Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems

AKM Iqtidar Newaz[†], Nur Imtiazul Haque[††], Amit Kumar Sikder[†],
Mohammad Ashiqur Rahman[††], and A. Selcuk Uluagac[†]
[†]*Cyber-Physical Systems Security Lab,* [††]*Analytics for Cyber Defense Lab*
Department of Electrical and Computer Engineering
Florida International University, Miami, USA
{anewa001, nhaqu004, asikd003, marahman, suluagac}@fiu.edu

***Abstract*—The increasing availability of healthcare data requires accurate analysis of disease diagnosis, progression, and realtime monitoring to provide improved treatments to the patients. In this context, Machine Learning (ML) models are used to extract valuable features and insights from high-dimensional and heterogeneous healthcare data to detect different diseases and patient activities in a Smart Healthcare System (SHS). However, recent researches show that ML models used in different application domains are vulnerable to adversarial attacks. In this paper, we introduce a new type of adversarial attacks to exploit the ML classifiers used in a SHS. We consider an adversary who has partial knowledge of data distribution, SHS model, and ML algorithm to perform both targeted and untargeted attacks. Employing these adversarial capabilities, we manipulate medical device readings to alter patient status (disease-affected, normal condition, activities, etc.) in the outcome of the SHS. Our attack utilizes five different adversarial ML algorithms (HopSkipJump, Fast Gradient Method, Crafting Decision Tree, Carlini & Wagner, Zeroth Order Optimization) to perform different malicious activities (e.g., data poisoning, misclassify outputs, etc.) on a SHS. Moreover, based on the training and testing phase capabilities of an adversary, we perform white box and black box attacks on a SHS. We evaluate the performance of our work in different SHS settings and medical devices. Our extensive evaluation shows that our proposed adversarial attack can significantly degrade the performance of a ML-based SHS in detecting diseases and normal activities of the patients correctly, which eventually leads to erroneous treatment.**

***Index Terms*—Smart Healthcare System, Smart Medical Devices, Adversarial Machine Learning**

I. Introduction

Each year the healthcare industry is generating data at a staggering rate that is expected to reach 1,656 zettabytes by 2025 [1]. This healthcare data have clinical, financial, and operational value if analyzed properly to extract important features. In this context, Machine Learning (ML) can be an effective tool for managing healthcare data. Recently, several ML-based healthcare applications have been proposed to perform different healthcare functionalities, such as disease detection, early diagnosis, treatment plan, and antidote discovery [2], [3]. The integration of ML to identify and analyze clinical parameters has surely improved both the efficiency and quality of healthcare. The nature of feature extraction from the dataset of interest in ML can lead to patient-specific treatments and support, which can eventually lead to a reduction in medical costs and establish a better patient-doctor relationship. The global market of ML solutions in the healthcare sector is projected to reach $34 billion by 2025 [4].

ML plays a vital role in the healthcare industry, such as the development of the new medical procedure, processing patient data, and the treatment of chronic diseases. In parallel to this progress, ML models exhibit unpredictable and overly confident behavior outside of the training distribution, and adversarial examples are a subset of this broader problem. The discovery of adversarial examples has exposed vulnerabilities in the state-of-the-art ML systems [5]–[7]. An adversarial example is an input engineered to cause misclassification in the ML algorithms. Indeed, adversarial ML has gained much popularity in the healthcare domain because of the possible limitations of the current ML models. For instance, an adversary may add new adversarial data to a healthcare ML model to falsely classify a hypothyroid patient [8]. Researchers also reported several adversarial attacks against ML model in medical image processing to alter the results by adding noises and misclassify a benign mole as malignant with high confidence [9], [10]. However, the healthcare system is aggressively adapting ML model to improve disease detection and patient treatment without explicitly addressing these reported adversarial attacks.

In this paper, we present a new type of adversarial attacks that detects the pitfalls of a ML-based application in a smart healthcare system (SHS). In a SHS, multiple medical devices can be connected with each other to share the patient's vitals and detect normal and disease-affected activities by correlating various body functions. Our attack targets the underlying ML model used in a SHS for threat detection, disease recognition, and normal activity identification. In our attack, an adversary utilizes state-of-the-art adversarial attacks (HopSkipJump, Fast Gradient Method, Carlini & Wagner, Decision Tree, Zeroth Order Optimization) against the ML-based SHS to perform both white-box and black-box attacks [11]. We show that the proposed attack can successfully misclassify the patient's state and manipulate the outcome to a specific state by utilizing only partial knowledge of the ML model. To evaluate our attack, we consider a SHS correlating a patient's vitals collected from 8 different smart medical devices and detecting 11 benign scenarios, including 6 normal and 5 disease-affected activities using different ML models. Our evaluation shows that the proposed attack can achieve 32.27% accuracy drop and 20.43% success rate in untargeted and targeted attacks, respectively, leading to misclassification of the normal user activities and disease-affected scenarios in a SHS.

