Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices

Nicola Paoletti
Royal Holloway, University of London, UK

Zhihao Jiang
ShanghaiTech University, China

Md Ariful Islam
Texas Tech University, USA

Houssam Abbas
University of Pennsylvania, USA

Rahul Mangharam
University of Pennsylvania, USA

Shan Lin
Stony Brook University, USA

Zachary Gruber
Stony Brook University, USA

Scott A. Smolka
Stony Brook University, USA

ABSTRACT

An Implantable Cardioverter Defibrillator (ICD) is a medical device used for the detection of potentially fatal cardiac arrhythmias and their treatment through the delivery of electrical shocks intended to restore normal heart rhythm. An ICD *reprogramming attack* seeks to alter the device's parameters to induce unnecessary therapy or prevent required therapy. In this paper, we present a formal approach for the synthesis of ICD reprogramming attacks that are both *effective*, i.e., lead to fundamental changes in the required therapy, and *stealthy*, i.e., are hard to detect. We focus on the *discrimination algorithm* underlying Boston Scientific devices (one of the principal ICD manufacturers) and formulate the synthesis problem as one of multi-objective optimization. Our solution technique is based on an Optimization Modulo Theories encoding of the problem and allows us to derive device parameters that are optimal with respect to the effectiveness-stealthiness tradeoff. Our method can be tailored to the patient's current condition, and readily generalizes to new rhythms. To the best of our knowledge, our work is the first to derive systematic ICD reprogramming attacks designed to maximize therapy disruption while minimizing detection.

CCS CONCEPTS

• **Security and privacy**; • **Theory of computation** → *Logic and verification*; • **Applied computing** → *Life and medical sciences*;

KEYWORDS

Medical device security, Reprogramming attack, Implantable Cardioverter Defibrillator, Arrhythmia discrimination, Model-based attack synthesis.

ACM Reference Format:
Nicola Paoletti, Zhihao Jiang, Md Ariful Islam, Houssam Abbas, Rahul Mangharam, Shan Lin, Zachary Gruber, and Scott A. Smolka. 2019. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. In *10th ACM/IEEE International Conference on Cyber-Physical Systems (with CPS-IoT Week 2019) (ICCPS '19), April 16–18, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3302509.3311044

ICCPS '19, April 16–18, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6285-6/19/04…$15.00
https://doi.org/10.1145/3302509.3311044

1 INTRODUCTION

An *Implantable Cardioverter Defibrillator* (ICD) is a medical device for the detection and treatment of potentially fatal arrhythmias such as ventricular tachycardia (VT) and ventricular fibrillation (VF). ICDs run embedded software that processes intracardiac signals, called *electrograms* (EGMs), to detect arrhythmias and deliver appropriate therapy in the form of electrical shocks. ICD software implements so-called *discrimination algorithms* which comprise multiple discrimination criteria (discriminators) for the detection and classification of arrhythmia episodes based on the analysis of EGM features such as ventricular intervals and signal morphology.

ICD discriminators feature a number of programmable parameters that, if adequately configured, ensure minimal rates of arrhythmia mis-classification [17]. In contrast, wrongly configured parameters can result in unnecessary shocks (*false positive* classification errors), which are painful and damage the cardiac tissue, and even worse can prevent required therapy (*false negatives*), leading to sudden cardiac death.

An ICD *reprogramming attack* is one that alters the device's parameters to induce mis-classification and inappropriate or missed therapy. Reprogramming attacks can significantly compromise patient safety, with high-profile patients being obvious targets (e.g. former US Vice President Cheney had his pacemaker's wireless access disabled to prevent assassination attempts [21]). Seminal work by Halperin et al. [9] demonstrated that ICDs can be accessed and reprogrammed by unauthorized users using off-the-shelf software radios. More recently, over half a million cardiac devices have been recalled by the FDA for security risks related to wireless communication [8], and researchers managed to gain control of a pacemaker/ICD by exploiting vulnerabilities in the device's remote monitoring infrastructure [22]. These incidents confirm that vulnerabilities in implantable cardiac devices exist, and a thorough investigation of cyber-attacks on ICDs is needed to improve device safety and security.

In this paper, we present a formal approach for the automated synthesis of ICD reprogramming attacks that are both *effective*, i.e., lead to fundamental changes in the required therapy, and

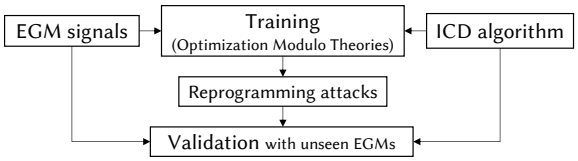


Figure 1: Overview of our method for synthesis of stealthy reprogramming attacks on ICDs.

stealthy, i.e., involve minimal changes to the nominal ICD parameters. Stealthy attacks are therefore difficult to detect and even if detected, would most likely be attributed to a clinician's error in configuring the device. We follow a model-based approach, as the attacks are not evaluated on the actual hardware but on a model of the ICD algorithm. We focus on the *Rhythm ID* algorithm implemented in Boston Scientific ICDs (one of the principal ICD manufacturers), which was compiled from device manuals and the medical literature [6, 28]. Slight variations on the discriminators used and computations performed by Rhythm ID are also found in the algorithms of the three other major ICD manufacturers. Thus, focusing on Rhythm ID does not limit the applicability of our approach.

Our method, illustrated in Figure 1, synthesizes device parameters that are optimal with respect to the effectiveness-stealthiness tradeoff (i.e., lie along the corresponding Pareto front). We formulate this synthesis problem as one of multi-objective optimization, and solve it using *optimization modulo theories* (OMT) techniques [5], an extension of SMT for finding models that optimize given objectives. OMT is uniquely suited to solve this problem, because the problem is combinatorial in nature (parameters can be configured from a finite set of values), and is also constrained by the behavior of the ICD algorithm, which can be adequately encoded as SMT constraints. The synthesized reprogramming attacks yield optimal effectiveness and stealthiness with respect to a set of *training EGM signals*. We employ the method of [12] to generate synthetic EGMs with prescribed arrhythmia. This allows the attacker to synthesize malicious parameters tailored to the victim's cardiac condition.

Why optimized attacks? The objective of this paper is to show that ICDs are vulnerable to stealthy reprogramming attacks. While it is already known that incorrect parameter values can lead to incorrect therapy, our work formally establishes to what *degree* these parameters need to be manipulated to produce *injurious* incorrect therapy, and device designers should be made aware of these results. We remark that our approach does not provide an exhaustive recipe for ICD attacks, as the actual algorithms on-board devices usually contain more decision branches than we have chosen to model, and indeed more than is described in the open literature. See Section 3 for further details about real-life attacks and countermeasures.

In summary, our main contributions are the following.

- We introduce, to the best of our knowledge, the first method for deriving systematic reprogramming attacks on cardiac devices designed to maximize therapy disruption while minimizing the likelihood of detection.
- We formulate the problem of synthesizing malicious parameters as a multi-objective optimization problem.
- We present a method, based on OMT techniques and an efficient SMT encoding of the ICD algorithm, for precisely solving this optimization problem.
- We evaluate our approach by synthesizing attacks for 19 different arrhythmias (i.e., *condition-specific* attacks), as well as more generic attacks (*condition-agnostic*) that are suitable when the attacker has little knowledge of the victim's condition. Our results demonstrate that some arrhythmias are particularly vulnerable, as only minor changes to the detection thresholds are sufficient to prevent the required therapy.
- We show that our approach is suitable for real-world attacks as it readily generalizes to unseen signals (i.e., *test EGMs*), representing the unknown EGMs of the patient.

