IEEE Open Journal of
Circuits and Systems

Received 28 June 2022; revised 3 September 2022; accepted 19 September 2022. Date of publication 28 September 2022;
date of current version 17 October 2022.

Digital Object Identifier 10.1109/OJCAS.2022.3210163

RF Analog Hardware Trojan Detection Through Electromagnetic Side-Channel

JOHN KAN, YUYI SHEN (Graduate Student Member, IEEE),
JIACHEN XU (Graduate Student Member, IEEE), ETHAN CHEN, JIMMY ZHU (Fellow, IEEE),
AND VANESSA CHEN (Member, IEEE)

Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

This article was recommended by Associate Editor M. Johnston.

CORRESPONDING AUTHOR: J. KAN (e-mail: johnkan@andrew.cmu.edu)

This work was supported in part by the National Science Foundation under Grant 1952907, Grant 1953801, and Grant 2028893;
and in part by the Data Storage Systems Center (DSSC) at Carnegie Mellon University.

ABSTRACT With the advent of globalization, hardware trojans provide an ever-present threat to the security of devices. Much of the research to date has centered around documenting and providing detection methods for digital trojans. Few, however, have explored the space of trojans in the RF/analog front end. Two hardware trojans, an analytical analysis of the trojan impacts on two different types of amplifiers, and an unsupervised ML detection method for edge IOT applications using magnetic tunnel junction sensors for side-channel monitoring are explored. A classification autoencoder for anomaly detection is presented with an accuracy of greater than 90% with both single tone and BLE data is presented.

INDEX TERMS Hardware security, RF/Analog, magnetic tunnel junction sensor, classification autoencoder, FPGA.

I. INTRODUCTION

HARDWARE trojans and their detection methods remain a subject of study within both academia and industry. While increased globalization opened the door for greater cooperation and development, it also opened the way for nefarious actors to impact hardware circuitry development at multiple areas of the development and manufacturing process [1], [2], [3], [4], [5]. Responding to the call from various governments and private entities, a number of published sources within the current literature present much-needed solutions not only to address the development of hardware trojans, but also their detection methods. Much of the cataloged trojans and their detection methods remained in the digital domain, that is, the trojans themselves afflicted digital logic circuits. Recently, however, the U.S. government has continued to put out a call for continued research in securing the RF/Analog domain. While previous literature addresses a number of areas within the analog domain, study into the development and impact of trojans in the RF/analog domain is needed to provide trust for these devices. To address some perceived gaps in the research, this paper presents novel trojans in the RF/Analog domain and an electromagnetic (EM) side channel hardware trojan detection method for implementation at the edge. In particular, this paper presents detailed mathematical analysis of two trojans that impact in the RF/analog domain, indicates the impacts of the trojans, presents a readout circuit utilizing an EM side-channel sensing approach with spin-tunnel junction technology, provides a novel sensor with results than can be extended to the conclusions of this study, and applies a classification autoencoder with results that can utilized for real-time, semi-supervised trojan detection and classification through the power domain at the RF-front end.

II. HARDWARE TROJAN OVERVIEW

A. ANALOG AND DIGITAL TROJAN CATALOGUE

Past work on hardware trojan design can be divided into four areas: trojan trigger design, payload design, footprint optimization, and benchmarking tools [4]. For this work, a non-exhaustive list of works targeting digital and analog integrated circuits were surveyed by payload mechanism and impact in Table 1. Although a great deal of attention has been paid to the development, injection, and impact of trojans impacting the digital domain (DTs) [3], few trojans whose

TABLE 1. Hardware trojans.

Reference	Domain	Vector	Impact
ISCAS '18 [6]	Analog	IC	PF
MWSCAS '20 [7]	Analog	EC	PE, LI
VLSI '17 [8]	Analog	EC	PE, PF
SP '16 [9]	Digital	EC	PF
ICCAD '09 [10]	Digital	EC	LI

IC = Initial Conditions; EC = Extra Components; PF = Expected Performance; PE = Power Efficiency; LI = Leaked Info.

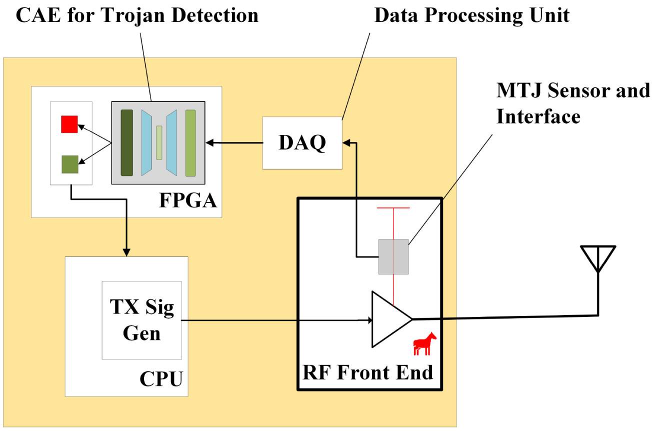

FIGURE 1. System Block Diagram. This figure indicates the goal for this project, namely, that within some sort of SOC this particular system would monitor the RF/analog front end to provide real-time security against hardware trojan threats.

payloads target RF and analog blocks appear to have been published in the literature. Furthermore, it is important to note that in the language of trojans, an analog trojan may be considered a trojan that impacts the particular domain to which it belongs, be it analog or digital. Most of the trojans labeled "DT" in the table above would be classified as "analog" trojans, that is, they operate off of an analog principle or fact, but they impact a digital integrated circuit or system. In this paper, "digital" trojan versus "analog" trojan are presented to portray the difference between what sort of integrated circuit domain the particular trojan impacts. In particular, the trojans we present are analog in nature, but they also impact analog circuits, thus making them analog trojans.

Analog trojans (ATs) that have been reported span a broad spectrum of attack vectors. The Power/Area/Architecture/Signature Transparent (PAAST) trojans presented in [6] have the capacity to severely impact system functionality while leaving no footprint on system performance when not triggered. By applying undesirable modes of operation in electronic designs as trojans, such as additional oscillatory modes in oscillator circuits, no impact is made on the system until a side-channel trigger is applied, enabling the trojans to evade detection by measurements or simulations [6]. Another type of trojan based on added circuitry hidden beneath a ground plane and transmission line within an analog RF IC, is presented in [7]. By adding an extra MOSFET to the input line of the power amplifier, the AT can impact the operation of the power amplifier, while remaining hidden.

TABLE 2. Hardware trojan detection methods.

Reference	Trojan Domain	Detection Method	Time
[2]	Digital	Optical	PFab
[11]	Digital	Delay Analysis	PFab
[12]	Digital	Delay Analysis	PFab
[13]	Digital	IC Fingerprinting via Side-channel	PFab
[14]	Analog and Digital	Current Integration	PFab

PFab = Post-Fabrication.

A detailed power analysis was required in order to detect the trojan hidden within the ASIC itself. Another type of trojan presented in this table is based in both analog and digital circuitry, such as the SOC trojan found in [8]. The line between AT and DT begins to be blurred with various systems-on-chip (SOC) that exhibit a mixture of digital and analog circuitry that can be targeted by trojans, such as in [8]. Still, within the present literature the trojans in these SOCs are themselves analog, and their manifestation and targeted issue usually lie within the analog domain.

B. A BRIEF SURVEY OF TROJAN DETECTION METHODS

Efforts to discover and document trojans are also generally accompanied by detection methods. Non-destructive methods such as golden chip comparison, logic testing, built-in self tests (BIST) and optical detection methods have been applied with varying levels of success [5]. Table 2 details various methods utilized to detect both analog and digital trojans. Most of these tests must be conducted shortly after fabrication, which is a common injection point for trojans [5]. Methods presented in the past works listed here generally do not provide real-time detection and supervision of potentially trojan-filled devices, but rather, post-fabrication discovery and mitigation.

A number of different methods were previously developed for detecting trojans embedded in digital circuitry such as optical backscattering, golden chip comparison, and built-in self tests. Side-channel analysis (SCA) utilizes signals which indirectly indicate the operation of the IC have been utilized for both those attempting to gain malicious access and those attempting to secure ICs [15]. As such, SCA is a key method for the detection of trojans. In particular, methods such as adaptive channel estimation (ACE) [16] have been presented to explore utilizing side-channel fingerprints as a way of detecting trojans at the analog front-end.

C. MACHINE LEARNING AND HARDWARE TROJAN DETECTION

Previous work experimented with various statistical and machine learning based embedded hardware trojan detection and defense methods. Beginning in 2008, statistical methods have been applied to various digital logic circuits, such as in [17], where one-class clustering analysis was utilized to discover trojans embedded in digital logic hardware. Furthermore, some methods utilized artificial neural networks (ANN)s, embedded into the ICs for which they were designed to detect various anomalies [1], [18]. These methods require efficient area and power constraints as exemplified in the ANN in [18], and hence, one key area of

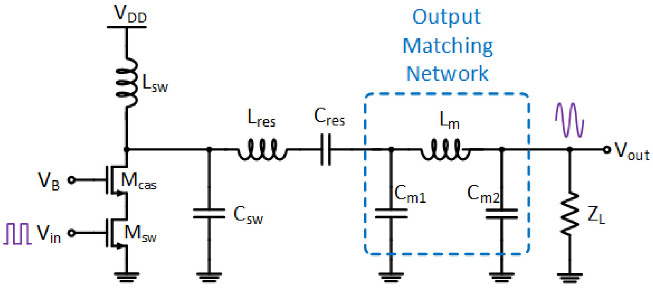

FIGURE 2. A schematic depicting a canonical Class E PA topology.

research that requires further study is the on-chip detection of trojans.

Both classical machine learning and deep learning based methods require one-class classification for trojan prevention. In one-class classification methods, it is assumed that there is a normal operation, which is documented, and the abnormal operation, which is unknown. Clustering analysis, such as that conducted in [17] primarily determines whether the principal component of that particular signature is similar to those of the pre-determined and known normal operation.