Contributions: In summary, our contributions are three-fold:

- We present an adversarial ML-based, data-driven attack to identify flaws in the underlying ML model in a SHS. We consider two different attacks: training phase attack (poisoning) and testing phase attack (evasion).

- We implemented four white-box attacks and one black-box attack utilizing 5 different adversarial ML algorithms.
- We evaluated our proposed attack against a ML-based SHS consisting of 8 different smart medical devices. Our extensive evaluation illustrates that it can successfully downgrade the accuracy of a ML model in a SHS.

Organization: We provide an overview of adversarial attacks in healthcare systems in Section II. The detailed overview of SHS, as well as the adversarial ML, is provided in Section III. In Section IV, we explain our threat model considering adversary goals, capabilities, and methodology. We explore the feasibility of our attack in the case of various SHS scenarios in Section V. We conclude the paper in Section VI.

II. Related Work

In this section, we discuss existing adversarial attacks to SHSs and explain how our attack model is different from the existing ones.

In recent years, several adversarial attacks on ML models have been reported by the researchers where the adversaries manipulate the input data distribution to cause incorrect classification in the output. Most of the adversarial attacks target the input data distribution to alter the training data and misclassify the output [8], [12]. Again, different ML models (e.g., Neural networks, deep neural networks, etc.) are exposed to small modifications of the input data during test time which can be used by the adversaries to manipulate the data distribution by enforcing poisoned test data [13]. In ML-based healthcare systems, adversaries mostly try to alter the data distribution in multi-layer ML classifier to change the predicted disease [9]. Most of the attacks in healthcare domain target medical imaging data to alter the predicted disease. Mirsky et al. showed how an attacker could use a 3D conditional Generative Adversarial Network (GAN) to add or remove evidence of medical conditions from volumetric (3D) medical scans [14]. Taghanaki et al. tested two state-of-the-art Deep Neural Network (DNN) for chest X-ray images against ten different adversarial attacks and showed that DNN in medical image classification is vulnerable against adversarial attacks [15]. Kim et al. presented how a subtle universal adversarial perturbations can be added to a medical image to alter the predicted labels with high confidence [16]. In a recent work, researchers proposed a novel approach for generating adversarial examples to attack Convolutional Neural Network (CNN) based segmentation models for medical images [17]. Another type of adversarial attacks can be executed in Electronic Health Records (EHR) to change the future treatment plan of a patient. Choi et al. used GAN to learn the distribution of real-world multi-label discrete EHR and generate adversarial patient records [18]. Sun et al. presented a framework that identifies the susceptible locations in a time sequence medical records utilizing adversarial attacks on deep predictive models [19].

Difference with Existing Attacks: Our threat model is an entirely new approach to attack a ML-based SHS. While most of the works consider single medical device data or stored healthcare data to apply adversarial attacks, in our work, we consider a connected multi-device SHS utilizing ML models for disease and normal activity identification. The main differences between existing attacks and our approach can be noted as follows: (1) Most of the prior works focus on manipulating data distribution of a ML model in offline mode. Here, we consider a realtime ML-based SHS and manipulate the input data distribution of different medical devices to perform the adversarial attacks. (2) While the existing attacks focus on generating multi-label discrete EHR [18], we detect the minimum number of the compromised medical devices in a SHS to initiate an attack. (3) Rather than depending on training model and full knowledge of the data distribution of ML algorithm, we consider the threshold of data injection for a successful attack using partial knowledge of the data range [12]. (4) Different from identifying susceptible locations in offline medical records, our attack finds the critical devices in realtime SHS to perform adversarial attacks which has higher impact on the patient's status and overall treatment [19]. (5) Unlike adversarial attacks on medical images [15], [9], [16], [17], we consider a ML-based SHS as our attack target/platform.

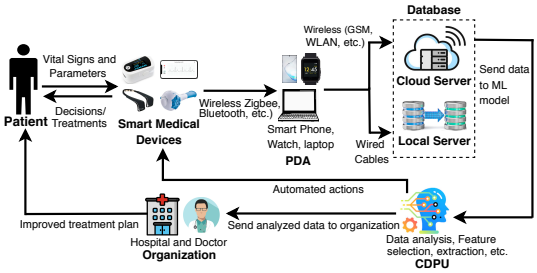


Fig. 1: An example of a smart connected healthcare system.