2 BACKGROUND

ICDs are battery-powered devices implanted under the pectoral muscles in the chest and connected to the cardiac muscle through one (in single-chamber ICDs) or two (dual-chamber) leads that sense the electrical activity of the heart and deliver electrical defibrillation shocks when dangerous arrhythmia is detected (see Figure 2). Shocks are delivered through shocking coils located along the ventricular lead. ICDs also support anti-tachycardia pacing and cardiac pacing functions [19].

Sensed electrical signals are called *intracardiac electrograms (EGMs)*, which in a dual-chamber ICD are of three types: *atrial and ventricular EGMs*, describing the local, near-field electrical activity in the right atrium and ventricle, respectively; and the *shock EGM*, a far-field signal that gives a global view of the electrical activity, measured from the shock coil to the ICD can.

ICD discrimination algorithms are responsible for detecting tachycardia episodes and initiating adequate therapy based on the sensed EGMs. These algorithms are embedded in the device and employ signal-processing methods such as peak detection to identify cardiac events; viz. electrical activation of the atria and ventricles (heart beats). Therapy delivery depends on a number of discrimination criteria to distinguish between potentially fatal Ventricular Tachy-arrhythmias (VT) and non-fatal Supra-Ventricular Tachy-arrhythmias (SVTs).

Since an ICD only has three signals, there are a limited number of features that can be used as discriminators. Atrial rate, ventricular rate, and far-field ventricular morphology are the core features that all major ICD manufacturers employ [25]. To generalize to a large variety of physiological conditions and to avoid "over-fitting" the algorithm to known conditions, device manufacturers have adopted simple discriminators and decision tree-like to distinguish between SVT and VT.

2.1 ICD Discrimination Algorithm

Figure 3 illustrates the *Rhythm ID* algorithm implemented in Boston Scientific (BSc) ICDs. The algorithm consists of a number of discriminators arranged in a decision tree-like structure, where each discriminator depends on one or more programmable parameters. Leaves of the tree determine whether or not therapy is delivered during the current cardiac cycle.

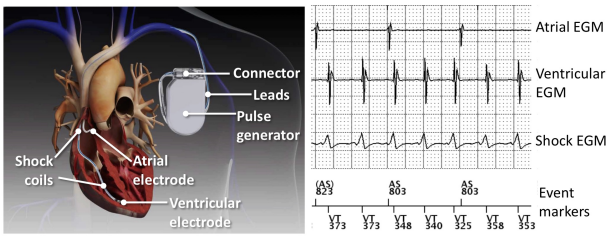


Figure 2: Left: illustration of a dual-chamber ICD. Right: sensed atrial, ventricular and shock electrograms. Event markers label sensed impulses (AS: atrial, VT: ventricular tachycardia) and corresponding intervals in milliseconds.

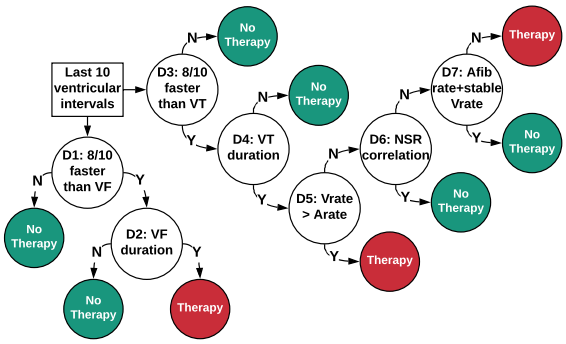


Figure 3: Discrimination tree of the Boston Scientific *Rhythm ID* algorithm. White nodes denote discrimination criteria. Any sequence of decisions eventually leads to either delivering (red) or not delivering (green) the therapy.

The parameters of the algorithm are given in Table 1. We consider the description of the Rhythm ID algorithm by Jiang et al. [12], where the authors provided a MATLAB implementation of the algorithm based on the manufacturer's manuals and the medical literature [6, 28]. This implementation faithfully captures the behavior of the Rhythm ID algorithm, as it was validated by demonstrating conformance to a BSc commercial ICD device on 11 test cases. The algorithm and its discriminators, described next, are executed at each ventricular event, which marks the end of the corresponding cardiac cycle.

D1, 8/10 faster that VF: this discriminator is true iff at least eight out of the last ten ventricular intervals (i.e., the time between two consecutive ventricular beats) are shorter than the programmable threshold VF_{th}. **D1** detects the onset of arrhythmia (VF in this case), as a high ventricular rate is a strong indication of VF. If **D1** is true, therapy is delivered only if the VF episode persists, which is checked by discriminator **D2**.

D2, VFduration: when in VF duration mode, the algorithm checks that at least six out of the last ten ventricular intervals are below VF_{th}, and that the last interval is below VF_{th}. If this criterion is not met, the algorithm exits the VF duration mode as the episode did not persist, and thus requires no therapy. If this criterion stays true for the entire VF duration (parameter VFdur), then therapy is given.

Name	Description	**Nominal** (Programmable)
VF_{th} (BPM)	VF detection threshold	**200** (110 : 5 : 210, 220 : 10 : 250)
VT_{th} (BPM)	VT detection threshold	**160** (90 : 5 : 210, 220)
$AFib_{th}$ (BPM)	AFib detection threshold	**170** (100 : 10 : 300)
VFdur (s)	Sustained VF duration	**1.0** (1 : 0.5 : 5, 6 : 1 : 15)
VTdur (s)	Sustained VT duration	**2.5** (1 : 0.5 : 5, 6 : 1 : 15, 20 : 5 : 30)
$NSRcor_{th}$	Rhythm Match score	**0.94** (0.7 : 0.01 : 0.96)
stb (ms^2)	Stability score	**20** (6 : 2 : 32, 35 : 5 : 60, 70 : 10 : 120)

Table 1: Parameters of the Rhythm ID algorithm, including nominal and programmable values [6]. AFib: atrial fibrillation. $n : k : m$ denotes the sequence $n, n + k, n + 2k, \ldots, m$. Thresholds are programmed in BPM (beats per minute) but the algorithm employs the corresponding time duration.

D3, 8/10 faster that VT: this discriminator is analogous to **D1**, but uses the VT threshold VT_{th}.

D4, VTduration: this discriminator is analogous to **D2**, but uses the VT threshold VT_{th} and the duration parameter VTdur. The difference with **D2** is that in this case, therapy is not given immediately at the end of the duration timer; rather, the algorithm ensures that the episode is not mistaken for SVT, as illustrated below.

D5, V rate > A rate: it is true iff over the last ten cardiac cycles, the average ventricular rate is at least 10 BPM faster the average atrial rate. If true, **D5** indicates that tachycardia originated in the ventricles and thus must be treated. Otherwise, the algorithm inspects **D6** and **D7**.

D6, NSR correlation: this criterion, also called *Rhythm Match*, compares the morphology of the far-field shock EGM with that of a pre-computed normal sinus rhythm (NSR) template. The two signals being similar suggests that the arrhythmia originated in the atria, indicating SVT (no therapy). In particular, for at least three out of the last ten cardiac cycles, the two signals should have a so-called feature correlation coefficient (FCC) greater than parameter $NSRcor_{th}$. The FCC is computed by looking at the voltages of the two signals at prescribed time-points. See [6] for more details on the computation of the FCC.

D7, AFib rate and stable Vrate: if **D6** does not hold, **D7** makes the final decision on the therapy. The device diagnoses SVT if at least six out of the last ten atrial intervals are shorter than threshold $AFib_{th}$ (suggesting that the tachycardia originated in the atria) and the ventricular rhythm is stable, i.e., the last ten ventricular intervals have variance below parameter stb. Otherwise, VT is diagnosed and therapy is delivered.

We reiterate that discriminators **D1**–**D7**, or slight variations thereof, are found in other ICD manufacturers' algorithms. Thus, our method apply to other devices as well.