III. PROPOSED HARDWARE TROJAN ANALYSIS

The nature of analog trojans necessitates the application of design knowledge by the trojan designer in their construction and placement. The specific impact a trojan exerts on an analog system is often depends on the system's circuit-level structure and mode of operation. A description of the design and operation of the Class E PA DUT is thus provided to contextualize the following analysis of the tested analog trojans' effects. Hardware trojans were created to afflict the Class E PA because of the access to the circuit level design and layout, similar to how an untrusted foundry may also have the available information.

A. REVIEW OF POWER AMPLIFIER TOPOLOGY: CLASS E DESIGN

Class E mode amplifier circuits exist in a middle ground between conventional transconductance-based Class A-C PAs and hard-switched "digital" Class-D-style PAs. Although the transistor of a Class E amplifier is operated as a switch, the switching waveforms take on a distinctly analog appearance as the reactive output network shapes the transistor voltage and current in such a way as to minimize overlap between the two waveforms. An example of such a circuit is displayed in Fig. 2, with cascoded devices to reduce the drain-to-source voltage stress.

The ideal Class E mode is defined as exhibiting a non-sinusoidal and assymmetric transistor voltage waveform that increases from and returns to ground at the turn-off and turn-on instances respectively, with zero-slope at the turn-on instance [19]. This has the effect of fully discharging the transistor drain capacitance by the time of the turn-on instance and avoiding any current surges from the drain capacitance into the transistor itself, lowering switch transition loss. Furthermore, the ramp rate of the transistor voltage waveform at the turn-off instance is limited by said drain capacitance, decreasing losses due to the finite turn-off time of the device.

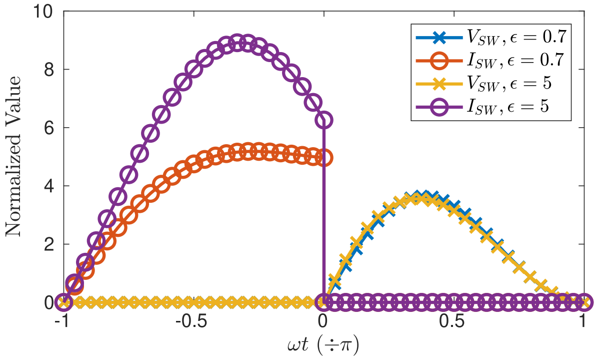

FIGURE 3. Classic Class E mode waveforms are depicted for differing values of ϵ.

Class E amplifier designs can be parameterized around the design point $\epsilon = \omega_0\sqrt{L_{SW}C_{SW}}$, where ϵ is the ratio of the operating frequency to the switching tank resonant frequency. Characteristic transistor voltage and current waveforms for differing values of ϵ are shown in Fig. 3, with voltage normalized to VDD and current to $\omega C_{SW} VDD$.

B. ANALOG HARDWARE TROJANS ATTACKER MODELS

1) GENERAL PRINCIPLES IN DEVELOPING HARDWARE TROJANS FOR THE ANALOG FRONT END

Because analog circuits are generally simpler to analyze via various golden chip and optical methods, ATs must be well-placed and difficult to find. Furthermore, they must specifically target their particular circuit topology, and are difficult to generalize. Hence, intimate knowledge of the circuit targeted for trojan injection must be held by the designer of the trojan. Similar to the motivated attacker implied in the PAAST trojans, the trojan designer will want to inject trojans that appear to be similar to circuit elements introduced by an experienced hardware designer [6]. Furthermore, by understanding the exact parameters of the circuit to which they desire to impact, they can shift these parameters to undermine the intended functionality of the circuit.

2) THREAT MODEL

The threat model for this particular trojan would be an untrusted foundry [4]. If this particular class-E integrated circuit is sent to an untrusted foundry, a malicious actor in the fabrication process could cause the IC to be compromised or to malfunction. Including a built-in self test (BIST) for trust would enable the designers and those that utilize the device to have trust that the system has been properly secured. Protection for other threat models could also be considered. As system-on-chips (SOCs) become increasingly more complicated, third-party intellectual property (3PIPs) are required to be utilized for ease and timeliness of delivery [4]. Untrusted 3PIPs could also enable the injection fo

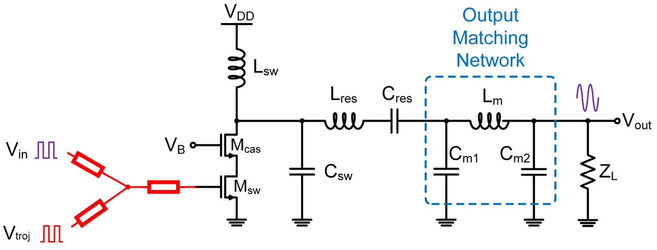

FIGURE 4. Input Signal Coupling Trojan diagram.

hardware trojans into the design process. Because of the nature of the side-channel analysis, not only can various trojans directly impacting the RF/analog domain be discovered, but also trojans caused by various types of unknown signals, perhaps generated in the digital domain such as in [8] could be revealed.

3) DEVELOPING HARDWARE TROJANS FOR A CLASS E PA

An ideal PA properly operating in Class E is held to four boundary conditions that minimize switching transition loss: (1) transistor drain voltage rises from zero at the turn-off instance, (2) transistor drain voltage returns to zero at the turn-off instance, (3) transistor drain voltage exhibits zero slope at turn-off, and (4) transistor drain voltage rises at the turn-off instance with a finite slope limited by drain capacitance. An attacker targeting the PA operation may cause any of these boundary conditions to be violated through activating a trojan within the structure of the circuit, either impacting the overall power dissipation of the PA or the power output by malforming the transistor waveforms. For example, a trojan may be used to detune the output matching network, altering the amount of fundamental frequency power that can be extracted from the switching transistor. But in determining where the trojan can most effectively be placed, the trojan designer must not only look for what can have the greatest impact, but also that which can be feasibly hidden from optical and golden chip testing. Because inductors themselves generally are optically easy to inspect (and take up the most area within the IC itself), they are difficult to target. MIM capacitors are also generally quite large. However, chip designs often have extra dummy transistors on chip in the design phase. Hence, extra transistors that are maliciously connected are conceivable trojans which may be added to a particular system.

C. SWITCH TROJAN

The switch trojan presented here is similar to [16] which operates based on some key bit that is stolen by digital trojans elsewhere in an RFIC. Instead of switching output of the power amplifier on and off, however, this trojan intermittently disconnects the transistor to disrupt the output signal, based on externally generated signals. This naturally pulses the power consumption profile of the amplifier in time with trojan activation, making for a signature that may be detected by monitoring power traces. A switched trojan utilizing extra transistors placed on an IC could cause large issues in the operation of the device, and if controlled externally, could escape notice from both optical and golden chip testing methods.

D. COUPLED INPUT SIGNAL

Due to the greater complexity found in certain RFICs especially similar to those found in [20] and [21], an attacker could couple an external signal into the PA. This would cause the additional signal to be combined with the desired RF output signal, disrupting the transmitted signal. The types of external signals an attacker could apply for this purpose include (1) signals located in other frequency bands originating from the other transmit paths within a multiband multistandard wireless transceiver, and (2) phase shifted copies of the PA input signal.

Generally, the sum of two modulated signals may be written as:

$$\Phi_n(t) = \omega_n t + \phi_n(t) \tag{1}$$

$$x(t) = r_1(t)\cos(\Phi_1(t)) + r_2(t)\cos(\Phi_2(t)) \tag{2}$$

Equating $r_2(t) = r_1(t) - K$, this equation can be re-written to be a product of cosines:

$$x(t) = 2r_1 \cos\left(\frac{\Phi_1 + \Phi_2}{2}\right)\cos\left(\frac{\Phi_1 - \Phi_2}{2}\right) - K\cos(\Phi_2) \tag{3}$$

If the amplitudes of the original PA input signal at ω_1 and coupled external signal at ω_2 are close to one another, the K term drops out, and the sum becomes a simple product term. Furthermore, should the frequency bands of the two signals lie close to one another as would be the case in (2) and often applicable in (1), the overall input signal to the PA becomes a high frequency signal located at the average of the two signals' carrier frequencies that is modulated by a much lower frequency cosine.

1) COUPLED INPUT SIGNAL: LINEAR PA ANALYSIS

Making these simplifying assumptions, the impact of this form of trojan on the power profile of a classical reduced conduction angle PA of class A - C type may be determined by applying them to the canonical PA circuit as shown in Fig. 5. Considering the behavior of such a circuit during a single period of the high frequency component while the trojan is triggered, the input voltage at the PA transistor can be written by assuming $C_1 = \cos(\frac{\Phi_1 - \Phi_2}{2})$ to be quasi-static and omitting the K term:

$$V_{in} = V_q + 2r_1 C_1 \cos\left(\frac{\Phi_1 + \Phi_2}{2}\right) \tag{4}$$

Defining $\hat{V}_{in}$ to be normalized such that $\hat{V}_{in} = \frac{V_{in} - V_{th}}{V_{sat} - V_{th}}$ where $V_{in} = V_{sat}$ saturates the transistor drain current to I_{max} as in [22], a closed form expression may be derived for the DC current consumption of a canonical Class A-C type PA by assuming a sinusoidal normalized input voltage of RF amplitude A_{in} such that $\hat{V}_{in} =$

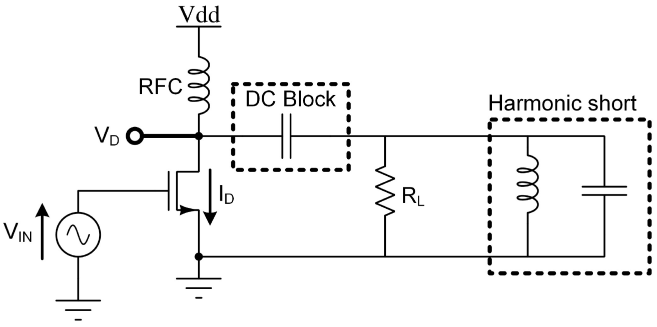

FIGURE 5. The schematic for a classical linear PA is shown, with a LC tank shorting out higher order harmonics.