III. Background

In this section, we discuss different components of SHS and different design assumptions that we have considered in our attack. In addition, we discuss the impact of adversarial ML in the context of a SHS.

A. Connected Smart Healthcare System Overview

A SHS can consist of single or a group of smart medical devices (e.g., wearables, wireless devices, implantable devices, etc.) to collect data from a patient's body to provide improved treatments and realtime monitoring [20]. SHS considers different medical (patient's vitals) and non-medical parameters (e.g., physical status, location of the patient, etc.) to understand the overall condition of the patient and provides realtime monitoring. Smart medical devices take the vital signs as an analog signal, convert them to a digital signal and send information as a network packet to a Personal Digital Assistant (PDA) via wireless technology (Bluetooth, Zigbee, etc.), which can consist of a smartphone, a laptop, or a smartwatch. PDA works as a user interface and forwards data to a database (e.g., cloud server, local server, etc.). The database sends data to a Central Data Processing Unit (CDPU), which uses a ML model to select and extract features from the dataset. CDPU runs the ML model to detect patient disease status, normal activity of a patient, and threats in a SHS. Next, it takes automated actions (e.g., pushing a new dose of medicine, change of medication, etc.) to provide improved treatment for the patient and sends analyzed data to the authorized entity (hospital or doctor). Finally, the doctor sends a notification to the patient with updated treatment plans for his health status. Figure 1 shows an example of SHS with several smart medical devices (e.g., EEG, ECG, pulse oximeter, etc.) connected to a patient's body to collect data for different vital signs (e.g., blood oxygen, neural activities, heart rate, etc.).

B. Adversarial ML in a SHS

To understand how adversarial ML works in a SHS, we assume a ML model M is trained over input pair (X,y) from

the data distribution μ with a randomized training procedure of randomness r (e.g., random weight initialization, dropout, etc.). Here, inputs (X,y) are smart medical devices data that are collected and preprocessed before running into the model. The model parameters θ is represented as follows: $\theta \leftarrow train(M, X, y, r)$. In a white-box attack on this ML model, an adversary has complete knowledge about the model (M) used for classification, and the algorithm used in training. He has access to the training data distribution (μ) and is able to identify the feature space where the model is vulnerable. In a black-box attack, an adversary does not have knowledge about the model but uses information about the settings or past inputs to analyze the vulnerability of the ML model.

IV. Threat Model

In this section, we discuss the adversarial goals and attackers capabilities considered in our proposed attack. We also describe our attack assumptions focusing on a use case scenario and explain the attack methodology in a ML-based SHS.

A. Adversarial Goals

An adversary attempts to provide an input to a ML-based SHS to alternate the decision of the ML classifiers and manipulate the automated decisions of the SHS. We categorize the goals of an adversary in the following three categories based on the impact on the classifier output integrity:

- Untargeted Attack (UA): An adversary tries to alter the output of the ML classifiers of a SHS from the actual output. For example, a high blood pressure patient will be predicted to a different disease resulting in a malicious treatment plan.
- Targeted Attack (TA): An adversary attempts to alter the input data of the ML classifier to change the output to a specific class. For instance, a high blood pressure patient's data is manipulated in such a way that it is predicted as a specific condition of abnormal oxygen levels.
- Targeted Device Attack: The adversary tries to find the minimum number of devices needed to be compromised to launch an attack in a ML-based SHS. For instance, to change the status of a stressed patient to a heart attack, it may only need to compromise the oxygen saturation device.

B. Adversarial Capabilities

To understand the attack scope of our work, we assume a patient (P) admits to a hospital having breathing problems (abnormal oxygen level) (Figure 2). For emergency monitoring, a SHS consisting of an ECG device, a sphygmomanometer, a pulse oximeter, and an EEG is placed on P for cardiac, blood pressure, oxygen level, and neurological activity monitoring, respectively. This SHS uses a ML-based classifier to predict the disease and perform realtime monitoring of the patient. We assume that the devices are working correctly, and no compromised device is in the system. Here, we introduce a new adversarial attack that can exploit a SHS and change the patient's disease status to provide the wrong treatment. To perform the adversarial attack on the ML model, we consider following capabilities for the adversary:

- Data Distribution: As different smart medical devices generate different ranges of data, an adversary may know partial data distribution of the devices. Based on this, it may manipulate a data value within a certain threshold to change patient disease status or normal activity pattern.
- SHS Architecture: An adversary may know the complete or partial architecture of the SHS including number of devices, device correlation, etc. to perform an attack.
- Output Label: An adversary may know the output labels (e.g., disease states and normal activity pattern) of the ML model to initiate an attack.
- ML Model: Different ML models (e.g., Decision Tree, Random Forest, etc.) are used for disease and user activity detection. An adversary may have the knowledge of underlying ML model used to classify the patient status.

