.S., Defend, B., Morgan, W., Fu, K., Kohno, T., Maisel, W.H.: Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. In: 2008 IEEE Symposium on Security and Privacy (sp 2008). pp. 129–142. IEEE (2008)
57. Han, X., Hu, Y., Foschini, L., Chinitz, L., Jankelson, L., Ranganath, R.: Deep learning models for electrocardiograms are susceptible to adversarial attack. Nature medicine **26**(3), 360–363 (2020)
58. Hirano, H., Minagi, A., Takemoto, K.: Universal adversarial attacks on deep neural networks for medical image classification. BMC medical imaging **21**, 1–13 (2021)

59. Hu, L., Zhou, D.W., Guo, X.Y., Xu, W.H., Wei, L.M., Zhao, J.G.: Adversarial training for prostate cancer classification using magnetic resonance imaging. Quantitative Imaging in Medicine and Surgery **12**(6), 3276 (2022)
60. Insulet: Omnipod-5 (2025), URL: https://www.omnipod.com/what-is-omnipod/omnipod-5, Last accessed: Apr 18, 2025
61. Jee, E., Lee, I., Sokolsky, O.: Assurance cases in model-driven development of the pacemaker software. In: International Symposium On Leveraging Applications of Formal Methods, Verification and Validation. pp. 343–356. Springer (2010)
62. Jetley, R., Iyer, S.P., Jones, P.: A formal methods approach to medical device review. Computer **39**(4), 61–67 (2006)
63. Jiang, Z., Pajic, M., Connolly, A., Dixit, S., Mangharam, R.: Real-time heart model for implantable cardiac device validation and verification. In: 2010 22nd Euromicro Conference on Real-Time Systems. pp. 239–248 (2010). https://doi.org/10.1109/ECRTS.2010.36
64. Joel, M.Z., Umrao, S., Chang, E., Choi, R., Yang, D., Duncan, J., Omuro, A., Herbst, R., Krumholz, H.M., Aneja, S., et al.: Adversarial attack vulnerability of deep learning models for oncologic images. MedRxiv **1**, 2021 (2021)
65. Khan, M.A., Quasim, M.T., Alghamdi, N.S., Khan, M.Y.: A secure framework for authentication and encryption using improved ECC for IoT-based medical sensor data. IEEe Access **8**, 52018–52027 (2020)
66. Klonoff, D.C., Ho, C.N., Ayers, A., Abdel-Malek, A.: FDA Interoperability Designation—Creating Options for People With Diabetes and Pump Companies: Regulatory, Technological, and Commercial Perspectives. Journal of Diabetes Science and Technology (2024). https://doi.org/10.1177/19322968241271304, https://doi.org/10.1177/19322968241271304
67. Kramer, D.B., Baker, M., Ransford, B., Molina-Markham, A., Stewart, Q., Fu, K., Reynolds, M.R.: Security and privacy qualities of medical devices: An analysis of FDA postmarket surveillance. PloS one **7**(7), e40200 (2012)
68. Lal, S., Rehman, S.U., Shah, J.H., Meraj, T., Rauf, H.T., Damaševičius, R., Mohammed, M.A., Abdulkareem, K.H.: Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. Sensors **21**(11), 3922 (2021)
69. Leveson, N.: Engineering a safer world: Systems thinking applied to safety. MIT press (2011)
70. Leveson, N., Samost, A., Dekker, S., Finkelstein, S., Raman, J.: A systems approach to analyzing and preventing hospital adverse events. Journal of Patient Safety **16**(2), 162–167 (2020)
71. Levy-Loboda, T., Sheetrit, E., Liberty, I.F., Haim, A., Nissim, N.: Personalized insulin dose manipulation attack and its detection using interval-based temporal patterns and machine learning algorithms. Journal of Biomedical Informatics **132**, 104129 (2022)
72. Li, C., Raghunathan, A., Jha, N.K.: Hijacking an insulin pump: Security attacks and defenses for a diabetes therapy system. In: IEEE International Conference on e-Health Networking, Applications and Services. pp. 150–156 (2011)
73. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In: Advances in Neural Information Processing Systems. vol. 36, pp. 51991–52008. Curran Associates, Inc. (2023)
74. Li, Y., Liu, S.: Adversarial attack and defense in breast cancer deep learning systems. Bioengineering **10**(8), 973 (2023)

75. Lin, Y.C., Hong, Z.W., Liao, Y.H., Shih, M.L., Liu, M.Y., Sun, M.: Tactics of adversarial attack on deep reinforcement learning agents. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 3756–3762. International Joint Conferences on Artificial Intelligence Organization (2017)
76. Lyell, D., Wang, Y., Coiera, E., Magrabi, F.: More than algorithms: an analysis of safety events involving ML-enabled medical devices reported to the FDA. Journal of the American Medical Informatics Association **30**(7), 1227–1236 (2023)
77. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **110**, 107332 (2021)
78. Mangaokar, N., Pu, J., Bhattacharya, P., Reddy, C.K., Viswanath, B.: Jekyll: Attacking medical image diagnostics using deep generative models. In: IEEE EuroS&P. pp. 139–157 (2020)
79. Mason-Blakley, F., Habibi, R., Weber, J., Price, M.: Assessing stamp EMR with electronic medical record related incident reports: case study: manufacturer and user facility device experience database. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI). pp. 114–123. IEEE (2017)
80. Meiseles, A., Rosenberg, I., Motro, Y., Rokach, L., Moran-Gilad, J.: Adversarial Vulnerability of Deep Learning Models in Analyzing Next Generation Sequencing Data. In: IEEE BIBM. pp. 464–468 (2020). https://doi.org/10.1109/BIBM49941.2020.9313421
81. Menon, K., Bohra, V.K., Murugan, L., Jaganathan, K., Arumugam, C.: COVID-19 Diagnosis from Chest X-Ray Images Using Convolutional Neural Networks and Effects of Data Poisoning. In: ICCSA. pp. 508–521 (2021)
82. Mitre: Conexus Telemetry Protocol vulnerability, URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-6538, Last accessed: Apr 18, 2025
83. Mitre: Philips MRI 1.5T and MRI 3T vulnerability (1), URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-26262, Last accessed: Apr 18, 2025
84. Mitre: Shekar Endoscope vulnerability (1), URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2017-10722, Last accessed: Apr 18, 2025
85. Mitre: Sony IPELA E Series Camera vulnerability (1), URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-3938, Last accessed: Apr 18, 2025
86. Mitre: Windows 7 vulnerability (2), URL: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-5921, Last accessed: Apr 18, 2025
87. Mitre: Common Vulnerabilities and Exposures (CVE) Database (2024), URL: https://cve.mitre.org/, Last accessed: Apr 18, 2025
88. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., Jha, N.K.: Systematic poisoning attacks on and defenses for machine learning in healthcare. IEEE journal of biomedical and health informatics **19**(6), 1893–1905 (2014)
89. Nielsen, C., Tuladhar, A., Forkert, N.D.: Investigating the Vulnerability of Federated Learning-Based Diabetic Retinopathy Grade Classification to Gradient Inversion Attacks. In: International Workshop on Ophthalmic Medical Image Analysis. pp. 183–192 (2022)
90. Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H., Berger, C.: Engineering Safety Requirements for Autonomous Driving with Large Language Models. arXiv preprint arXiv:2403.16289 (2024)
91. Nouri, A., Cabrero-Daniel, B., Torner, F., Sivencrona, H., Berger, C.: Welcome Your New AI Teammate: On Safety Analysis by Leashing Large Language Models. In: Proceedings of the IEEE/ACM CAIN '24. p. 172–177 (2024)