$Q + (1 - Q)A_{in}\cos(\theta)$. Applying the physically justifiable cubic model for MOSFETs where $I_d(\theta) = I_{max}(3\hat{V}_{in}^2 - 2\hat{V}_{in}^3)$ prior to saturating continuously at I_{max} as V_{in} approaches and exceeds V_{sat} [22] yields the following expression for I_{DC}, where α denotes the transistor conduction angle and $\beta = 2\arccos(\frac{1}{A_{in}})$ denotes the saturation angle during which transistor current saturates at I_{max} due to V_{in} exceeding V_{sat}:

$$I_{DC} = \frac{I_{max}}{2\pi}(\beta + A + B + C + D) \tag{5}$$

$$A = (\alpha - \beta)\left[3Q^2 - 2Q^3 + \frac{3}{2}(1 - 2Q)(1 - Q)^2 A_{in}^2\right] \tag{6}$$

$$B = \left(\sin\frac{\alpha}{2} - \sin\frac{\beta}{2}\right)\left[12Q(1 - Q)^2 A_{in} - 3(1 - Q)^3 A_{in}^3\right] \tag{7}$$

$$C = \frac{3}{2}(\sin\alpha - \sin\beta)(1 - 2Q)(1 - Q)^2 A_{in}^2 \tag{8}$$

$$D = -\frac{1}{3}\left(\sin\frac{3\alpha}{2} - \sin\frac{3\beta}{2}\right)(1 - Q)^3 A_{in}^3 \tag{9}$$

Sweeping A_{in} for this expression reveals the existence of two regions of operation, one in which I_{DC} increases smoothly with a square-law-like curve, and one in which I_{DC} saturates as the input voltage sweeps the transistor voltage past V_{sat}. Thus, we consider two cases for the amplitudes of the original PA input signal and coupled external signal: (1) in which they add up to barely saturate the I_{DC} curve and (2) in which they add to drive the transistor voltage well past V_{sat}. The first scenario may be modeled by setting $A_{in} = \cos(\frac{\Phi_1 - \Phi_2}{2})$, while the second may be modeled by adding a factor of two such that $A_{in} = 2\cos(\frac{\Phi_1 - \Phi_2}{2})$. Assuming modulation is not present on either input signal and setting $Q = 0.1$ to correspond to Class AB operation and $\omega_2 = 0.999\omega_1$ where $\omega_1 = 1$ for convenience yields the curves shown in Fig. 6.

It can be seen from these curves that a low frequency pseudo-sinusoid at $\omega_1 - \omega_2$ is imposed on top of the variations in DC current that result from any pre-existing amplitude modulation within the input signals, yielding a signature capable of passing through the drain-side RF choke of a PA to be sensed via power profiling. The peak of this pseudo-sinusoid is clearly compressed to an extent determined by the total amplitudes of the PA input signal and coupled external signal, rendering the shape of the power signature in time dependent on the specific circumstances of trojan activation.

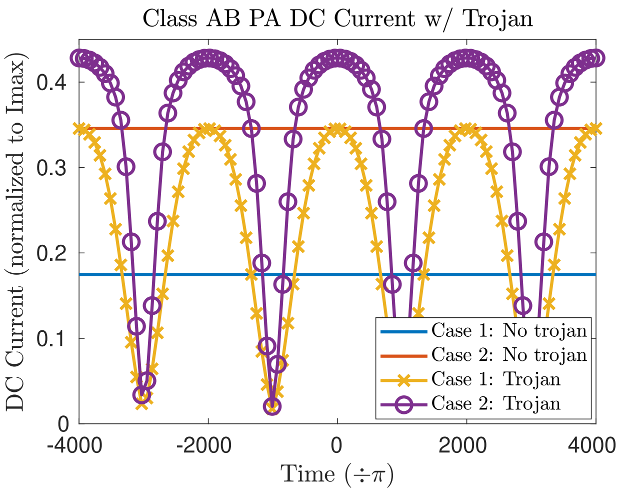

FIGURE 6. Calculated normalized DC current draw of Class AB operation is shown over time for two cases of coupled input signal amplitudes: (1) in which the total amplitude barely drives the PA to compression and (2) in which the total amplitude drives the PA well into compression.

Adding back the K term of $-K\cos(\Phi_2)$ to V_{in} to account for significantly different amplitudes between the original PA input signal and coupled external signal results in the following expression for $\hat{V}_{in}$:

$$\hat{V}_{in} = Q + (1 - Q)\frac{2r_1 C_1 \cos\left(\frac{\Phi_1 + \Phi_2}{2}\right) - K\cos(\Phi_2)}{V_{sat} - V_q} \tag{10}$$

Applying the cubic MOSFET current model and numerically integrating for I_{DC} results in the traces shown in Fig. 7. The low frequency pseudo-sinusoid persists in the case of a coupled external signal with a significantly lower amplitude than the PA input signal, although the amplitude of oscillation falls significantly such that the waveform approaches the "No Trojan" cases illustrated in Fig. 6. Compression of the sinusoid peak, as with the equal amplitude case, is controlled by the total amplitudes of the PA input signal and coupled external signal.

2) COUPLED INPUT SIGNAL: CLASS E PA ANALYSIS

Determining the impact of the coupled input signal trojan on a Class E PA design requires a different approach due to the nature of switching PA operation. Because all amplitude information is lost during the operation of a standalone Class E PA, we must turn to calculating the instantaneous frequency as a result of the addition of an external signal to the original input. Combining this information with basic assumptions of the nature of the PA driver amplifier enables the modification of the classical Class E PA boundary conditions for numerically determining a solution to the on-state and off-state differential equations governing Class E PA behavior.

Inspecting the envelope of the addition of two sinusoids of similar frequencies and different amplitudes reveals the existence of "nodes" and "antinodes" where the amplitude

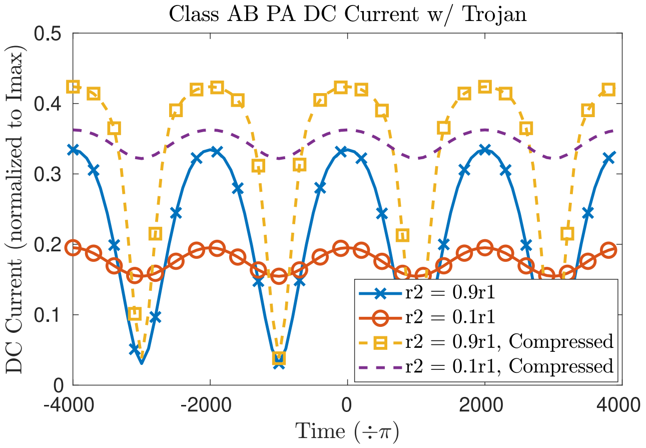

FIGURE 7. Calculated normalized DC current draw of Class AB operation with unequal input and coupled signal amplitudes is shown over time for two cases of coupled input signal amplitudes: (1) in which the total amplitude barely drives the PA to compression and (2) in which the total amplitude drives the PA well into compression.

of the overall waveform displays valleys and peaks. The relative heights of these valleys and peaks are dependent on the relative amplitudes of the two sinusoids, with the severity of the variations in overall waveform amplitude increasing the closer the amplitudes of the component sinusoids are. The driver amplifiers of a Class E design seek to shape the input signal into a trapezoidal waveform so as to effectively drive the PA transistor as a switching element, but would clearly fail at this task at the nodes. This effect may be modeled through linearly scaling the PA transistor on-resistance with envelope amplitude.

Revisiting the Class E PA boundary conditions for an arbitrary frequency requires accounting for the possibility that the PA transistor drain capacitance is not fully discharged by the time of the turn-on instance, and factoring in the on-resistance scaling effect naturally requires the switch model of the PA transistor to be modified with a finite resistance. Simplifying the latter possibility by assuming the drain capacitance discharge occurs quickly in comparison to the switching period and applying the requirements for the PA drain voltage and inductor current to be continuous results in the following refined normalized boundary conditions where $t = 0$ is defined to be the turn-off instant and $\omega t = \pi$ is defined to be the turn-on instant: (1) $\hat{V}_D(0)$ must be defined to avoid an under-constrained solution, (2) $\frac{d\hat{V}_D}{dt}(0) = w\hat{I}_{SW}(0^-)$, and thirdly:

$$\frac{\hat{V}_D}{dt}(\pi) = \omega\left[2\hat{I}_0\cos(\phi) + \hat{I}_{SW}\left(0^-\right) + \frac{\pi C_{sw}}{\epsilon^2}\left(\hat{V}_D(0) - 1\right)\right] \tag{11}$$

The fourth boundary condition may be derived from the fundamental frequency Fourier coefficients for the normalized drain voltage $\hat{V}_D(t)$, the cosine coefficient a_1 and the sine coefficient b_1 as calculated from the general solution for $\hat{V}_D(t)$. Assuming that the admittance of the PA load, which is composed of the LC series resonator, output matching network, and load resistance, is known to be Y_L at the operating frequency gives the following expressions for the output current phase ϕ and the normalized output current magnitude $\hat{I}_0$:

$$\phi = \arg(Y_L(a_1 + b_1 j)) \tag{12}$$

$$\hat{I}_0 = \frac{1}{\omega C_{SW}}||Y_L|| \times ||a_1 + b_1 j|| \tag{13}$$

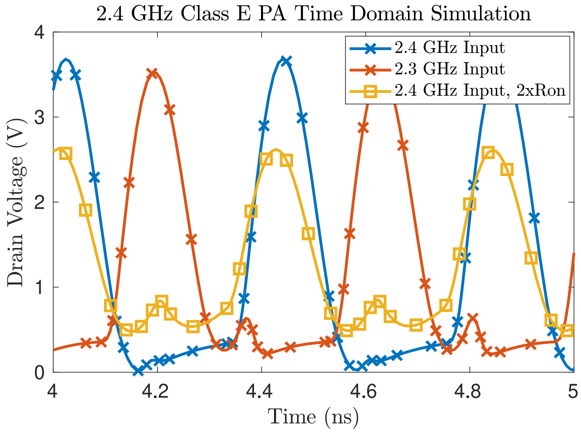

FIGURE 8. Simulated drain voltage of Class E PA designed for 2.4 GHz operation in 65nm PDK with on-resistance scaling, and altered input frequency.