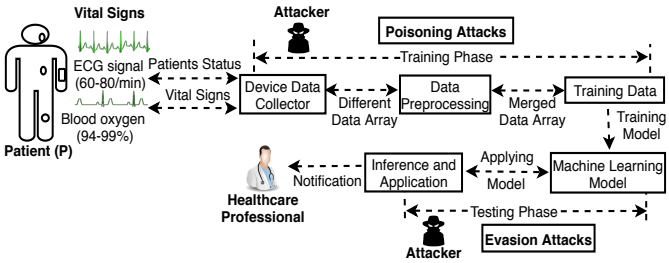


Fig. 2: An example of adversarial attack in a SHS.

C. Attack Methodology

A ML-based SHS can be considered as a data processing pipeline where the vital signs collected from the patients are analyzed to diagnose the disease and provide necessary treatment. A primary sequence of operations of the SHS is illustrated in Figure 2. A device data collector module collects data representing patient's vital signs and statuses from different smart medical devices and forwards them to the data preprocessing module. A data preprocessing module samples the data according to the corresponding sampling frequencies and saves them as an array. For example, a heart rate monitoring device monitors the heart rate of a patient in a minute, whereas the ECG device monitors the cardio-vascular state of a patient in every 10 seconds. The sampled data are used to train the ML model of the SHS for disease detection and realtime monitoring. The training data are labeled with different diseases and benign states (e.g., high blood pressure, low sugar, etc.) to understand the data pattern for different scenarios. In the testing phase, physiological data collected from the patients are analyzed based on the previously trained ML model to detect different diseases or benign states of the patient. Here, our attack methodology can be defined in the context of the data processing pipeline. An adversary attempts to manipulate either the collection or the processing of data to corrupt the ML model, hence altering the original output. We divide the attack into the following two categories in a SHS:

Poisoning Attack: This type of attack takes place during the training time of ML in a SHS (Figure 2). In this attack, an adversary carefully manipulates the training data to compromise the whole learning process. As explained in Section IV-B, an adversary may know the data distribution of the ML model and can change the value of the input data to a certain threshold. Here, the adversary can use data injection, modification, and logic corruption methods to manipulate the training data. In data injection, an adversary does not have any access to the training data and algorithm but can augment a new data to the training set. In data modification, an adversary poisons the training data directly by modifying the data before it is used for training the target model. In logic corruption, an adversary has

TABLE I: Devices and parameters considered for monitoring health condition [20]–[23].

Device Monitoring Type	Model	Feature Parameter Value	Database
Heart Rate and Blood Pressure	QuadioArm	60-100 beats per minute, Systolic (120 mm Hg) and Diastolic (80 mm Hg)	Fetal ECG Synthetic Database, Data.Gov
Blood Glucose	MiniMed™ 670G Insulin Pump System	70-130 mg/dl	UCI ML Database of Diabetes
Blood Oxygen	iHealth Air Wireless Pulse Oximeter	Oxygen Saturation Level $\geq 94\%$	Pattern Analysis of Oxygen Saturation Variability
Respiratory and Sweating Rate	QuardioCore	12-20 Breaths per minute, 0.5μ/min/cm^2	BIDMC PPG and Respiration Dataset
Blood Alcohol	Scram Continuous Alcohol Monitoring (Cam)	0.08 g/dl	StatCrunch Dataset
Blood Hemoglobin	Germaine AimStrip Hb Hemoglobin Meter	12.3-17.5 g/dl	Hemoglobin Data in DHS Survey
Neural Activity	Emotiv Insight	Delta (0.5-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (16-24 Hz) & EEG data	EEG Data
Sleep and Human Motion	Fitbit Versa Smart Watch	REM and NREM sleep cycle	The CAP Sleep Database

the ability to tamper the ML model. Using these capabilities, an adversary can affect the overall learning process of the ML model of the the SHS to misdiagnosed the test data which could lead to mistreatment of a patient.