92. Office, U.G.A.: Medical Device Cybersecurity: Agencies Need to Update Agreement to Ensure Effective Coordination. Tech. Rep. GAO-24-106683, United States Government Accountability Office (GAO) (December 2023), "https://www.gao.gov/assets/d24106683.pdf"
93. O'Neil, M.M.M.: Application of CAST to hospital adverse events. Ph.D. thesis, Massachusetts Institute of Technology (2014)
94. OpenAPS: URL: https://openaps.org, Last accessed: Apr 18, 2025
95. Pajic, M., Mangharam, R., Sokolsky, O., Arney, D., Goldman, J., Lee, I.: Model-driven safety analysis of closed-loop medical systems. IEEE Transactions on Industrial Informatics **10**(1), 3–16 (2012)
96. Palo Alto Networks: 6 New Vulnerabilities Found on D-Link Home Routers (2020), URL: https://unit42.paloaltonetworks.com/6-new-d-link-vulnerabilities-found-on-home-routers, Last accessed: Apr 18, 2025
97. Pattison, J., Dungan, K.M., Faulds, E.R.: Supporting the use of a person's own diabetes technology in the inpatient setting. Diabetes Spectrum **35**(4), 398–404 (2022)
98. Pereira, D.P., Hirata, C., Nadjm-Tehrani, S.: A STAMP-based ontology approach to support safety and security analyses. Journal of Information Security and Applications **47**, 302–319 (2019)
99. Qi, Y., Dong, Y., Khastgir, S., Jennings, P., Zhao, X., Huang, X.: STPA for Learning-Enabled Systems: A Survey and A New Practice. In: Proceedings of IEEE ITSC '23. pp. 1381–1388 (2023)
100. Qi, Y., Zhao, X., Khastgir, S., Huang, X.: Safety Analysis in the Era of Large Language Models: A Case Study of STPA Using ChatGPT. arXiv preprint arXiv:2304.01246 (2023)
101. Santu, S.K.K., Feng, D.: Teler: A general taxonomy of LLM prompts for benchmarking complex tasks. arXiv preprint arXiv:2305.11430 (2023)
102. Shanahan, M., McDonell, K., Reynolds, L.: Role play with large language models. Nature **623**(7987), 493–498 (2023)
103. Souza, F.G., Pereira, D.P., Pagliares, R.M., Nadjm-Tehrani, S., Hirata, C.M.: WebSTAMP: A web application for STPA & STPA-Sec. In: MATEC Web of Conferences. vol. 273, p. 02010. EDP Sciences (2019)
104. Sun, M., Tang, F., Yi, J., Wang, F., Zhou, J.: Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 793–801 (2018)
105. Thomas, J., Leveson, N.G.: Performing hazard analysis on complex, software-and human-intensive systems. In: Proc. of the 29th ISSC Conference about System Safety (2011)
106. Tidepool: Supported Devices, URL: https://www.tidepool.org/devices, Last accessed: Apr 18, 2025
107. Tosun, F.E., Teixeira, A., Ahlén, A., Dey, S.: Detection of bias injection attacks on the glucose sensor in the artificial pancreas under meal disturbance. In: American Control Conference (ACC). pp. 1398–1405. IEEE (2022)
108. US-CERT: Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) Alerts, URL: https://ics-cert.us-cert.gov/alerts, Last accessed: Apr 18, 2025
109. U.S. FDA: ABMD Software, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K213760, Last accessed: Apr 18, 2025

110. U.S. FDA: Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices, URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices, Last accessed: Apr 18, 2025
111. U.S. FDA: Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations, URL: https://www.fda.gov/media/184856/download, Last accessed: Apr 18, 2025
112. U.S. FDA: d-Nav System, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K181916, Last accessed: Apr 18, 2025
113. U.S. FDA: Dario Blood Glucose Monitoring System, URL: https://www.accessdata.fda.gov/cdrh_docs/pdf15/K150817.pdf, Last accessed: Apr 18, 2025
114. U.S. FDA: DreaMed Advisor Pro, URL: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K191370.pdf, Last accessed: Apr 18, 2025
115. U.S. FDA: GI Genius, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN200055, Last accessed: Apr 18, 2025