2.2 Generation of Synthetic EGMs

Discrimination algorithms utilize two elements of EGMs for feature extraction: timing of atrial and ventricular events, and morphology of far-field ventricular events. Jiang et al. [12] have developed a heart model that can generate realistic synthetic EGMs that can be used to evaluate the safety and efficacy of discrimination algorithms. The timing of heart events is generated by a timed-automata model

of the electrical conduction system of the heart [13], which allows simulating cardiac dynamics under different parameter settings. The morphology of far-field ventricular events is sampled from a large database of real patient EGM records [1]. EGM signals are synthesized by overlaying the sampled EGM morphology templates on the sequence of cardiac events generated by the timed model.

Finally, different arrhythmias are reproduced by running the model on different parameters. For example, a generic SVT arrhythmia has ventricular intervals in the range of $[280, 530]$ ms; then, EGMs for a specific SVT arrhythmia are synthesized by uniformly sampling parameters from a sub-interval of this range.

Jiang et al. generated synthetic EGMs for the 19 arrhythmias of the RIGHT clinical trial [3], a trial designed to evaluate the BSc discrimination algorithm. The validity and faithfulness of these EGMs were validated by electrophysiologists. In this paper, we therefore use the same synthetic EGM dataset.

3 ICD ATTACK MODEL

We present a model-based approach to synthesizing reprogramming attacks on ICDs, where the attacks are not evaluated on the actual physical device but on a model of the device. The BSc algorithm model that we consider faithfully reproduces the behavior of the real device in terms of arrhythmia discrimination and therapy, as discussed in Section 2. In an ICD reprogramming attack, the attacker manipulates the parameter values of the victim's ICD to cause harm while going undetected. These two objectives are respectively called *effectiveness* and *stealthiness*, and are formalized in Section 4.

An attack is effective when it compromises the decision of the discrimination algorithm to introduce false negatives (FN), i.e., prevent a required therapy during VF/VT, or false positives (FP), i.e., introduce inappropriate therapy during SVT. These are called *FN attacks* and *FP attacks*, respectively. Our attack model is concerned with inducing at least one compromised decision, which suffices to cause adverse or even fatal effects: depriving a patient of treatment for VF can lead to sudden cardiac death, while inappropriate shocks can result in injurious cardiac tissue remodeling and cause significant psychological distress [12]. Note that the unaltered parameters can themselves have a low rate of inappropriate or missed therapy [23], which is, however, negligible compared to that of malicious parameters.

In our attack model, stealthiness depends on the clinician's ability to detect the attack. We are therefore interested in finding malicious parameters that exhibit small deviations from the clinical settings of the victim's ICD, changes that are difficult for the clinician to notice or that can be mistaken for human error. In fact, deviations from the default settings are the norm, as ICD parameters are adjusted by the clinician on a regular basis during follow-up visits. The victim has no means to monitor their ICD parameters outside of clinic, and upon experiencing unusual activity by the ICD, s/he will likely seek medical aid rather than suspect a cyber-attack. Hence, the in-clinic setting is of primary interest. Moreover, the victim will likely be unable detect the attacker on the spot, because an ICD attack does not typically induce adverse outcomes immediately but with some delay, depending on the frequency that the victim experiences arrhythmia and the probability that the reprogrammed parameters mis-classify that arrhythmia.

Reprogramming attacks are synthesized in an offline *training phase*, which allows the attacker to obtain malicious parameters with optimal effectiveness and stealthiness with respect to a set of training EGM signals. Such parameters are derived by solving a multi-objective optimization problem over a set of logical constraints describing the behavior of the discrimination algorithm over the training signals. We solve the problem using SMT-based techniques that are guaranteed to find optimal parameter values along the effectiveness-stealthiness Pareto front (see Sections 4 and 5). This is a computationally intensive task, better performed offline.

To evaluate how the attack generalizes with previously unseen signals, which mimic the unknown EGM of the victim, we *validate* the parameters synthesized in the training phase *using a disjoint test dataset*.

We assume that the attacker has no knowledge of the victim's ICD parameters, and thus their best strategy is to train the attack by assuming that the default parameters correspond to the nominal values (Table 1). Therefore, the stealthiness computed under nominal parameters might deviate from that under the actual victim's parameters. This discrepancy, however, is limited by the fact that condition- or patient-specific parameters tend to be close to the nominal ones, which are generally considered safe [17].

Due to limited availability of real patient signals, we choose to work with *synthetic EGMs*, even though our approach supports both. The EGM generation method of Section 2.2 gives the attacker a crucial advantage. If the attacker knows the victim's specific arrhythmia, then they can generate a training dataset of synthetic signals for that arrhythmia. We call such attacks *condition-specific*. We will also consider more generic datasets that include signals for different arrhythmias (*condition-agnostic attacks*), suitable when the attacker has little knowledge of the victim's condition. Our method, however, supports any choice of training EGMs, e.g., EGMs reproducing a desired level of inter-patient variability.

Open-loop (i.e., fixed) EGM signals are adequate for our purposes because successful attacks do not affect the signals in a significant way: when the attack prevents a required shock for an EGM with arrhythmia, the arrhythmia persists and the EGM is unaffected; when it introduces inappropriate shocks during an already normal rhythm, the EGM is also unaffected, as shocks restore the electrical activity of the heart to normal.

Real-world attacks. We discuss additional assumptions that would make our model-based method suitable to real-world attacks using radio signals via software-defined radios. First, the attacker must know the ICD model of the victim, so that it can select the appropriate discrimination algorithm to use in the training phase. The ICD model can be revealed by sending discovery signals to the device (as shown in [9]), or from the victim's medical records. To change the parameter settings, the attacker also must know the communication protocol of the ICD, which can be reverse-engineered as also shown in [9]. In our work, we focus on a single discrimination algorithm. Due, however, to the universality of discriminators, our approach can be easily adapted to other algorithms.

Second, the radio antenna transmitting the attack signals must be physically close to the victim. To do so, the attacker could approach

the victim (e.g., in a crowded space) or hide the transmitter and leave it running in proximity of the victim.

Countermeasures. A possible countermeasure is to store a copy of the physician-programmed values both in a hospital database and in a secure memory location on the device. The currently programmed values are regularly checked against the stored, golden values. Any discrepancy leads to an alarm. A more general countermeasure is to secure device access through an authentication token (smart card, NFC device, etc.) that shares a secret key with the device [27]. Finally, a simple attack detection method would be to alert the patient (e.g., with a beep) whenever a communication happens with the device [9].

4 ICD ATTACK SYNTHESIS PROBLEM

We formalize the problem of synthesizing ICD reprogramming attacks as a multi-objective optimization problem that seeks to find ICD parameters optimizing two contrasting objectives: *effectiveness*, in terms of maximizing therapy disruption; and *stealthiness*, in terms of making the attack difficult to detect.

For a set X, let X^* denote the Kleene closure of X. For a sequence $\mathbf{x} \in X^*$, $|\mathbf{x}|$ denotes its length and, for $k = 0, \ldots, |\sigma| - 1$, $\mathbf{x}[k] \in X$ denotes its $k+1$-st element. Let $\mathrm{Sig} \subseteq \mathbb{R}^{m*}$ be the set of m-dimensional, finite-length, discrete-time *cardiac signals*. For signal $\mathbf{s} \in \mathrm{Sig}$, $\mathbf{s}[k]$ gives the values of the atrial, ventricular and shock EGMs ($m = 3$) at the $k+1$-st sample of the signal.

ICD parameters are tuples $\mathbf{p} = (p_1, \ldots, p_n)$, where $p_i \in P_i$ is the value of the i-th parameter, and P_i is its finite domain. For each parameter, there is a finite set of programmable values; see Table 1. We denote with $\mathbb{P} = \bigtimes_{i=1}^{n} P_i$ the set of possible parameterizations.