Evasion Attack: In evasion attack, adversary tries to deceive the SHS by enforcing adversarial samples during the testing phase. An adversary does not have any influence over the training data but can access the ML model to obtain sufficient information. As a result, it attacks the ML model and manipulates the model to misclassify the patient status in a SHS. Generally, evasion attacks are classified in two categories: *white-box attack* and *black-box attack*.

(1) *White-Box Attacks:* In a white-box attack, an adversary has complete knowledge about the ML model used in a SHS (e.g., type of neural network along with the number of layers). The adversary knows the algorithm used in the training phase and can access the training data distribution. Moreover, it knows the parameters of fully trained architecture. We perform HopSkipJump, Fast gradient method, Carlini & Wagner, and Decision tree-based attack in the ML model of a SHS.

- HopSkipJump Attack: HopSkipJump is an efficient query algorithm and a powerful decision-based generator of adversarial examples which utilizes distance-vector as a hyper-parameter [24]. In our attack, we obtained optimal solution by using Chebyshev distance, which calculates the largest magnitude among each element of a vector to determine the distance between adversarial and the original samples.
- Fast Gradient Method (FGM) Attack: The fast gradient method uses the gradient of the underlying model to find adversarial examples [5]. The original input is manipulated by adding or subtracting a small error in the direction of the gradient with the intent to change the behavior of the learning model. In our model, we considered an attacker's capability (threshold) as adding a small error in the direction of gradient to temper classification of the model.
- Carlini & Wagner (C&W) Attack: This is the state-of-the-art white-box attack in which targeted adversarial attacks are considered as an optimization problem to take advantage of the internal configuration of ML model [25]. In our model, we found Euclidean distance to be optimal for computing the difference between adversarial and original examples.
- Crafting Decision Tree Attack: Here, an adversary exploits the underlying tree structure of the decision tree classifier model [26]. For a given sample and a tree, an adversary searches for leaves with different classes in the neighborhood of the leaf corresponding to the decision tree's original prediction for the sample.We performed this attack for enumerating the minimum number of devices that need to be compromised for initiating an attack.

(2) *Black-Box Attack:* In a black-box attack, an adversary does not know the ML model in the SHS and uses the information about settings or past inputs to analyze the vulnerability of the model. We perform the Zeroth Order Optimization attack in a SHS as a black-box attack.

- Zeroth Order Optimization (ZOO) Attack: ZOO attack is based on a coordinate descent method using only the zeroth-order oracle (without gradient information), which can effectively attack the black-box ML model. This is called a state-of-the-art black-box attack [27].

V. Exploring Various Attack Scenarios

To demonstrate the feasibility of our proposed attack, we performed the poisoning and evasion attacks successfully against a ML-based SHS. We consider several research questions to evaluate our proposed adversarial attack.

RQ1 What is the impact of proposed attack after performing poisoning attack in a SHS? (Sec V-B)

RQ2 What is the impact of our attack after performing untargeted attack to find most significant devices? (Sec V-C)

RQ3 What is the impact of proposed white-box and black-box attacks on underlying ML model? (Sec V-D)

RQ4 What is the performance of ML model in a SHS after performing threshold-based attacks? (Sec V-E)

TABLE II: Device-activity correlation in SHS [20], [28], [29]

		ECG	SW	BP	GL	BR	OX	SM	HG	AL	NA	HM
Disease Type	High Blood Pressure	-	✓	✓	✓	-	✓	✓	✓	✓	✓	-
	High Cholesterol	-	✓	✓	✓	-	✓	-	✓	-	✓	-
	Excessive Sweating	✓	✓	✓	✓	-	✓	-	✓	-	✓	✓
	Abnormal Oxygen Level	✓	-	✓	✓	✓	✓	✓	-	-	✓	✓
	High or Low Blood Sugar	✓	✓	✓	✓	-	✓	-	✓	-	✓	-
Activity Type	Sleeping	✓	-	✓	✓	✓	✓	-	-	-	-	-
	Walking	✓	✓	-	✓	✓	✓	-	✓	-	✓	✓
	Stress	✓	✓	✓	-	✓	-	-	-	-	✓	-
	Exercise	✓	✓	✓	✓	✓	✓	-	-	-	✓	✓
	Heart-Attack	✓	✓	-	-	✓	-	-	-	-	✓	-
	Stroke	✓	-	✓	-	-	-	-	✓	-	✓	✓