Examining these boundary conditions brings us to the conclusion that changes to the PA transistor on-resistance as a result of a coupled external signal significantly impact the operation of the PA, as $\hat{I}_{SW}(0^-)$ is inversely proportional to on-resistance. However, due to the design of the PA reactive drain network for bringing the drain voltage close to zero with zero-slope at the nominal turn-on instance, small variations in operating frequency do not greatly impact PA operation, and thus the power consumption [19]. This can be seen in the simulation results shown in Fig. 8, where the drain voltage of a Class E PA circuit designed for 2.4 GHz in a 65 nm PDK is simulated for a 2.3 GHz input frequency and for the impact of doubling the PA transistor on-resistance. Clearly, the latter change has by far the greater effect on the PA.

Repeating the simulation for an input signal composed of two sinusoids of amplitudes r_1 and r_2 and frequencies ω_1 and ω_2 where $\omega_2 = 0.999\omega_1$ represents the coupled external signal from the trojan results in the power supply current shown in Fig. 9. All frequency content in the plot above 100 kHz is filtered out to represent the low-pass characteristic of a power profile readout sensor. For a coupled external signal with amplitude equal to the input signal, the power supply current clearly droops in time with the valleys and peaks of the envelope of the overall input waveform, as concluded earlier. Furthermore, the coupled external signal with a significantly smaller relative amplitude only yields a small change from the power supply current that occurs during ordinary operation. Thusly, as the amplitude of the coupled external signal approaches that of the input signal, the power signature of the trojan becomes more distinguishable through a power profiling sensor system.

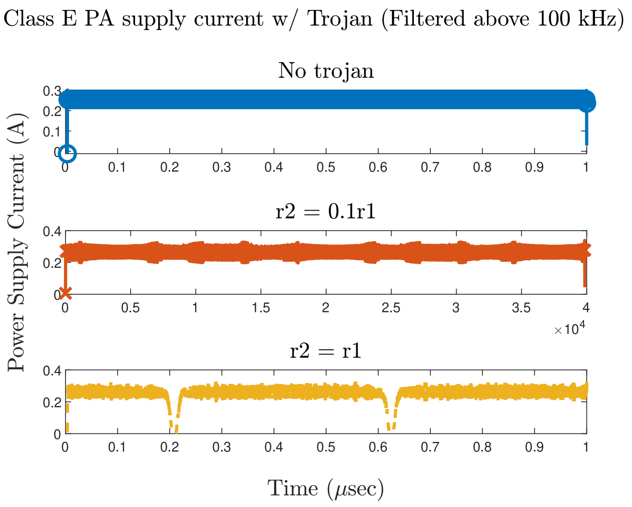

FIGURE 9. Simulated power supply current of Class E PA designed for 2.4 GHz operation in 65nm PDK is shown for three cases: (1) no trojan (2) the amplitude of the coupled external signal is 1/10th of the input signal's (3) the amplitude of the coupled external signal is equal to the input signal's.

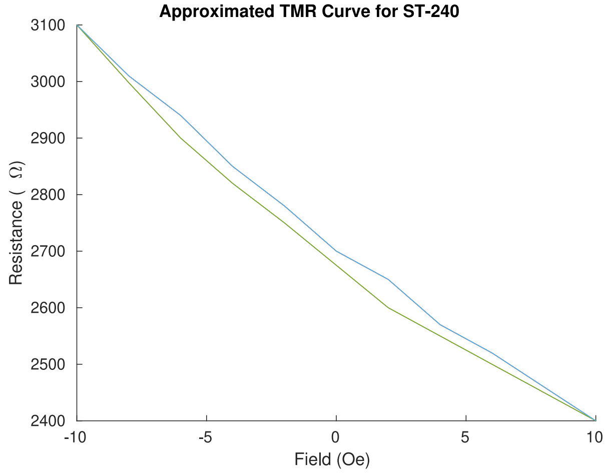

FIGURE 10. A graphical representation of a typical MicroMagnetics TMR curve. Note that this is similar to the manufactures' data test of the MTJ before being shipped to customer [24].

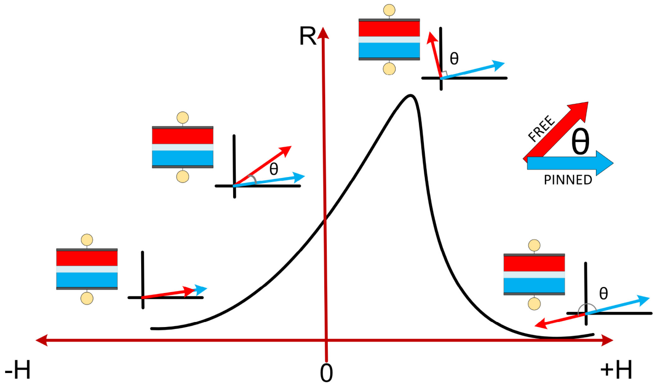

FIGURE 11. Typical TMR curve illustrating the relationship between the pinned and fixed layer of an MTJ in response to various H-fields and their impact on device resistance.

IV. SENSOR INTERFACE, INTEGRATION, AND TESTING

Built-in self-testing circuits for side-channel protection such as [8] generally involve placing some sort of resistive component directly into the current carrying line of the device. While placement of a sensing resistor on the current-carrying line is a viable solution, it requires some finite, dynamic power loss related to the resistive element. Recently, tunnel magnetoresistive (TMR) research advanced the sensitivity and viability of utilizing TMR for various sensing applications. TMR sensors change resistance based upon the strength of the magnetic field, and thus do not require placement directly into the current carrying or signal path. Hence, an opportunity to provide low-power non-invasive sensing in regards to the signal and current carrying line arises due to the TMR sensors unique properties. An interface for a commercial sensor was developed to collect data and test the overall effectiveness of a neural network for detecting and classifying the two aforementioned trojans. Note that all of these were tested on a PCB.

A. MTJ SENSOR PHYSICS

TMR sensors are based upon magnetic tunnel junction (MTJ) sensors are based upon the relationship between resistance and magnetic field intensity H that arises from tunneling magnetoresistance. TMR sensors are generally constructed using three different layers, a free layer, a pinned layer, and a barrier layer that provides a junction between the free and pinned layers. The pinned and free layers are constructed with magnetoresistive materials that change their resistance based on the intensity of the magnetic field strength through the sensor [23]. Thus, the device resistance changes as a function of the angle between the fixed and pinned layers of the MTJ. Because of the quantum tunnel effect, electrons will have a higher or lower probability of travelling through the intermediary film due to the alignment of the layers. Hence, the general formula for the angle-dependent resistance of a general MTJ sensor is derived in Eq. (14). The sensitivity of the MTJ plays a role in determining the placement of the sensor. The COTS sensor utilized in this particular study had a sensitivities given in the range of 0.7-2%/Oe of resistance change with a 0.21% hysteresis. At zero field, the resistance of the system was approximately 2.5 $k\Omega$. A typical TMR curve for the Micromagnetics ST-240 [24] is displayed in Fig. 10. Note that this is approximate and is based upon general Micromagnetics data. Note that the important portions of this graph are its curvature based on hysteresis and the slight nonlinearity of the spin tunnel junction sensor.

$$R(\theta) = \frac{1}{2} TMR * i \frac{R * A}{W * h} cos\left(\theta_p - \theta_f\right) \tag{14}$$

Fig. 11 illustrates a general magnetic flux-resistance curve, and the relationship to the angle between the fixed and pinned layers. Note that there are two distinct regions of operation. Previous MTJ development centered around memory technology, that is, providing a sharp transition between one and zero based on an applied magnetic field. Hence, around some set value of H, the resistance curve sharply transitions from a low resistance to a high resistance, as illustrated by 11.

1) MTJ SENSING APPLICATIONS

MTJ sensors have been used for a diverse range of applications in the literature including proximity sensors, current measurement sensors [23], and large-scale power systems [25]. Because the sensor transforms the magnitude and direction of the magnetic flux passing through its sensing region into an easily measured resistance, it is a prime candidate for non-invasive sensing applications. The sensor itself does not need to be directly inserted into the DUT's signal chain, and simply acquires electromagnetic information from the proximity of the DUT. Of course, a MTJ sensor embedded within an IC for security applications would require a supervisory circuit either through pinouts on the IC or as a built-in current sensor (BICS), similar to [16]. However, such a system would differ from [16] in two key ways - there would be no energy losses within the signal chain due to inserted components, and backend data processing, analysis and classification algorithms would be implemented digitally. Hence, the nature of the MTJ as well as its small form factor provide a promising new expansion to various BICSs already in use.