A *discrimination algorithm* is a function $d : \mathbb{P} \to (\mathrm{Sig} \to \mathbb{B}^*)$, where $\mathbb{B}^*$ is the set of Boolean sequences. For parameters $\mathbf{p} \in \mathbb{P}$ and signal $\mathbf{s} \in \mathrm{Sig}$, $d(\mathbf{p})(\mathbf{s})$ is a Boolean-valued sequence, called a *therapy signal*, with as many elements as the number of cardiac cycles in $\mathbf{s}$. For $k < |d(\mathbf{p})(\mathbf{s})|$, $d(\mathbf{p})(\mathbf{s})[k]$ is true if the ICD decides to deliver therapy at the k-th cycle, and is false otherwise. Recall from Section 2 that the discrimination algorithm is only invoked at each ventricular event (corresponding to the end of a cardiac cycle), and thus intermediate time points between two ventricular events are not relevant. Note that we do not consider ICD parameters that affect the detection of ventricular events, meaning that the length of a therapy signal $d(\mathbf{p})(\mathbf{s})$ is constant for any $\mathbf{p} \in \mathbb{P}$.

Effectiveness. Let $\mathbf{p}^* = (p_1^*, \ldots, p_n^*) \in \mathbb{P}$ be the default parameters of ICD algorithm d, and $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{P}$ be particular attack parameters. The *effectiveness* of $\mathbf{p}$ is evaluated over a (training or test) dataset of signals $S \subseteq \mathrm{Sig}$, and is denoted by $f_e(\mathbf{p}, S)$.

Per our description of the attack model (Section 3), we define effectiveness as the proportion of signals in S where an FN attack (preventing required therapy) or an FP attack (delivering inappropriate therapy) occurs:

$$f_e(\mathbf{p}, S) = \frac{1}{|S|} \cdot \sum_{\mathbf{s} \in S} I\left(R_{th}(d, \mathbf{p}, \mathbf{s}) \neq R_{th}(d, \mathbf{p}^*, \mathbf{s})\right), \tag{1}$$

where I is the indicator function and $R_{th}(d, \mathbf{p}, \mathbf{s})$ is the *therapy reachability* value, describing whether or not therapy is administered at any point in signal $\mathbf{s}$ for parameters $\mathbf{p}$:

$$R_{th}(d, \mathbf{p}, \mathbf{s}) = \bigvee_{k=0}^{|d(\mathbf{p})(\mathbf{s})|-1} d(\mathbf{p})(\mathbf{s})[k]. \tag{2}$$

Therapy reachability is motivated by the fact that we employ synthetic EGMs reflecting a number of arrhythmogenic (VF/VT-like) and non-arrhythmogenic (SVT-like) situations, with the former requiring therapy and the latter requiring that such therapy not be delivered. We deem an attack successful on an EGM if the EGM is mis-classified in this manner. In practice, FN attacks during VF or VT can be fatal (these arrhythmias can lead to sudden cardiac death [12]) and thus, are more dangerous than FP attacks during SVT. Nevertheless, in our definition of effectiveness, we do not need to assign different weights to these two attacks because the datasets that we consider contain either VT/VF-like EGMs (subject to FN attacks only) or SVT-like EGMs (subject to FP attacks only).

Stealthiness. An attack is considered stealthy when the deviation between the reprogrammed $\mathbf{p}$ and the default parameters $\mathbf{p}^*$ is small. To capture this deviation, we introduce a measure of *parameter distance* to minimize for optimal stealthiness. Since ICD parameters can be only programmed to a finite set of values, we quantify the distance between two parameters as the number of programmable values separating them.

For $i = 1, \ldots, n$, let $P_i = \{p_1^i, \ldots, p_{n_i}^i\}$ be the programmable values for the i-th ICD parameters. W.l.o.g. assume that the values $p_1^i, \ldots, p_{n_i}^i$ are ordered. Rewrite the default parameters as $\mathbf{p}^* = \left(p_{I_1^*}^1, \ldots, p_{I_n^*}^n\right)$ and the attack parameters as $\mathbf{p} = \left(p_{I_1}^1, \ldots, p_{I_n}^n\right)$, i.e., I_i^* is the index of the element of P_i corresponding to the value of the i-th parameter in $\mathbf{p}^*$. I_i is defined in an analogous way for $\mathbf{p}$. Then, the distance between $\mathbf{p}$ and $\mathbf{p}^*$ is defined as:

$$f_s(\mathbf{p}) = \max_{i=1,\ldots,n} \left|I_i - I_i^*\right|. \tag{3}$$

We explain (3) with an example. Suppose that the i-th parameter is VTdur from Table 1, which can be programmed to any value in the set $P_i = \{1, 1.5, \ldots, 5, 6, \ldots, 15, 20, \ldots, 30\}$. We set $\mathbf{p}^*$ using the nominal value of 2.5 for VTdur, which corresponds to the 4-th element of P_i. Hence, $I_i^* = 4$. Consider attack parameters $\mathbf{p}$ where VTdur is set to 4.5, i.e., the 8-th value of P_i ($I_i = 8$). The distance relative to VTdur is the number of programmable values separating the default setting (2.5) and the attack (4.5), which is given by $\left|I_i - I_i^*\right| = |8 - 4| = 4$. Indeed, the two are separated by four programmable values (3, 3.5, 4, 4.5). The overall distance is the maximum separation over all ICD parameters.

This notion of distance assumes that parameters admit a linear order, which is the case for all numeric parameters of the BSc ICD algorithm. For categorical parameters, one could either assign the same distance to all categories different from the nominal one, or repeat the synthesis for each category.

Optimal stealthy attacks. We formulate the synthesis of stealthy reprogramming attacks as a multi-objective optimization problem where we seek to optimize effectiveness and stealthiness (maximize f_e and minimize f_s) of the parameters w.r.t. a set of training EGMs. Multi-objective optimization allows one to derive the optimal trade-off between multiple, possibly contrasting objectives, implying that we do not need to assume any weight or priority ordering for the

objectives. The result of this analysis is a so-called *Pareto front*, i.e., a set of non-dominated points in the objective space of possible effectiveness and parameter distance values.

Problem 1 (Reprogramming attack synthesis). *For effectiveness objective f_e and distance objective f_s, training set of signals $S \subseteq \mathsf{Sig}$, find the set $\mathbf{P}$ of Pareto-optimal parameters, i.e.:*

$$\mathbf{P} = \{\mathbf{p} \in \mathbb{P} \mid \nexists \mathbf{p}' \in \mathbb{P}.\ (f_e(\mathbf{p}', S) > f_e(\mathbf{p}, S) \wedge f_s(\mathbf{p}') \leq f_s(\mathbf{p})) \vee (f_e(\mathbf{p}', S) \geq f_e(\mathbf{p}, S) \wedge f_s(\mathbf{p}') < f_s(\mathbf{p}))\}. \quad (4)$$

Consider for instance two parameters $\mathbf{p}_1$ and $\mathbf{p}_2$, such that for some S, $f_e(\mathbf{p}_1, S) = 0.5$, $f_e(\mathbf{p}_2, S) = 0.7$, $f_s(\mathbf{p}_1) = 5$, and $f_s(\mathbf{p}_2) = 5$. $\mathbf{p}_2$ has better effectiveness than $\mathbf{p}_1$ and same distance, so $\mathbf{p}_2$ dominates $\mathbf{p}_1$, meaning that $\mathbf{p}_1$ cannot be in the Pareto-optimal front. $\mathbf{p}_2$ is in the Pareto-optimal front if there are no parameters that dominate it.

To quantify how well the attacks generalize to unseen data, we introduce a *validation score* defined as the average deviation of the attack effectiveness between training and test data.