A. Environment Setup and Methodology

We consider a SHS consisting of eight different devices detecting five different disease states and six different regular activities. We collected data from eight different smart medical devices available on various public healthcare databases that gave us ten vital signs: the heart rate (ECG), blood pressure (BP), blood glucose (GL), oxygen (OX) saturation, blood hemoglobin (HG), breathing rate (BR), alcohol level (AL), neural activity (NA), human motion (HM) and sleep monitoring (SM) (Table I). These vital signs are represented by fifteen different features and the correlation among these features is presented in figure 3. We considered the ranges of different vital signs of humans (e.g., heart rate, blood pressure, etc.) as the normal state. Our considered SHS uses different ML algorithms to detect five different disease scenarios (high blood pressure, high cholesterol, excessive sweating (SW), abnormal oxygen level, abnormal blood sugar) and six regular user activities (sleeping, walking, exercise, stress, heart attack, and stroke situation) based on the collected device data. The device-activity correlation is shown in Table II.

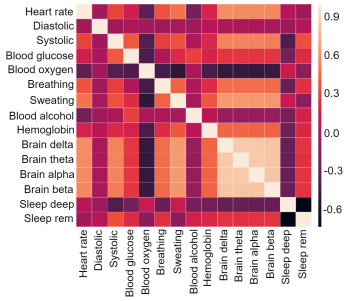


Fig. 3: Correlation matrix of selected features used in SHS.

As the underlying ML model of the SHS, we considered four different ML algorithms (Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN), Logistic Regression (LR)) because of multi-class classification and hierarchical patterns in our dataset. We collected 17,000 data instances for healthy and disease-infected people and divided our data into two sections, where 70% of the collected dataset was used to train the ML model, and 30% of the collected data was used for testing [30]. We executed the attack in this SHS by simulating evasion and poisoning attacks based on the attack methodology described in Section IV. For a poisoning attack, we considered that an adversary poisoned the training data while he/she had the access to the ML model. Here, an adversary manipulates the output status of the patient to alter the detected disease or normal user activity could be manipulated into a disease state. To perform the evasion attack, we considered that an adversary adjusted malicious samples during the testing phase. For evasion attacks, we also conducted white-box and black-box attacks by performing gradient-based, decision boundary-based, and coordinate-based attacks to the ML algorithms of the SHS. We built a simulation environment for SHS in *MATLAB* using digital signal processing toolbox and adversarial robustness toolbox [31] (Python library) for simulating different white-box and black-box attacks. We used *Keras* and *Scikit-Learn* library for training and testing the dataset in Google Colab platform.

TABLE III: Performance of data poisoning attacks.

	Before Attack	Accuracy Drop		
Underlying Algorithms	**Accuracy**	**10% poisoning**	**20% poisoning**	**30% Poisoning**
RF	95.37	1.65	2.03	2.15
DT	90.16	4.31	15.88	27.31
ANN	91.42	10.28	13.14	27.32
LR	88.11	9.28	18.93	28.21

In the evaluation, we used accuracy drop of the SHS's ML model from the actual accuracy as our performance metric. While accuracy refers to the effectiveness of the ML model in detecting different patient's states, adversarial samples lead to a decrease in the accuracy of the ML model. For targeted attacks, we considered attacker success rate as we performed the targeted attack in a specific portion of the total dataset.

B. Evaluation with manipulated training data

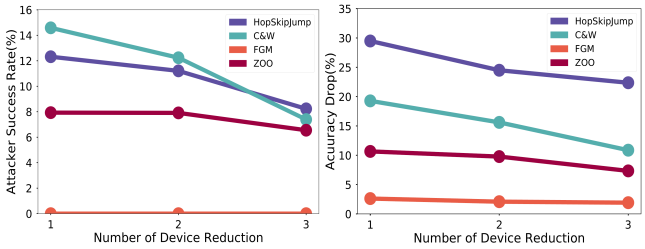
In a SHS, an adversary can have access to the CDPU in the training phase and inject bad data into the underlying ML model's training pool. In that case, the adversary can learn the training model and inject bad data (poisoning attack) to manipulate the decision of the SHS. In our poisoning attack, we have altered the vital status of 10%, 20%, and 30% training data for compromising SHSs' model performance. Table III presents the performance of the poisoning attack in a SHS. We used RF, DT, ANN, LR as our underlying ML algorithms for training the SHS. The SHS achieves highest accuracy of 95.37% in detecting different patient's states using RF. After poisoning the training dataset, the accuracy of the SHS dropped 2.15% for 30% data poisoning. We can also observe that our approach achieved highest accuracy drop of 18.93% and 28.21% for 20% and 30% data poisoning, respectively, if the underlying ML model is LR. In the case of ANN with 10% data poisoning, the highest accuracy drop is 10.28%.