B. SENSOR INTERFACE

The readout circuit used to acquire data from the MTJ sensor is shown in Fig. 1. Magnetic flux from the power line of the DUT alters the resistance of the MTJ, which is placed in a simple resistor divider to generate an output voltage signal. A commercial INA821 instrumentation amplifier is used to amplify the resultant signal, and an ADC samples the signal. The resultant data samples are fed into an FPGA for detection of anomalies through the use of a machine learning module. These systems were implemented on a PCB, which was then connected externally to an FPGA. Note that in this case, a resistor divider circuit along with a an instrumentation amplifier was utilized to provide the gain for the sensor. The resistor divider (Eq. (15)) output voltage is amplified by the instrumentation amplifier (INA821) with some selected resistor. The gain of the instrumentation amplifier was set to 50. Hence, the readout circuit was developed for sensing small changes in the resistance of the MTJ sensor.

$$V_{sense} = V_{ref}\frac{R_{fixed}}{R_{fixed} + R_{sense}} \tag{15}$$

C. SENSOR INTEGRATION

The MTJ must be placed at a distance from the DUT such that the range of the DUT's magnetic field is within the sensitivity range of the sensor. Placing the sensor into the actual device could have several advantages which would outweigh a number of disadvantages. Protection would be added by creating an IP that provides supervisory information concerning the operation of the IC by utilizing the sensor plus the interface circuitry. While it is possible this particular IP could be tampered with, the MTJ and interface circuitry would be necessarily quite small in form factor and mostly likely quite difficult for those within a untrusted fabrication unit to know to damage this particular unit. Furthermore, this MTJ could be fabricated in a trusted environment before packaging, providing an increased amount of trust. Assuming the sensor could be placed on-chip at an arbitrary distance from the current carrying line to be monitored, the magnetic field strength observed at the sensor would be found by estimation of the magnetic field in the near-field. An approximate closed-form expression can be found by assuming that the current carrying line is a collection of multiple conductive planes at some depth d constrained with some width w, with the current divided equally over each of the current-carrying planes. This leads to an expression in which the magnetic field at the sensor distance d_{sensor} is (16) [26], assuming that the sensor is in the near-field. If the device is placed on the IC interconnect, the H field sensed at that location can be approximated utilizing Eq. (16). Assuming the current is produces an H-field that is within the sensing range of the MTJ, there is the potential that the MTJ can be utilized as a non-invasive BIST for monitoring the power signals.

$$H = \frac{I}{\pi * w} arctan\left(\frac{w}{2h}\right) \tag{16}$$

Due to the inability to integrate a MTJ sensor on-chip with the DUTs that were used for this study, a COTS MTJ sensor was used to emulate an integrated on-chip sensor. To emulate the magnetic field experienced by an on-chip sensor, it is necessary to determine an equivalent magnetic field strength corresponding to the COTS sensor. In the far-field case, magnetic field strength emanating from a current carrying line is proportional to the inverse of the radial distance from a source of electromagnetic radiation (Eq. (17)). An inductor with a particular geometry was wound around the sensor to produce a field similar to that which would represent an optimal design through integration of the MTJ into the IC design. This geometry was selected in consideration of the field strength produced by a current carrying wire (Eq. (17)), which was used to choose the inductor area and turn number. Fig. 12 indicates the dynamic response associated with a range of possible inductance values. Note that with an instrumentation amplifier gain of approximately 50, set by an external resistor on the INA821, approximately 20 turns were required to adequately sense currents smaller than 100 mA. Utilizing the current amplification scheme improved dynamic range of the system.

$$B_{sensor} = \frac{\mu_0 In}{2\pi R}, n = \text{number of turns} \tag{17}$$

D. EXPERIMENTAL METHOD AND RESULTS

The dataset for the hardware trojan side-channel detection method was generated by implementing and testing the sensing circuit and data acquisition on a PCB, B-field magnification schemes, and the two hardware trojans described

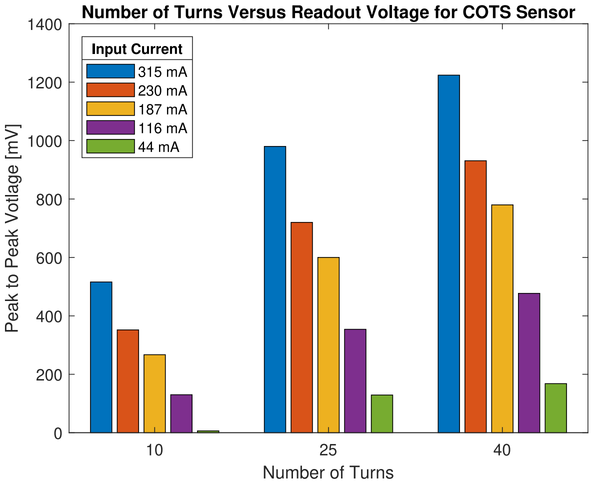

FIGURE 12. Selection of the number of turns required to properly simulate the magnetic field.

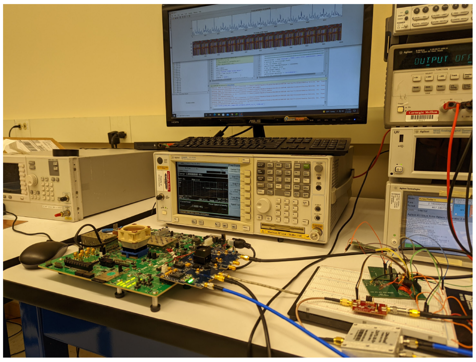

FIGURE 14. Testbench for the Lab. Signals to be transmitted are input into the PA (the COTS PA is pictured here) to which the trojans have been added. The data from the MTJ and the output of the PA are used to monitor and determine the impacts of the trojans, respectively.

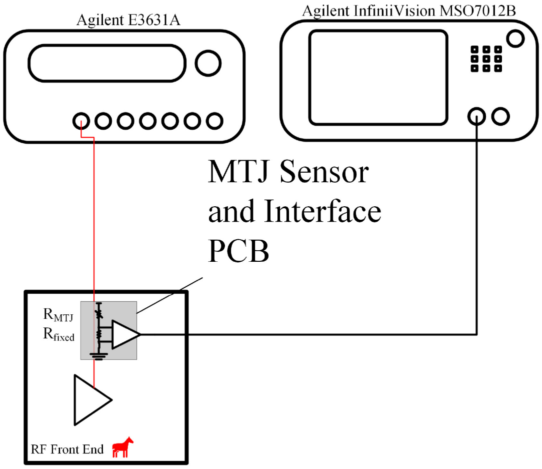

FIGURE 13. Testbench for the Lab. Signals to be transmitted are input into the PA (the COTS PA is pictured here) to which the trojans have been added. The data from the MTJ and the output of the PA are used to monitor and determine the impacts of the trojans, respectively.

in Section II, a block diagram of which is shown in Fig. 13. In particular, the hardware trojans were developed by placing a transistor on the high side of the power amplifier, to simulate a power switch trojan, and by injecting an additional signal into the PA input line with a commercial power combiner. For each hardware trojan tested and for each type of input signal, the transmitted output of the power amplifier and the side-channel monitored signal were measured. Single-tone and Bluetooth modulated input signals were used. Furthermore, to determine the ability of the detection system to classify various trojans, various input tone magnitudes and frequencies were tested. The data collection was repeated for each B-field magnification scheme. A typical curve for the normal and trojan operation of the circuit as sensed by the MTJ conditions is highlighted in Fig. 14. Note that in this case, the figure illustrates "normal" operation, that is, when the PA is transmitting data, when the known trojan is operative, and when the unknown trojan is operative. In this particular experiment, we utilized the "switch trojan" as the unknown trojan and the "power combiner" trojan the known trojan. Note that because of the orientation of the device based upon the magnetic field, higher voltage drop on the readout circuit is akin to lower power consumption, while lower normalized values refer to increased power consumption. Normal operation is between these two values.

For this paper, the trojans developed and tested on a PCB were connected to the TI COTS PA and the Class E PA. Note that because of the narrowband operation of the class E PA, BLE data with a with a center of frequency 2.4 GHz was tested. The COTS PA was tested with single tone inputs at 2.4 GHz. In this case, a toroid inductor placed on the current carrying line was utilized to magnify the field and the instrumentation amplifier had a gain of two. Note that because the input frequency impacts the operation of the amplifier, the current draw differs and can be clearly distinguished. Fig. 4 illustrates the apparent changes in current draw for a coupled input signal. The single-tone inputs were generated using a frequency generator, while the Bluetooth modulation signals were generated by utilizing a MATLAB script and an AD9082 transmitter. The transmitted Bluetooth output signal from the PAs were tested utilizing an RF DAC front end and MATLAB demodulation scheme and the EM side-channel data was sampled at a rate of 400kSa/s utilizing an oscilloscope. A picture of the board level interface is described in Fig. 15. Data was then processed, quantized, and normalized before training and testing For training and testing the neural network, the sections where the trojans were on and off were labeled as anomalies. The dataset was expanded by quantizing the data into various quantization levels, ranging from four to sixteen bits. These quantized datasets were then used to empirically determine the most effective quantization level required for accurate detection and classification of a particular trojan. Additionally, to indicate the dynamic range of the system, each magnification scheme was evaluated for

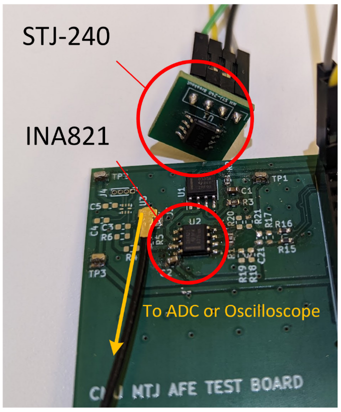

FIGURE 15. Picture of PCB level interface. The STJ-240 and INA821 interfaced through the cable pictured with the oscilloscope.

a single tone input as well as modulation scheme to determine the amount of B-field magnification and amplifier gain required for the system to effectively determine the existence of trojans.

V. RF FRONT END SECURITY: ANOMALY DETECTION WITH NEURAL NETWORKS

Anomaly detection algorithms have been applied to a number of applications, from large data-center scale application to the low-powered devices in the Internet of Things. For applications at the edge, a number of different anomaly detection algorithms applied to the Internet of Things. Note that the table includes a variety of both "heavy" and "light" -weight algorithms, that is, those that are more complex and require more energy usage, as opposed with those that are lighter in energy usage. Furthermore, a number of papers in present literature present anomaly detection and hardware trojan detection in various integrated circuits, including those at the edge.