Given a training set S, a set of Pareto-optimal parameters $\mathbf{P}$ with respect to S, and a test set S', we define the validation score as: $\sum_{\mathbf{p} \in \mathbf{P}} (f_e(\mathbf{p}, S') - f_e(\mathbf{p}, S))/|\mathbf{P}|$. Positive values indicate that the parameters $\mathbf{P}$ have better performance with unseen data than with training data, whereas negative values imply the opposite. Note that the validation score need not consider stealthiness because this is independent of the signals.

5 OMT ENCODING

In this section, we present a solution method for the reprogramming attack synthesis problem (Problem 1). We formalize the behavior of the BSc discrimination algorithm in the framework of Satisfiability Modulo Theories (SMT) [2], within which the ICD algorithm is described as a set of first-order formulas over some (decidable) background theory. Parameters are represented as uninterpreted constants in the SMT encoding, and parameter synthesis corresponds to finding a satisfiable assignment to those constants, i.e., a so-called model. In particular, we formulate Problem 1 as an Optimization Modulo Theories (OMT) problem, i.e., an extension of SMT for finding models that optimize given objectives [5].

The synthesis of optimal reprogramming attacks is difficult, as it entails solving a combinatorial multi-objective optimization problem (non-continuous, non-convex) constrained by the behavior of the discrimination algorithm, which cannot be captured by simple (in)equality constraints. Therefore, classical optimization methods such as linear or convex programming are not suitable, while non-linear optimization techniques such as genetic algorithms would provide only sub-optimal solutions. In contrast, OMT is uniquely suited to solve this problem, as the ICD algorithm can be adequately encoded as SMT constraints and the parameters found by OMT are guaranteed to be optimal.

Since we are interested in analyzing the behavior of the algorithm offline over a fixed set of EGM signals, we can pre-compute for each signal the non-linear operations underlying some of the discriminators, such as the Rhythm Match score. This allows us to encode the problem over the decidable theory of quantifier-free linear integer real arithmetic (SMT QF_LIRA). Importantly, we pre-compute only the operations that are not affected by the ICD parameters, meaning that our encoding accounts for all possible behaviors induced by different parametrizations.

W.l.o.g. assume that the training dataset S is indexed. The behavior of the algorithm for the j-th signal is described by a sequence of symbolic states $s_{j,0}, \ldots s_{j,N_j}$, one for each cardiac cycle, where N_j is the number of cycles in the j-th signal. The evolution of the discrimination algorithm over the training signals is characterized by the following formula (inspired by bounded model checking [4]):

$$\mathsf{paramRanges} \wedge \bigwedge_{j=1}^{|S|} \left(\mathsf{Init}(s_{j,0}) \wedge \bigwedge_{k=0}^{N_j - 1} T(k, s_{j,k}, s_{j,k+1}) \right) \quad (5)$$

where paramRanges is a predicate describing the programmable values of the ICD parameters (see Table 1); $\mathsf{Init}(s_{j,0})$ is the predicate for constraining the initial state of the algorithm, and $T(k, s_{j,k}, s_{j,k+1})$ is the transition relation determining from the current state and cardiac cycle, the admissible states of the algorithm at the next cycle. In our case, T is deterministic, i.e., for fixed $s_{j,k}$ and k, there exists only one state $s_{j,k+1}$ such that $T(k, s_{j,k}, s_{j,k+1})$ holds. The transition relation describes the behavior of the discrimination algorithm presented in Section 2, see [19] for its full SMT QF_LIRA encoding. In (5), states $s_{j,k}$ are implicitly existentially quantified.

In the BSc algorithm, the state $s_{j,k}$ for the j-th signal and k-th cardiac cycle is represented by

$$s_{j,k} \stackrel{\text{def}}{=} (\mathsf{VFd}_{\mathsf{j,k}}, \mathsf{VTd}_{\mathsf{j,k}}, \mathsf{tVF}_{\mathsf{j,k}}, \mathsf{tVT}_{\mathsf{j,k}}) \in \mathbb{B} \times \mathbb{B} \times \mathbb{Z}^{\geq} \times \mathbb{Z}^{\geq},$$

where $\mathsf{VFd}_{\mathsf{j,k}}$ and $\mathsf{VTd}_{\mathsf{j,k}}$ tell whether or not the algorithm is, respectively, in the VF duration and VT duration mode, with $\mathsf{tVF}_{\mathsf{j,k}}, \mathsf{tVT}_{\mathsf{j,k}}$ being the clocks that keep track of time spent in the respective modes. The clocks are digital ($\in \mathbb{Z}^{\geq}$) and measure the time in milliseconds.

For any signal j, the initial state of the algorithm is given by the following Init predicate

$$\mathsf{Init}(s_{j,0}) = \neg\mathsf{VFd}_{\mathsf{j,k}} \ \wedge\ \neg\mathsf{VTd}_{\mathsf{j,k}} \ \wedge\ \mathsf{tVF}_{\mathsf{j,k}} = 0 \ \wedge\ \mathsf{tVT}_{\mathsf{j,k}} = 0,$$

indicating that the algorithm is in neither duration mode and that the clocks are set to zeros.

The value of the therapy signal is not part of the state but is encoded by the state predicate $\mathsf{Th}_{\mathsf{j,k}}$ (see [19] for its SMT encoding), describing whether or not therapy is given at the k-th cycle in the j-th signal. Thus, for signal $\mathbf{s}_j$ and fixed parameters $\mathbf{p}$, $\mathsf{Th}_{\mathsf{j,k}}$ is a symbolic representation of $d(\mathbf{p})(\mathbf{s}_j)[k]$.

An example path of the BSc algorithm encoding is given below. $s \xrightarrow{k} s'$ denotes a transition between states s and s' at the k-th cardiac cycle, i.e., such that $T(k, s, s')$ holds.

$$\ldots (\bot, \bot, 0, 0) \xrightarrow{13} (\bot, \top, 0, 0) \xrightarrow{14} (\bot, \top, 0, 309) \ldots \xrightarrow{25} (\bot, \top, 0, 2317) \xrightarrow{26} (\bot, \bot, 0, 0)$$

The transition at $k = 13$ marks the start of VT duration (VTd passes from $\bot$ to $\top$). The algorithm stays in VT duration for 13 more cardiac cycles during which the episode persists, until it reaches the end of the timer: at the start of the 26-th cycle the VT clock evaluates to $\mathsf{tVF} = 2317$, but at the end of the cycle, the clock would exceed the VT duration parameter which, in this example, is set to the nominal value VTdur = 2500 milliseconds.[1] At this point,

[1] To have a concrete path, we fixed an interpretation for the ICD parameters.

it delivers therapy and resets the VT clock, going back to state $(\perp, \perp, 0, 0)$.

Effectiveness and stealthiness encoding. We show how to encode effectiveness maximization as a MaxSMT problem. For each signal j, we define the following soft constraint:

$$\mathsf{effective}_j = \left(Rth_j^* = \neg \bigvee_{k=0}^{N_j-1} \mathsf{Th}_{j,k} \right), \tag{6}$$

where Rth_j^* is the therapy reachability value (telling whether or not therapy is administered at any point) for signal j and default parameters. Rth_j^* can be pre-computed for efficiency. $\bigvee_{k=0}^{N_j-1} \mathsf{Th}_k$ is the therapy reachability for the attack parameters, and thus, $\mathsf{effective}_j$ is true if the attack disrupts the default therapy. Note that maximizing the effectiveness f_e defined in (1) is equivalent to maximizing the number of $\mathsf{effective}_j$ constraints satisfied. Hence the MaxSMT formulation.