C. Significant device identification in different attacks

A SHS uses the correlation between different medical devices to identify the different vital status of a patient. We used this correlation to perform crafting decision tree untargeted attacks to identify the most significant devices in a SHS (Table IV). We considered the untargeted attack as it uses the whole training sample to generate adversarial examples. We can observe that for an untargeted attack to change the disease state from stroke to abnormal oxygen level, and hemoglobin devices are compromised. Similarly, glucose device is affected while changing disease state from high cholesterol to stroke. In summary, to perform an untargeted attack on underlying the ML model (DT) glucose, heart rate, and oxygen saturation devices are the most affected devices.

TABLE IV: Crafting DT for untargeted attack.

Attack	Current State	Final State	Affected Device	Device Count
Untargeted	High Cholesterol	Stroke	Glucose	1
	High Blood Pressure	Stroke	Glucose	1
	Abnormal Oxygen Level	Stroke	Glucose, Heart rate	2
	Stroke	Abnormal Oxygen Level	Glucose, Hemoglobin	2
	Sleeping	Drunk	Alcohol	1
	Stress	Heart Attack	Blood Oxygen	1



(a) Targeted attack (b) Untargeted attack

Fig. 4: Impact of device reduction in adversarial attacks.

Figure 4a illustrates the impact of device reduction in the performance of untargeted attack. We can observe reducing number of devices drops the success rate of the proposed attack significantly. For removing one device (glucose) and two devices (glucose, blood oxygen), we achieved the highest success rate of 14.59% and 12.24% using C&W attack. For removing three (glucose, oxygen, heart rate) devices, HopSkipJump can achieve the highest success rate of 8.23%. For an untargeted attack (Figure 4b), HopSkipJump and FGM achieved the highest and lowest accuracy drop respectively for removing 1, 2, and 3 devices from the SHS. In summary, performance of adversarial attacks decrease with the reduction of devices in SHS.

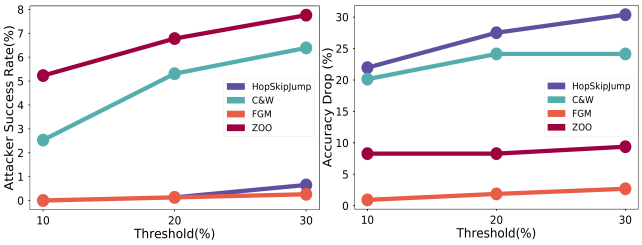
D. Evaluation with Different Attack Algorithms

To evaluate the performance of the proposed adversarial attack, we performed white box and black box attacks in a SHS using different adversarial algorithms. From Table V, we can observe the highest accuracy drop (32.27%) and success rate (15.68%) can be achieved using HopSkipJump for untargeted attack and targeted attack respectively. On the contrary, FGM achieved the lowest accuracy drop and success rate while the

underlying model was LR. For the black-box, we performed both untargeted and targeted ZOO attacks and obtained 12.29% accuracy drop and 8.22% success rate respectively, where the underlying model was RF.

TABLE V: Performance of white-box and black-box attacks.

Adversary Capability	Attack Algorithms	Underlying ML model	Actual Accuracy	Accuracy Drop (UA)	Success Rate (TA)
White-box	HopSkipJump	DT	90.16	32.27	15.68
	C&W	ANN	85.14	24.14	20.43
	FGM	LR	88.11	3.02	0.26
Black-box	ZOO	RF	90.21	12.29	8.22



(a) Targeted attack (b) Untargeted attack

Fig. 5: Threshold based attacks on a SHS.

E. Evaluation with Considering Different Thresholds

We performed threshold-based attack as described in Section IV. From Figure 5a, we can observe that for an targeted attack HopSkipJump and FGM failed to achieve any success rate for 10% threshold-based attack. ZOO attack achieved highest success rate for all three threshold-based attacks in a SHS. During an untargeted attack (Figure 5b), ZOO attack algorithms had same accuracy drop of 8.28% for both 10% and 20% thresholds. FGM had the lowest accuracy drop for all three attack. In summary, HopSkipJump achieved highest accuracy drop for threshold-based attacks.