A. DETECTION AND CLASSIFICATION OF TROJAN MODELS

Malicious trojans must be discovered within a particular hardware system and hence, the methods described in Section II-B generally deal with one-class classification, even from the first paper published on the subject [17]. One-class classification generally looks at operation considered "normal" from a golden chip and then classifies whether or not the system itself is secure. In this case, an autoencoder was utilized to provide self-supervised determination of whether a particular time series occurrence of the trojan was detected. Furthermore, the nature of the autoencoder could be utilized to provide various amounts of trust to certain types of waveforms. Hence, a classification autoencoder, which provides a one-class classification of whether or not a system has seen a trojan, as well as a determination of the type of previously known trojan detected is presented in this paper.

B. CLASSIFIER AUTOENCODER FOR CLASSIFICATION AND TROJAN DETECTION

1) AUTOENCODER

The autoencoder is a well-known learning model with a number of variations ([27], [28]). A simple autoencoder has three main sections: (1) encoder (MLP), (2) latent variables (z), and (3) decoder (MLP), highlighted in Fig. 17. Data (denoted x) is propagated through the encoder until it reaches the latent variables, which provide a lossy encoding. The latent variables (z) are then decoded to generate some variable $\hat{x}$ that imitates the original input. Because of their generative property, autoencoders can be used to generate data to imitate the original input signal, albeit with some information lost due to lossy compression. Such a generated signal provides an opportunity to train the model to adapt to the introduction of new, unknown trojans or operational modes where traditional MLP classifiers could not. Because of the relatively lightweight structure of such an unsupervised learning model, the autoencoder becomes a good match for providing defense against unknown signals, and in this case, become quite helpful in detecting signals from analog trojans. Furthermore, this paper also utilizes the structure of the autoencoder to provide an MLP classifier using the encoder.

Autoencoders can take a number of different forms, and can be built form both convolutional neural networks, but the most simple would be the version based upon the MLP or deep neural networks. Defining $\phi(x)$ to be the encoder and $\psi(x)$ to be a decoder, the autoencoder is defined by Eq. (18).

$$\hat{x} = \phi(\psi(x)) \tag{18}$$

Typical training of the autoencoder is based upon the difference between the generated data $\hat{x}$ and the original input data. Typically, mean-square error (MSE) is used to define the loss function between the real and generated data. Furthermore, for this particular application, the training assumes that some trojans may be already known (or recorded), and hence, data may be available with information known concerning these trojans. Hence, it is useful to utilize the Classification Autoencoder (CAE) not only to classify the various trojans, but also to provide protection for unknown trojans or operating states. In order to train the autoencoder for both classification and generative models, the training algorithm 17 is presented based on the mathematics found in [29].

$$\mathcal{L} = ||x - \hat{x}||^2 \tag{19}$$

2) AUTOENCODER FOR CLASSIFICATION ALGORITHM

An algorithm for training the CAE is presented in 17. Note that for this experimental design, it is assumed that there are three types of signals in two groupings: (1) known signals and (2) unknown signals. Known signals are further categorized into normal operation signals and trojan operation signals. Hence, the classifier aims to determine not only

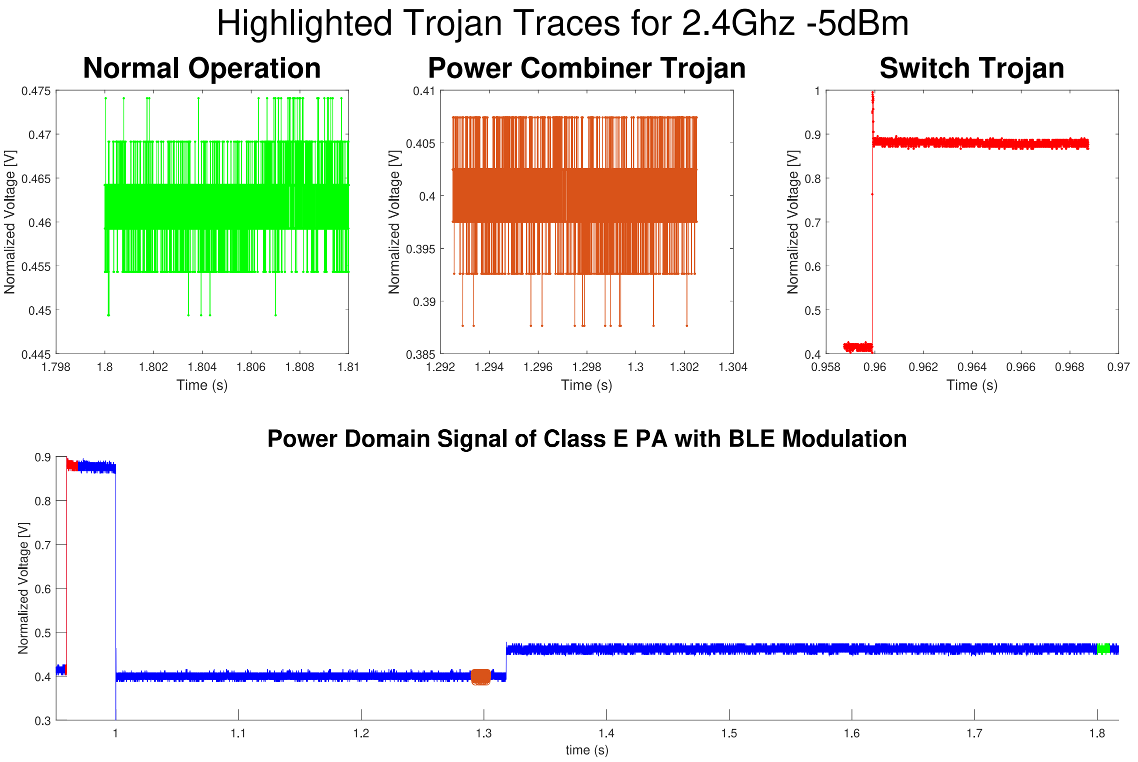

FIGURE 16. Output signal for a normalized, quantized MTJ readout circuit measurement. Note that the spikes are due to the switch trojan, while the pulses are due to the pcomb trojan.

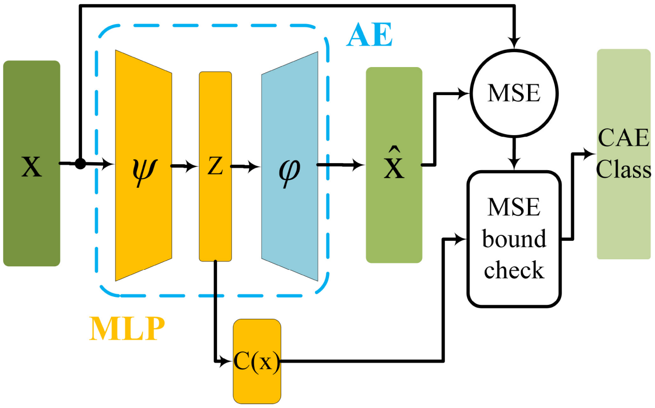

FIGURE 17. Diagram of the Classification Autoencoder. A typical autoencoder comprises of an MLP encoder (ψ), and an MLP decoder (ϕ). The generated output is compared to the input to the encoder.

whether a signal is abnormal in comparison to known normal operation, but also to classify the signals within the "normal" range as either a normal operation signal or a trojan operation signal. In particular, training is first carried out by first dividing the dataset up into known and unknown data. In the training phase, known data, including both normal operation and a switch trojan was used to train the MLP classifier, but not the autoencoder. Instead, losses from the output of the autoencoder are only backpropagated in training for those samples of the normal or typical data type. Then, in the testing and validation stage, known and unknown data were combined to test the ability of the classifier to determine an abnormal and normal signal. In particular, the algorithm outlines what will occur when the loss function is high, which would indicate a higher probability that the generated data is not similar to the prior training data, and when the classifier indicates that the trojan is classified as a non-trojan. In this scenario, the typical response of the MLP would be to classify as a non-trojan, as it attempts to classify data it has yet to encounter. In order to determine the boundary between the signals not previously known and the signals, the boundary was empirically determined. Fig. 18 graphically illustrates the determination of the decision boundary in the training and testing phase of the system. Note that the known trojan and the normal operation signals are both expected to have lower loss, which will be discussed in a following section.

In order to determine the boundary for which signal could be trusted, a small subset of the data passed through the AE in order to determine the magnitude of the expected loss for the various types of input vectors. The losses were then found to have some distribution, with some expected value. The weighted average of the distributions, in particular, the expected value from the known distribution with highest loss amongst the known distributions and the unknown distribution with the highest overall loss, were used to set a decision boundary. As pictured in Fig. 18, the decision boundary indicates where the MLP can be trusted to make a classification, shaded green in the figure. By properly training the