Parameter distance is encoded as an uninterpreted integer constant to minimize, dist. Recall that we measure distance between two parameters as the number of programmable values separating them, and that in BSc ICDs, any parameter has a finite number of numeric programmable values. It follows that dist has a finite domain, i.e. $\mathsf{dist} \in \{0, 1, \ldots, \mathsf{dist}_{\max}\}$. [2]

We encode dist in an implicit way, that is, we do not add constraints for (3) but we restrict the parameter domains conditioned on the distance value as follows:

$$\bigwedge_{s=0}^{\mathsf{dist}_{\max}} \mathsf{dist} \leq s \Rightarrow \left(\bigwedge_{i=1}^{n} p_L^i \leq \mathsf{P}_i \leq p_U^i \right), \tag{7}$$

where P_i is the SMT encoding of the i-th parameter, $L = \max\left\{I_i^* - s, 1\right\}$, and $U = \min\left\{I_i^* + s, n_i\right\}$. In other words, p_L^i is the s-th closest left neighbor of P_i's default value, p_U^i is its s-th closest right neighbor. Therefore, $p_L^i \leq \mathsf{P}_i \leq p_U^i$ restricts the domain of P_i to values with distance at most s, from which the correctness of (7) follows. Below we show part of the concrete instantiation of (7) relative to VTdur:

$$(\mathsf{dist} \leq 0 \Rightarrow (\ldots \wedge 2500 \leq \mathsf{VTdur} \leq 2500 \wedge \ldots)) \wedge$$
$$(\mathsf{dist} \leq 1 \Rightarrow (\ldots \wedge 2000 \leq \mathsf{VTdur} \leq 3000 \wedge \ldots)) \wedge$$
$$(\mathsf{dist} \leq 2 \Rightarrow (\ldots \wedge 1500 \leq \mathsf{VTdur} \leq 3500 \wedge \ldots)) \wedge \ldots$$

Synthesis of Pareto-optimal attacks. The OMT solver returns the set of Pareto-optimal objective values, i.e., the set of all (s, e) pairs such that $s = f_s(\mathbf{p})$ and $e = f_e(\mathbf{p}, S)$ for some Pareto-optimal parameter $\mathbf{p} \in \mathbf{P}$ w.r.t. training set S. For each (s, e), the solver computes a witness $\mathbf{p}'$ yielding that Pareto-optimal objective value. The synthesized parameters is the set of all such $\mathbf{p}'$. This implies that we synthesize a subset of $\mathbf{P}$ since the witness might not be unique, but do not exclude any (s, e) in the space of Pareto-optimal objectives.

6 RESULTS AND DISCUSSION

For the synthesis of condition-specific attacks, we employ synthetic EGMs for 19 different arrhythmias, generated as per Section 2.2, and apply our method to synthesize Pareto-optimal parameters using a training set of 100 signals for each arrhythmia. We validate the attacks with test sets of 50 signals per arrhythmia (disjoint from the training sets). Experiments suggested that 100 training signals provide a sufficiently complete representation of the signal space, as the performance with unseen test signals stays relatively constant for any training set size larger than 40. All EGMs have a duration of 30 seconds, but their lengths – given by the number of cardiac cycles – vary depending on the ventricular interval duration.

Arrhythmia	Effectiveness	Distance	|P|	V. score	Time	$\lvert\sigma\rvert$
1 SVT	0.338 [0.02,0.87]	15.5 [13,18]	6	-0.0217	776	57.59
2 SVT	0.397 [0.04,0.92]	15.5 [13,18]	6	-0.0433	459	58.19
3 VT	0.497 [0.01,1.00]	6.583 [1,13]	12	-0.0033	4776	90.48
4 VT	0.561 [0.01,1.00]	9.583 [4,16]	12	0.0025	8208	84.64
5 SVT	0.505 [0.01,1.00]	9.154 [1,17]	13	-0.0523	1894	64.3
6 SVT	0.298 [0.03,0.55]	10 [4,18]	9	0.02	455	61.03
7 VT	0.504 [0.01,1.00]	9.357 [2,16]	14	-0.0593	5270	84.36
8 SVT	0.170 [0.01,0.48]	9.5 [7,12]	6	-0.05	460	48.64
9 SVT	0 [0,0]	0 [0,0]	1	0	279	47.72
10 VT	0.565 [0.01,1.00]	7.091 [2,13]	11	-0.0518	4739	89.34
11 SVT	0.033 [0.01,0.06]	11 [10,12]	3	-0.0267	343	45.87
12 SVT	0.326 [0.01,0.75]	11.385 [3,18]	13	-0.0077	876	59.39
13 SVT	0.084 [0.01,0.20]	16 [14,18]	5	-0.036	363	50.38
14 SVT	0.067 [0.01,0.16]	15.333 [12,18]	6	-0.01	539	52.01
15 SVT	0.498 [0.01,0.92]	13.5 [11,16]	6	0.0083	374	51.23
16 VT	0.468 [0.02,0.99]	6 [1,11]	11	-0.0064	4419	89.06
17 VT	0.490 [0.05,1.00]	10.6 [6,16]	10	-0.004	2699	84.82
18 VT	0.517 [0.04,1.00]	10.7 [6,16]	10	-0.009	2489	84.45
19 VT	0.506 [0.04,1.00]	10.6 [6,16]	10	-0.02	2812	84.87

Table 2: Statistics for Pareto-optimal condition-specific attacks. Effectiveness and parameter distance are in the form $\mu[m, M]$ (mean μ, minimum m, maximum M objective function value for all solutions). |P| is the number of Pareto-optimal solutions. V. score is the validation score. Time is the runtime in seconds. $|\sigma|$ is the average length of the training signals.

We classify these 19 arrhythmias into two categories, VT and SVT, depending on whether or not the corresponding signals require ICD therapy under nominal parameters. In particular, we have 8 VT arrhythmias (subject to FN attacks) and 11 SVT arrhythmias (subject to FP attacks).

We also synthesize condition-agnostic attacks, suitable when the attacker has little knowledge of the victim's arrhythmia. We consider two attacks for generic VT and SVT arrhythmias, using training sets of 200 EGMs randomly sampled among the 8 VT-like arrhythmias and the 11 SVT arrhythmias, respectively. We validate the two attacks with disjoint test sets of 100 signals.

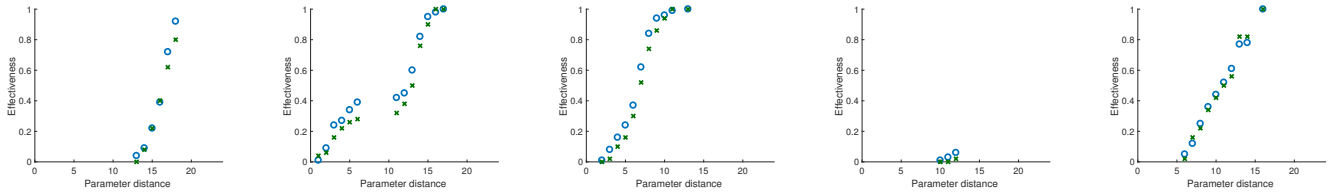
The method for synthetic EGMs was implemented in MATLAB. For parameter synthesis, we used the z3 SMT solver [5].

Condition-specific attacks. Table 2 provides statistics on the synthesized Pareto-optimal attacks. Figure 4 shows the Pareto-optimal fronts for a selection of representative arrhythmias (see [19] for the full set of plots and synthesized parameters).

The synthesized attacks attain validation scores that are either positive or very close to zero, indicating that the attacks generalize well with unseen data and, thus, would have comparable effectiveness on the unknown EGM of the victim.

As visible in Table 2, our method can derive effective FN attacks for all VT arrhythmias, since the corresponding Pareto fronts always contain a parametrization able to disrupt the therapy of all training signals (effectiveness 1), with the exception of arrhythmia 16 where the maximum effectiveness is 0.99. Not all attacks on VT

[2] $\mathsf{dist}_{\max} = \max_{i=1,\ldots,n} \max\left\{n_i - I_i^*, I_i^* - 1\right\}$, where n_i is the number of programmable values for the i-th parameter and I_i^* is the index of its default value.