VI. Conclusion

The healthcare industry is generating a vast amount of data for patients and ML is being used to analyze these data for different applications such as early disease diagnose, improved treatment plan, realtime patient monitoring, etc. However, the ML models used in different medical applications have weaknesses in terms of unpredictability and outside of the training distribution data. To address these shortcomings, in this paper, we presented a new adversarial ML-based attack to identify the pitfalls in the underlying ML model of a SHS. We evaluated our attack considering different adversarial settings and medical setup in a SHS. Our extensive evaluation shows that it can achieve very promising results for untargeted and targeted attacks to misclassify the normal user activities and disease affected scenarios in SHSs.

VII. Acknowledgment

This work was partially supported by US NSF-CAREER-CNS-1453647. The views here belong to the authors only.

References

[1] "Data age 2025," https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.
[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.
[3] A. I. Newaz, A. K. Sikder, L. Babun, and A. S. Uluagac, "Heka: A novel intrusion detection system for attacks to personal medical devices," in *2020 IEEE Conference on Communications and Network Security*, 2020.
[4] "Are you ready to rumble?" https://www.beckershospitalreview.com/healthcare-information-technology/ml-and-ai-are-taking-healthcare-by-storm-are-you-ready-to-rumble.html/, 2018.
[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[6] A. Abusnaina, A. K., H. A, J. Park, A. Anwar, and A. Mohaisen, "Adversarial learning attacks on graph-based iot malware detection systems," in *39th International Conference on Distributed Systems*. IEEE, 2019.
[7] M. H. Shahriar, N. I. Haque, M. A. Rahman, and M. Alonso Jr, "G-ids: Generative adversarial networks assisted intrusion detection system," *arXiv preprint arXiv:2006.00676*, 2020.
[8] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE journal of biomedical and health informatics*, 2014.
[9] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.
[10] A. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "A survey on security and privacy issues in modern healthcare systems: Attacks and defenses," *arXiv preprint arXiv:2005.07359*, 2020.
[11] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.
[12] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, 2019.
[13] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time." Springer, 2013.
[14] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "Ct-gan: Malicious tampering of 3d medical imagery using deep learning," *arXiv preprint arXiv:1901.03597*, 2019.
[15] S. A. Taghanaki, A. Das, and G. Hamarneh, "Vulnerability analysis of chest x-ray image classification against adversarial attacks," in *Understanding ML in Medical Image Computing Application*. Springer, 2018.
[16] H. Kim, D. C. Jung, and B. W. Choi, "Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: Adversarial attacks," *Journal of the Korean Society of Radiology*, 2019.
[17] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Intelligent image synthesis to attack segmentation cnn using adversarial learning," in *Simulation, Synthesis in Medical Imaging*. Springer, 2019.
[18] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," *arXiv preprint arXiv:1703.06490*, 2017.
[19] M. Sun, F. Tang, J. Yi, F. Wang, and J. Zhou, "Identify susceptible locations in medical records via adversarial attacks on deep predictive models," in *24th ACM SIGKDD on Data Mining*, 2018.
[20] A. I. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "Healthguard: A machine learning-based security framework for smart healthcare systems," in *IEEE SNAMS*, 2019.
[21] "Hypertension," https://catalog.data.gov/dataset/hypertension/, 2018.
[22] M. G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, and Hirshkowitz, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep," *Sleep medicine*, vol. 3, no. 2, pp. 187–199, 2002.
[23] A. F. Ashrafi, A. M. I. Newaz, R. Ajwad, M. M. Tanvee, and M. Mottalib, "A modified algorithm for dna motif finding and ranking considering variable length motif and mutation," in *2nd International Conference on Recent Trends in Information Systems*. IEEE, 2015.
[24] T. Klove, T.-T. Lin, S.-C. Tsai, and W. Tzeng, "Permutation arrays under the chebyshev distance," *IEEE Transactions on Information Theory*, 2010.
[25] F. Croce, J. Rauber, and M. Hein, "Scaling up the randomized gradient-free adversarial attack reveals overestimation of robustness using established attacks," *arXiv preprint arXiv:1903.11359*, 2019.
[26] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.
[27] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *ACM*, 2017.
[28] A. K. Sikder, L. Babun, H. Aksu, and A. S. Uluagac, "Aegis: a context-aware security framework for smart home systems," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019.
[29] A. K. Sikder, H. Aksu, and A. S. Uluagac, "A context-aware framework for detecting sensor-based threats on smart devices," *IEEE Transactions on Mobile Computing*, 2019.
[30] "Splitting the data into training and evaluation data," https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-the-data-into-training-and-evaluation-data.html/.
[31] "Adversarial robustness 360 toolbox (art) v1.0," https://github.com/IBM/adversarial-robustness-toolbox/.