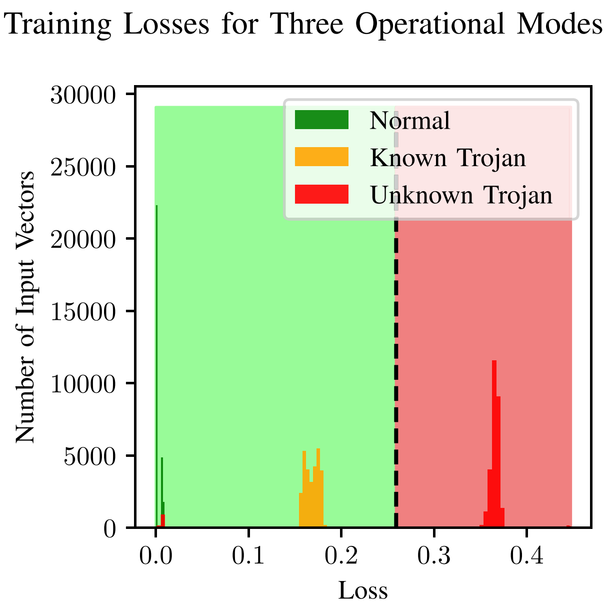

FIGURE 18. Graphical representation of the determination of the loss boundary for the CAE. Normal, known, and unknown trojan vectors were input, and some weighted average of the known and unknown trojans were then used to determine a decision boundary (black vertical line) to indicate, based on the generated data, what could be trusted.

```
Algorithm 1 Training the Classifier Autoencoder
Require: Training set $S = \{\{N, Tr\}, \{T_N, T_{Tr}\}\}$
  Initial value $\Theta_\psi, \Theta_\phi$
  Class output $C$
  Encoder $\phi$
  Decoder $\psi$
  Classifier $\mathcal{C}$
Ensure: $\hat{\Theta}_\psi, \hat{\Theta}_\phi$
  for $x$ in S do
    $C = \mathcal{C}((\psi(x)))$
    $\hat{x} = \phi(\psi(x))$
    $\mathcal{L}_{MLP}$ = BCEwithLogits($C$, $T_{Tr}$)
    $L_{AE} = ||x - \hat{x}||^2$
    for $l$ in $\mathcal{L}_{MLP}$ do ▷ backpropagate for normal signals
      if $T_l = T_N$ then
        $\theta_\psi^{k+1} \leftarrow \theta_\psi^k - l$
      end if
    end for
    $\theta_\phi^{k+1} \leftarrow \theta_\phi^k - l$
  end for
```

network, the autoencoder is able to provide a defense against unknown signals and operational modes. Furthermore, it is generally robust to noise, whereas something such as a comparator would require that all trojans would have the same impact - they would negatively impact the current to such an extent that the output would be quite noticeable. Unknown signals can be classified without prior recognition of the signals. Such self-supervised learning could also be utilized to determine whether there are issues in the chip generally, as well as whether multiple different types of trojans afflicting

```
Algorithm 2 Testing Algorithm
Require: Trained Parameters $\hat{\Theta}_\psi, \hat{\Theta}_\phi$
  Test set $T = \{\{N, Tr\}, \{T_N, T_{Tr}\}\}$
  Test set $T_u = \{\{U\}\}$
  Validation set $V = \{\{N, Tr\}, \{T_N, T_{Tr}\}\}$
  Validation set $V_u = \{\{U\}\}$
  $V > T$, $V_u > T_u$
Ensure: $\hat{\Theta}_\psi, \hat{\Theta}_\phi$
  for $x$ in $T, T_u$ do
    $\hat{x} = \phi(\psi(x))$
    $\mathcal{L}_{AE} = ||x - \hat{x}||^2$
  end for
  $b = mean(L_{AE,N}, L_{AE,U})$
  for $x$ in $V, V_u$ do
    $\hat{x} = \phi(\psi(x))$
    $C = \mathcal{C}(\psi(x))$
    if $||x - \hat{x}||^2 < b$ then $C = \mathcal{C}(\psi(x))$
    else $C = T_{Tr}$
    end if
  end for
```

TABLE 3. CAE accuracy and false positive improvement for TI PA single tone and Class E PA BLE signals.

	Class E PA BLE			TI PA Single Tone		
Model	FP	FN	Acc	FP	FN	Acc
CAE	0.5	0	99.7	1.7	0	98.2

the IC. Furthermore, if the number of chips is appreciably large and cost-prohibitive to replace, it could be that these sensors indicate the level of trust in these communication due to the particular indication from this system. One particular downside in this consideration is that, depending upon the number of known trojans, it maybe difficult to keep the size of the autoencoder and related portions small and energy-efficient to run on chip. However, it should be assumed that particular chips that will require protection would not have a large amount of known trojans, let alone any trojans at all, due to their simple nature.

3) POWER SIGNAL DATA SELECTION AND SOFTWARE TRAINING RESULTS

In software testing, the CAE provided considerable benefits to the protection of the circuit. Table 3 presents the results of training and testing the neural network with full precision. Single tone and Bluetooth results from Table 3 indicate that with the AE, trojans could be detected with accuracy up to 98% and 99.7%, respectively, providing a great deal of added protection. Hence, the algorithm for inference provides not only a way to detect trojans, but to provide some quantifiable increase in accuracy in classification of unknown data the previously available. Furthermore, Fig. 19 illustrates the robust noise rejection for the single tone case.

4) HARDWARE RESULTS FOR SINGLE TONE

The CAE was synthesized on a ZCU102EVM utilizing a MPSoC Zynq Ultrascale+ FPGA, with an autoencoder

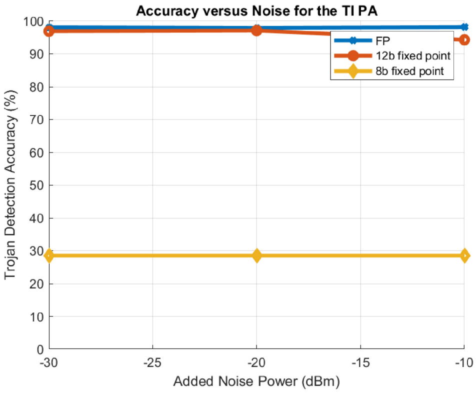

FIGURE 19. Robust noise rejection for the CAE. With various levels of simulated noise added to the signal, the a 12b fixed point CAE was able to classify both known and unknown signals with high accuracy.

TABLE 4. Resource utilization comparison for various autoencoder implementations.

Reference	FF	LUT	BRAM	DSP	Latency (ns)
Single Tone	29681	75925	0	623	1245
BLE	4571	6297	7	81	815
[28]	59100	23640	0.38 Mb	-	130
[29]	241522	45448	1068	1360	185

architecture of 32-16-8-16-32 and a two class classifier output, and the results are listed in Table 4, along with comparable FPGA implementations of similar anomaly detection algorithms utilizing autoencoders can be found in Table 4. Note that this work produces a high accuracy model for IOT anomaly detection with low latency and high accuracy. Furthermore, it has similar resolution requirements as [27] implementations. Furthermore, this particular design uses less resources than either [28] or [27].

5) HARDWARE RESULTS FOR BLE

A second model was developed for the BLE utilizing 16-bit fixed-point arithmetic. The model developed for the BLE input signals was an smaller overall model, with a 16-8-4-8-16 autoencoder for one-class classification and a two-class classifier output. Accuracy of trojan detection while utilizing fixed-point hardware was around 99% when approximately 25% of the data was unknown. Furthermore, hardware resources were lowest across the board and the latency was decreased in comparison with the single tone results.

C. NOISE CONSIDERATIONS AND LIMITATIONS

1) NOISE CONSIDERATIONS

Because of the nature of the system dealing with hardware, noise and environmental conditions must be considered a constant consideration. Note that there is some environmental noise that could have an impact. In order to indicate the ability of the system to properly classify whether a trojan existed or not wideband gaussian noise was injected into the single tone test signals. Fig. 17 illustrates the accuracy of the CAE for various levels of injected noise, ranging from -30 dBm to -10 dBm. Hence, it is important to note that even with some noise injected into the system, the autoencoder could still classify the signals as trojan and non-trojan with a high level of accuracy.

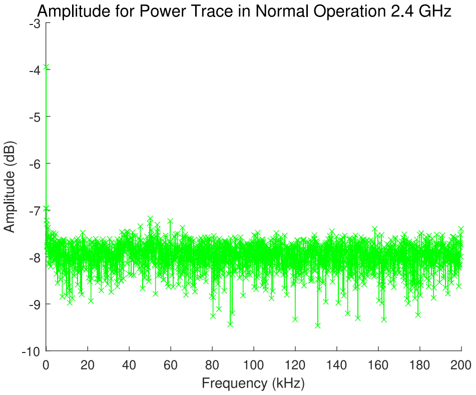

FIGURE 20. Fourier analysis of normal operation trace with multiple dormant trojans. Note that some signals, especially in the range of 50 to 60 kHz are visible, but 3dB smaller than the important traces, which mainly impact the DC signal from the power supply.

Noise in MTJ sensors is also a concern and does limit the overall operation of the device. Environment noise and its impacts on the MTJ were not the focus of this study and were noted in the experimental data collection. Fig. 20 illustrates the power profile of the system. Even in an environment in which a number of test devices were simultaneously working, the system was able to detect the changes in current based upon the trojan attack circuits due to the low magnitude of the signals that created various noise. It should be noted that large magnets could interfere with the device, and could even reset the device, which is a known limitation for MTJ sensors. In order to properly secure the circuit from outside magnetic interference, further study and magnetic shielding may be required.

2) LIMITATIONS

While the autoencoder could be used to provide a one-class classification that would defend against various trojans, it can be limited in the that it would require retraining once the signal is different from that which it recognizes, that is, as the environment changes or perhaps noise increases, the autoencoder may not be able to provide high levels of fidelity in the signal analysis. Hence, the weights used for the autoencoder may need to be updated as the noise and environmental requirements change. Furthermore, it is assumed that the number of known trojans on which the front half of the MLP is trained remains small as only so many trojans would be known or discovered beforehand. As the number of classification outputs increases, more latent states, and hence more neurons, may be required to classify the MLPs involved. This would increase the power and area of the device, which would lead to various issues when it comes to power requirements. This problem is generally mitigated by the fact that the autoencoder can be used to ensure particular operation signals are known (and trusted), while everything outside of those signals are considered untrusted and a trojan.

VI. CONCLUSION

This paper presents a system to detect hardware trojans in the analog/RF domain through the usage of magnetic tunnel junction sensors. Hardware trojans for a Class E and Class AB power amplifier were characterized and tested, and classifier-autoencoder was synthesized utilizing an FPGA for detection of anomalistic signals within time-series data from an RF power amplifier. The classification-autoencoder classified known normal, known trojan, and known unknown signals with approximately 97% accuracy with 12 bit fixed-point arithmetic even with up to 20dBm of injected noise, and 94% accuracy with 10dBm of injected noise. With floating point, the system was able to retain approximately 98% accuracy for all test cases. Furthermore, a lighter-weight 16 bit BLE CAE was also developed that also considerably improved accuracy of the MLP from 44% to 99.7%

REFERENCES

[1] Y. Jin, D. Maliuk, and Y. Makris, "Hardware trojan detection in analog/RF integrated circuits," in *Secure System Design and Trustable Computing*, C.-H. Chang and M. Potkonjak, Eds. Cham, Switzerland: Springer Int., 2016, pp. 241–268.