116. U.S. FDA: One Drop Blood Glucose Monitoring System, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K161834, Last accessed: Apr 18, 2025
117. U.S. FDA: Welldoc Bluestar System, URL: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190013.pdf, Last accessed: Apr 18, 2025
118. U.S. Food and Drug Administration: 510(k) Premarket Notification, https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmn.cfm
119. U.S. Food and Drug Administration: MAUDE - Manufacturer and User Facility Device Experience, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/search.cfm, Last accessed: Apr 18, 2025
120. U.S. Food and Drug Administration: MAUDE Adverse Event Report: IRHYTHM TECHNOLOGIES, INC ZEUS SYSTEM; COMPUTER, DIAGNOSTIC, PROGRAMMABLE, URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=19427744, Last accessed: Apr 18, 2025
121. U.S. Food and Drug Administration: Product Code Classification Database, URL: https://www.fda.gov/medical-devices/classify-your-medical-device/product-code-classification-database, Last accessed: Apr 18, 2025
122. U.S. Food and Drug Administration: Recalls, Corrections and Removals (Devices), URL: https://www.fda.gov/medical-devices/postmarket-requirements-devices/recalls-corrections-and-removals-devices, Last accessed: Apr 18, 2025
123. U.S. Food and Drug Administration: Zeus system, URL: https://www.accessdata.fda.gov/cdrh_docs/pdf22/K222389.pdf, Last accessed: Apr 18, 2025
124. Vargas, D.V., Su, J.: Understanding the one-pixel attack: Propagation maps and locality analysis. In: CEUR Workshop Proceedings. vol. 2640 (2020)
125. Wang, W., Yao, Y., Liu, X., Li, X., Hao, P., Zhu, T.: I can see the light: Attacks on autonomous vehicles using invisible lights. In: ACM SIGSAC CCS. pp. 1930–1944 (2021)
126. Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., Ren, K.: Threats to training: A survey of poisoning attacks and defenses on machine learning systems. ACM Computing Surveys **55**(7), 1–36 (2022)
127. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

128. Xu, Y., Tran, D., Tian, Y., Alemzadeh, H.: Analysis of cyber-security vulnerabilities of interconnected medical devices. In: 2019 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (2019)
129. Yaqoob, T., Abbas, H., Atiquzzaman, M.: Security Vulnerabilities, Attacks, Countermeasures, and Regulations of Networked Medical Devices—A Review. IEEE Communications Surveys & Tutorials **21**(4), 3723–3768 (2019)
130. Yoo, T.K., Choi, J.Y.: Outcomes of adversarial attacks on deep learning models for ophthalmology imaging domains. JAMA ophthalmology **138**(11), 1213–1215 (2020)
131. Young, W., Leveson, N.: Systems thinking for safety and security. In: Proceedings of ACM ACSAC '13. p. 1–8. Association for Computing Machinery (2013)
132. Young, W., Porada, R.: System-theoretic process analysis for security (STPA-SEC): Cyber security and STPA. In: STAMP Conference. pp. 27–30. MIT Press (2017)
133. Yu, J., Qiu, K., Wang, P., Su, C., Fan, Y., Cao, Y.: Perturbing BEAMs: EEG adversarial attack to deep learning models for epilepsy diagnosing. BMC Medical Informatics and Decision Making **23**(1), 115 (2023)
134. Zhou, X., Ahmed, B., Aylor, J.H., Asare, P., Alemzadeh, H.: Hybrid knowledge and data driven synthesis of runtime monitors for cyber-physical systems. IEEE Transactions on Dependable and Secure Computing (2023)
135. Zhou, X., Kouzel, M., Ren, H., Alemzadeh, H.: Design and validation of an open-source closed-loop testbed for artificial pancreas systems. In: 2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 1–12. IEEE (2022)