(a) Cond. 2 (SVT) (b) Cond. 5 (SVT) (c) Cond. 10 (VT) (d) Cond. 11 (SVT) (e) Cond. 17 (VT)

Figure 4: Pareto fronts for a selection of condition-specific reprogramming attacks (see [19] for the full set of arrhythmias). Blue dots: Pareto front obtained with training signals. Green crosses: effectiveness of the synthesized parameters on the test signals.

arrhythmias are, however, comparably stealthy (see Figure 4). For instance, for arrhythmia 10 a parameter distance of 7 ensures that the attack is effective with half of the training signals, while for arrhythmia 17, the same effectiveness level is obtained only at a distance of 11 from the nominal parameters (worse stealthiness).

In contrast, FP attacks on SVT arrhythmias are not all equally successful. For arrhythmia 5 we can find parameters with 100% effectiveness as well as stealthy attacks that e.g. are able to affect almost 40% of the signals with a distance of only 5. For arrhythmias 2 and 15 we obtain parameters with nearly 100% effectiveness but with poor stealthiness (the minimal distance of a Pareto-optimal attack is 13 and 11, respectively). Some EGMs turned out to be difficult to attack: for arrhythmia 11 the strongest attack affects only 6% of the signals and, for arrhythmia 9, no Pareto-optimal attacks exist but the trivial one that leaves the nominal parameters unchanged.

The reason why VT arrhythmias are easier to attack is that it takes only a minor increase to the VT and VF detection thresholds (parameters VF_{th} and VT_{th}) to make the ICD mis-classify a tachyarrhythmia episode. On the other hand, VF_{th} and VT_{th} must be reprogrammed to very low values in order for the ICD to classify a slow heart rate as VT/VF and induce unnecessary therapy. This is not always possible because in SVT arrhythmias, the heart rate is often below the lowest programmable values for VF_{th} (110 BPM) and VT_{th} (90 BPM), which explains why, for instance, no attack parameters exist that can affect arrhythmia 9. We remark that these results are *provably correct* because OMT is *guaranteed* to find Pareto-optimal attack parameters, when they exist.

Besides increasing VF_{th} and VT_{th}, the attacks on VT arrhythmias synthesized by our method tend to increase the VF and VT durations (VFdur and VTdur) thus reducing the probability that the ICD classifies an episode as sustained, which is a necessary condition for delivering therapy. For instance, the most effective attack for arrhythmia 10 has VF_{th} = 250 BPM, VT_{th} = 205 BPM, VFdur = 10 s, and VTdur = 13 s, against nominal values of 200, 160, 1, and 2.5, respectively. For some VT arrhythmias, the attacks also affect the VT zone-related parameters to make discriminators D6 and D7 more likely to be satisfied, thus tricking the ICD into classifying the episode as SVT.

Figure 5 compares nominal and reprogrammed parameters over an execution of the BSc algorithm at the start of a VF episode, using an EGM from arrhythmia 10. With nominal parameters, VF duration starts after the last 8/10 ventricular intervals faster than VF (see marker 1 in Fig. 5) and ends after an interval is found below

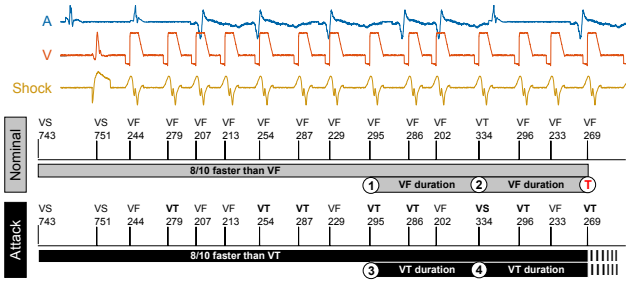
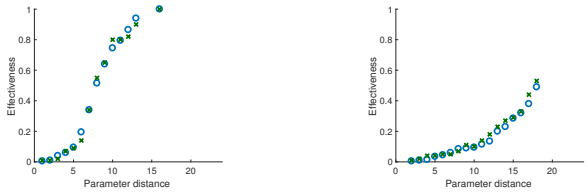


Figure 5: Execution of BSc ICD algorithm with nominal and attack parameters on atrial (A), ventricular (V), and shock EGMs from arrhythmia 10. Markers are: VF – sensed ventricular fibrillation, VT – tachycardia, and VS – normal rate. Intervals are in milliseconds. See text for a detailed explanation.

the VF threshold (see marker 2). A new VF duration can start right away, ending this time with a therapy (marker T). Here, the reprogramming attack sets VF_{th} = 240 BPM (250 ms), VF_{th} = 185 BPM (325 ms), and VTdur = 7 s. With the higher VF threshold, the attack leads to marking the VF episode as VT, triggering VT duration (marker 3). VT duration ends with one interval found below the reprogrammed VT threshold (marker 4). A new VT duration can start right away, but therapy is prevented due to the long VTdur.

Attacks on SVT arrhythmias follow the opposite strategy. All attacks tend to keep VF_{th}, VT_{th}, VFdur and VTdur to the minimum programmable values, thereby increasing the probability that slow heart rhythms are classified as sustained tachyarrhythmia. For some SVT arrhythmias the attacks also need to increase the Rhythm Match threshold, while the parameters of discriminator D7, $AFib_{th}$ and stb, appear to have little effect.

Condition-agnostic attacks. Pareto fronts for the condition-agnostic attacks on VT and SVT, hereafter referred to as VT attack and SVT attack, are shown in Figure 6. The corresponding parameters are available in Tables 22 and 23 of [19]. These attacks attain very good validation scores, comparable to the condition-specific case, suggesting that our method can generalize well also when trained with heterogeneous arrhythmias. The Pareto front for the VT attack has a similar profile to the condition-specific ones: the effectiveness is poor for parameter distance below 5, it has a sharp increase between distance 5 and 10, growing slowly after that up to reaching 100% success at distance 16. The attack strategy is the same discussed for the condition-specific case, yielding high values of VF_{th}, VT_{th}, VFdur, VTdur and stb, and low values of $NSRcor_{th}$ and $Afib_{th}$.



(a) VT. V. score: -0.0032 (b) SVT. V. score: 0.0179

Figure 6: Pareto fronts for condition-agnostic reprogramming attacks. Legend is as in Figure 4.

On the other hand, the parameters for the SVT attack reach a maximum effectiveness of 49% at distance 18, compatibly with the fact that condition-specific attacks are reasonably successful only for a subset of SVT arrhythmias. The attack strategy confirms our previous discussion, with the synthesized parameters having minimal values of VF_{th}, VT_{th}, VFdur and VTdur.

Performance and adequacy. The results of Table 2 show that synthesis for VT arrhythmias has a higher computational cost than for SVT, with runtimes ranging from 2489 to 8208 seconds against a range of 279 to 1894 seconds for SVT. The reason is that VT arrhythmias are characterized by shorter ventricular intervals, leading to more heart beats for the same EGM duration and thus, to longer signals. The path length and the number of training EGMs are indeed the main factors affecting the complexity of OMT-based synthesis.

Our approach is adequate in that the parameters found by OMT outperform those found by random search (RS). We ran RS for each arrhythmia and for the same runtime of OMT, and compared the area under the curve (AUC) of the Pareto fronts obtained with OMT and RS, with both training and test EGMs. Higher AUC values imply better performance. We remark that the parameters found by OMT are guaranteed to be Pareto-optimal with respect to training EGMs, and thus RS (or any other search method) cannot perform better on the training data. Indeed, RS yields AUC values strictly smaller than OMT for all arrhythmias but 18 and 19, for which RS and OMT produced the same Pareto fronts (see Table 24 of [19]). With test data, OMT outperforms RS on 11 arrhythmias, while the opposite happens only for three arrhythmias. These results confirm that OMT has superior performance also with unseen signals.