[2] B. Zhou *et al.*, "Hardware trojan detection using backside optical imaging," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 1, pp. 24–37, Jan. 2021.

[3] V. Govindan and R. S. Chakraborty, "Logic testing for hardware trojan detection," in *The Hardware Trojan War: Attacks, Myths, and Defenses*, S. Bhunia and M. M. Tehranipoor, Eds. Cham, Switzerland: Springer Int., 2018, pp. 149–182.

[4] K. Xiao, D. Forte, Y. Jin, R. Karri, S. Bhunia, and M. Tehranipoor, "Hardware trojans: Lessons learned after one decade of research," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 22, no. 1, pp. 1–23, 2017.

[5] S. Bhunia and M. M. Tehranipoor, *The Hardware Trojan War: Attacks, Myths, and Defenses*, 1st ed. Cham, Switzerland: Springer Publ. Company, Incorp., 2017.

[6] Q. Wang, D. Chen, and R. L. Geiger, "Transparent side channel trigger mechanism on analog circuits with PAAST hardware trojans," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2018, pp. 1–4.

[7] B. Gungor, M. Yazici, E. Salman, and Y. Gurbuz, "Establishing a covert communication channel in RF and mm-wave circuits," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, 2020, pp. 1072–1075.

[8] Y. Liu, Y. Jin, A. Nosratinia, and Y. Makris, "Silicon demonstration of hardware trojan design and detection in wireless cryptographic ICs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 4, pp. 1506–1519, Apr. 2017.

[9] K. Yang, M. Hicks, Q. Dong, T. Austin, and D. Sylvester, "A2: Analog malicious hardware," in *Proc. IEEE Symp. Security Privacy (SP)*, 2016, pp. 18–37.

[10] G. T. Becker, F. Regazzoni, C. Paar, and W. P. Burleson, "Stealthy dopant-level hardware trojans: Extended version," *J. Cryptograph. Eng.*, vol. 4, no. 1, pp. 19–31, 2014.

[11] D. Ernst *et al.*, "Razor: Circuit-level correction of timing errors for low-power operation," *IEEE Micro*, vol. 24, no. 6, pp. 10–20, Nov./Dec. 2004.

[12] J. Rajendran, V. Jyothi, O. Sinanoglu, and R. Karri, "Design and analysis of ring oscillator based design-for-trust technique," in *Proc. 29th VLSI Test Symp.*, 2011, pp. 105–110.

[13] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, and B. Sunar, "Trojan detection using IC fingerprinting," in *Proc. IEEE Symp. Security Privacy (SP)*, 2007, pp. 296–310.

[14] X. Wang, H. Salmani, M. Tehranipoor, and J. Plusquellic, "Hardware trojan detection and isolation using current integration and localized current analysis," in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Syst.*, 2008, pp. 87–95.

[15] L. Lin, W. Burleson, and C. Paar, "MOLES: Malicious off-chip leakage enabled by side-channels," in *IEEE/ACM Int. Conf. Comput.-Aided Design Dig. Tech. Papers*, 2009, pp. 117–122.

[16] K. S. Subramani, A. Antonopoulos, A. A. Abotabl, A. Nosratinia, and Y. Makris, "ACE: Adaptive channel estimation for detecting analog/RF trojans in WLAN transceivers," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, 2017, pp. 722–727.

[17] Y. Jin and Y. Makris, "Hardware trojan detection using path delay fingerprint," in *Proc. IEEE Int. Workshop Hardw. Orient. Security Trust*, 2008, pp. 51–57.

[18] Y. Jin, D. Maliuk, and Y. Makris, "Post-deployment trust evaluation in wireless cryptographic ICs," in *Proc. Des. Autom. Test Europe Conf. Exhibit. (DATE)*, 2012, pp. 965–970.

[19] K. C. Tsai, "CMOS power amplifiers for wireless communications," Ph.D. dissertation, Dept Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Dec. 2007.

[20] C. Park, Y. Kim, H. Kim, and S. Hong, "A 1.9-GHz CMOS power amplifier using three-port asymmetric transmission line transformer for a polar transmitter," *IEEE Trans. Microw. Theory Techn.*, vol. 55, no. 2, pp. 230–238, Feb. 2007.

[21] Y. Lee, C. Park, and S. Hong, "On-chip power combining method in CMOS power amplifier," in *Proc. Asia–Pacific Microw. Conf.*, 2008, pp. 1–4.

[22] H. Enzinger, K. Freiberger, and C. Vogel, "A joint linearity-efficiency model of radio frequency power amplifiers," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2016, pp. 281–284.

[23] H. Heidari, *Magnetic Sensors for Biomedical Applications* (IEEE Press Series on Sensors). Hoboken, NJ, USA: Wiley, 2019.

[24] *STJ-240 Single-Axis Magnetic Sensor, Rev. C.*, MicroMagnetics, Fall River, MA, USA, 2012.

[25] P. Shrawane and T. S. Sidhu, "Application of magnetic sensors for measurement of current phasors in power systems," in *Proc. IEEE Elect. Power Energy Conf. (EPEC)*, 2021, pp. 530–535.

[26] E. Chen, J. Kan, B.-Y. Yang, J. Zhu, and V. Chen, "Intelligent electromagnetic sensors for non-invasive trojan detection," *Sensors*, vol. 21, no. 24, p. 8288, 2021.

[27] E. Govorkova *et al.*, "Autoencoders on FPGAs for real-time, unsupervised new physics detection at 40 MHz at the large hadron collider," 2021, *arXiv:2108.03986*.

[28] D. J. M. Moss, D. Boland, P. Pourbeik, and P. H. W. Leong, "Real-time FPGA-based anomaly detection for radio frequency signals," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2018, pp. 1–5.

[29] L. Yu and X. Gao, "A robust classification-autoencoder to defend outliers and adversaries," 2021, *arXiv:2106.15927*.

JOHN KAN received the B.S. degree in electrical engineering from the University of Illinois Urbana–Champaign in 2019. He is currently pursuing the Ph.D. degree with EECS Lab, Carnegie Mellon University. As an undergraduate, he designed an educational swarm robotics platform and worked to build UAV platforms for cyber-physical system security. His interests include cyber-physical system security, data collection and processing, and low-power sensors.

YUYI SHEN (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Carnegie Mellon University in 2020, where she is currently pursuing the Ph.D. degree. She held an internship position with Apple Inc., in 2020. She is primarily interested in RFIC design with a focus on the application of RF circuits to security and device identification. She is a recipient of ISSCC Analog Devices Outstanding Student Designer Award in 2021 and the Ben Cook Graduate Fellowship in 2022.

JIACHEN XU (Graduate Student Member, IEEE) received the B.S. degree in computer engineering from Purdue University in 2020. He is currently pursuing the Ph.D. degree with Carnegie Mellon University. His interests lie in brain-inspired machine-learning algorithms and embedded system design for wireless applications. He is a recipient of ISSCC Analog Devices Outstanding Student Designer Award in 2022.

ETHAN CHEN is a Research Scientist with the Energy-Efficient Circuits and Systems Lab, Carnegie Mellon University. His research interests include neuromorphic computing, hardware security, and biomedical interfaces.

JIMMY ZHU (Fellow, IEEE) received the Ph.D. degree in physics from the University of California at San Diego in 1989. Prior coming to Carnegie Mellon in 1997, he had been an Assistant Professor and later an Associate Professor with the Department of Electrical Engineering, University of Minnesota from 1990 to 1996. He has authored and coauthored over 300 refereed papers in major international journals along with seven book chapters and has given over 90 invited papers at various major international conferences. He has graduated over 45 Ph.Ds. either in electrical and computer engineering, materials science and engineering, and physics. He holds 22 U.S. patents. Some of the awards that he has received include the McKnight Land Grant Professorship from University of Minnesota in 1992, the NSF Presidential Young Investigator Award in 1993, the R&D Magazine Top 100 Invention Award in 1996, the IEEE Magnetic Society Distinguished Lecturer in 2004, the Carnegie Mellon University Outstanding Research Award in 2010, the IEEE Magnetic Society Achievement Award, and the highest award of the IEEE Magnetic Society in 2011. He also received the ETA KAPPA NU Excellent Teaching Award in 2012 at Carnegie Mellon.

VANESSA CHEN (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2013.

She was with Qualcomm, San Diego, CA, USA, working on energy-efficient data-acquisition systems for mobile devices. From 2010 to 2013, she was with Carnegie Mellon. She focused her research on self-healing systems and high-speed ADCs and held a research internship position with IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, in 2012. She was an Assistant Professor with The Ohio State University, Columbus, OH, USA. She is currently an Assistant Professor of Electrical and Computer Engineering with Carnegie Mellon University. Her research interests focus on data conversion interfaces for machine learning, RF/analog hardware security, ubiquitous sensing and communication systems. She was a recipient of the NSF CAREER Award in 2019, the Analog Devices Outstanding Student Designer Award in 2013, and the IBM Ph.D. Fellowship in 2012. She is also an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and IEEE OPEN JOURNAL OF CIRCUITS AND SYSTEMS and a Guest Editor of the *ACM Journal on Emerging Technologies in Computing Systems*. She is a Technical Program Committee Member of the IEEE Symposium on VLSI Circuits, the IEEE Custom Integrated Circuits Conference, the IEEE Asian Solid-State Circuits Conference, and the IEEE/ACM Design Automation Conference.