7 RELATED WORK

The work of Halperin and colleagues [9] was the first to show that ICDs can be accessed and reprogrammed by unauthorized users using off-the-shelf hardware. As such, they demonstrate the physical feasibility of the attacks that we derive systematically in this work. Other attack examples from the cardiac domain include [15] and [7].

Our work leverages [12] for the generation of synthetic EGMs and the modeling of the ICD algorithm, but tackles the fundamentally different problem of designing stealthy attacks on ICDs, and uses formal (SMT-based) methods for solving it. The work in [16] synthesizes pacemaker parameters to ensure a safe rhythm and maximize robustness to parameter deviations. We tackle a different class of algorithms (found in ICDs), and study the problem of compromising device operation, as opposed to making it robust to parameter deviations.

Our work is complementary to methods for attack detection and identification in cyber-physical systems [10, 20, 26], state estimation from attack-prone sensor measurements [18, 24], and spoofing attack synthesis on general control systems [11].

8 CONCLUSIONS

We presented the first framework for systematically synthesizing reprogramming attacks on ICDs designed to maximize therapy disruption while minimizing detection. Such attacks can therefore be tailored to the victim's physiology and they readily generalize to unseen signals. This makes our approach suitable for real-world attacks.

For future work, we plan to evaluate synthesized attacks on a real ICD device, building on the hardware testbed for cardiac pacemakers of [14]. We will also investigate making ICD discrimination algorithms more resilient to such attacks.

REFERENCES

[1] Ann Arbor Electrogram Libraries. 2018. (2018). http://electrogram.com/
[2] C. W. Barrett, R. Sebastiani, S. A. Seshia, and C. Tinelli. 2009. Satisfiability Modulo Theories. *Handbook of satisfiability* 185 (2009), 825–885.
[3] R. D. Berger, D. R. Lerew, J. M. Smith, C. Pulling, and M. R. Gold. 2006. The Rhythm ID Going Head to Head Trial (RIGHT): design of a randomized trial comparing competitive rhythm discrimination algorithms in implantable cardioverter defibrillators. *Journal of cardiovascular electrophysiology* 17, 7 (2006), 749–753.
[4] A. Biere, A. Cimatti, E. Clarke, and Y. Zhu. 1999. Symbolic model checking without BDDs. In *Tools and Algorithms for the Construction and Analysis of Systems (LNCS)*, Vol. 1579. 193–207.
[5] N. Bjørner, A. D. Phan, and L. Fleckenstein. 2015. νZ-An Optimizing SMT Solver. In *Tools and Algorithms for the Construction and Analysis of Systems (LNCS)*, Vol. 15. 194–199.
[6] Boston Scientific Corporation. 2017. Implantable Cardioverter Defibrillator, reference guide (part number: 359407-003). (2017).
[7] S. Eberz, N. Paoletti, M. Roeschlin, M. Kwiatkowska, I. Martinovic, and A. Patanè. 2017. Broken hearted: How to attack ECG biometrics. In *Network and Distributed System Security Symposium (NDSS 2017)*.
[8] Food and Drug Administration. 2017. Implantable Cardiac Pacemakers by Abbott: Safety Communication. (2017). https://www.fda.gov/safety/medwatch/safetyinformation/safetyalertsforhumanmedicalproducts/ucm573854.htm
[9] D. Halperin, T. S. Heydt-Benjamin, B. Ransford, S. S. Clark, B. Defend, W. Morgan, K. Fu, T. Kohno, and W. H. Maisel. 2008. Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. In *IEEE Security and Privacy Symposium*. 129–142.
[10] X. Hei, X. Du, S. Lin, I. Lee, and O. Sokolsky. 2015. Patient infusion pattern based access control schemes for wireless insulin pump system. *IEEE Transactions on Parallel and Distributed Systems* 26, 11 (2015), 3108–3121.
[11] O. Inverso, A. Bemporad, and M. Tribastone. 2018. SAT-based synthesis of spoofing attacks in cyber-physical control systems. In *9th ACM/IEEE International Conference on Cyber-Physical Systems*. 1–9.
[12] Z. Jiang, H. Abbas, K. J. Jang, M. Beccani, J. Liang, S. Dixit, and R. Mangharam. 2016. In-silico pre-clinical trials for implantable cardioverter defibrillators. In *38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)*. IEEE, 169–172.
[13] Z. Jiang, M. Pajic, and R. Mangharam. 2012. Cyber–physical modeling of implantable cardiac medical devices. *Proc. IEEE* 100, 1 (2012), 122–137.
[14] Z. Jiang, S. Radhakrishnan, V. Sampath, S. Sarode, and R. Mangharam. 2014. Heart-on-a-Chip: a closed-loop testing platform for implantable pacemakers. (2014).
[15] D. F. Kune, J. Backes, S.S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *IEEE Security and Privacy*. 145–159.
[16] M. Kwiatkowska, A. Mereacre, N. Paoletti, and A. Patanè. 2015. Synthesising robust and optimal parameters for cardiac pacemakers using symbolic and evolutionary computation techniques. In *Hybrid Systems and Biology (LNCS)*, Vol. 9271. 119–140.
[17] A. J. Moss et al. 2012. Reduction in inappropriate therapy and mortality through ICD programming. *New England Journal of Medicine* 367, 24 (2012), 2275–2283.

[18] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas. 2014. Robustness of attack-resilient state estimators. In *5th International Conference on Cyber-Physical Systems*. 163–174.
[19] N. Paoletti, Z. Jiang, M. A. Islam, H. Abbas, R. Mangharam, S. Lin, Z. Gruber, and S. A. Smolka. 2018. Synthesizing Stealthy Reprogramming Attacks on Cardiac Devices. *CoRR* abs/1810.03808 (2018).
[20] F. Pasqualetti, F. Dörfler, and F. Bullo. 2013. Attack detection and identification in cyber-physical systems. *IEEE Trans. Automat. Control* 58, 11 (2013), 2715–2729.
[21] A. Peterson. 2013. Yes, terrorists could have hacked Dick Cheney's heart. *Washington Post* (2013).
[22] B. Rios and J. Butts. 2018. Understanding and Exploiting Implanted Medical Devices. Black Hat USA conference. (2018).
[23] Sedláček et al. 2015. The effect of ICD programming on inappropriate and appropriate ICD therapies in ischemic and nonischemic cardiomyopathy: the MADIT-RIT trial. *Journal of cardiovascular electrophysiology* 26, 4 (2015), 424–433.
[24] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada. 2017. Secure State Estimation for Cyber-Physical Systems Under Sensor Attacks: A Satisfiability Modulo Theory Approach. *IEEE Trans. Automat. Control* 62, 10 (2017), 4917–4932.
[25] I. Singer. 2001. *Interventional electrophysiology*.
[26] A. Tiwari, B. Dutertre, Jovanović D., T. de Candia, P. D. Lincoln, J. Rushby, D. Sadigh, and S. Seshia. 2014. Safety envelope for security. In *3rd international conference on High confidence networked systems*. ACM, 85–94.
[27] F. Xu, Z. Qin, C. C. Tan, B. Wang, and L. Qun. 2011. IMDGuard: Securing implantable medical devices with the external wearable guardian. In *IEEE Infocom*. 1862–1870.
[28] N. Zanker, D. Schuster, J. Gilkerson, and K. Stein. 2016. Tachycardia detection in ICDs by Boston Scientific. *Herzschrittmachertherapie+ Elektrophysiologie* 27, 3 (2016), 186–192.