electronics

Article

A Non-Destructive Method for Hardware Trojan Detection Based on Radio Frequency Fingerprinting

Siya Mi [1,2], Zechuan Zhang [1,*], Yu Zhang [3,4] and Aiqun Hu [2,5]

1 School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
2 Purple Mountain Laboratories, No. 9 Mouzhou East Road, Nanjing 211111, China
3 School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
4 Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China
5 School of Information Science and Engineering, Southeast University, Nanjing 211189, China
* Correspondence: zechuanzhang@seu.edu.cn

Abstract: Hardware Trojans (HTs) pose a security threat to the Internet of Things (IoT). Attackers can take control of devices in IoT through HTs, which seriously jeopardize the security of many systems in transportation, finance, healthcare, etc. Since subtle differences in the circuit are reflected in far-field signals emitted by the system, the detection of HT status can be performed by monitoring the radio frequency fingerprinting (RFF) of the transmitting signals. For the detection of HTs, a non-destructive detection method based on RFF is proposed in this paper. Based on the proposed method, the detection of HTs can be achieved without integrating additional devices in the receiver, which reduces associated costs and energy consumption. QPSK and triangular-wave signals are measured and identified via experimentation, and the results validate the proposed method. For identifying the presence and operating state of Trojan, the average accuracy achieved measures as high as 98.7%. Notably, with regard to capturing the moment of Trojan activation in the AES encryption circuit, the accuracy of the proposed method is 100% and can provide warning of the threat in a timely manner.

Keywords: hardware security; hardware Trojan detection; radio frequency fingerprinting; AES; USRP; triangular wave; QPSK

Citation: Mi, S.; Zhang, Z.; Zhang, Y.; Hu, A. A Non-Destructive Method for Hardware Trojan Detection Based on Radio Frequency Fingerprinting. *Electronics* **2022**, *11*, 3776. https://doi.org/10.3390/electronics11223776

Academic Editor: Wojciech Mazurczyk

Received: 28 September 2022
Accepted: 13 November 2022
Published: 17 November 2022

1. Introduction

Internet of Things (IoT) products are becoming prevalent globally due to the development of wireless devices and communication systems. IoT is a system where smart devices are connected as the basis for interconnection. In many cases, IoT devices collect confidential information and connect to the cloud via wireless communication, due to their mobility as well as computing and energy limitations. However, the propagation of wireless signals in the public domain renders them vulnerable to security flaws [1]. Therefore, the security and integrity of IoT communications have received much attention. In many studies, hardware Trojan (HT) attacks have been considered as a major security issue for integrated circuits (ICs). With the rapid development of the integrated circuit industry and the globalization of manufacturing, hardware suppliers often outsource product manufacturing to third-party vendors to reduce costs. However, some of the insecure third-party vendors may neglect potentially implanted HTs in the electronic products. HTs can be triggered under specific conditions, which can lead to serious errors in circuit function or leakage of information. Nowadays, integrated circuits are extremely complex, and the number of nodes in circuits is growing exponentially. It is easy to conceal the presence of HTs, whereas it is challenging to detect their presence in circuits.

Because of the threat of HTs, Trojan-detection methods are increasingly being emphasized. In the pre-silicon phase, static-detection techniques can extract and analyze the characteristics of the gate-level netlists to identify suspicious networks without logical or functional simulation, and machine learning methods can be used to classify the unknown networks into Trojan and normal networks efficiently [2–5]. However, these techniques make it difficult to build perfect models for all ICs. The detection techniques in the post-silicon phase can be divided into destructive and non-destructive detection techniques. Since destructive detection techniques are extremely expensive and time-consuming, non-destructive detection techniques tend to receive more attention. The dominant non-destructive detection is realized by testing or side-channel analysis. Logic testing is usually achieved by activating the Trojan payloads and detecting errors in the output. By generating input trigger vectors and Trojan detection vectors, HTs can be detected; however, the specific type of Trojan horse needs to be known in advance [6]. In order to synthesize Trojan circuits with the desired trigger probabilities, researchers have discussed methods to design combinatorial rare situations [7]. However, only the detection of combinatorial logic Trojans is considered. In order to detect Trojans, two or more detection methods can be used together [8]. In practice, the Trojan-detection methods based on logic testing have limitations due to the unknown nature of the Trojan, and it is impossible to traverse all the test vectors. The side-channel detection methods are realized by measuring various side-channel signals from the IC to identify differences caused by Trojan circuits [9]. Such information includes path delay, transient current, power signal, temperature profile, and EM radiation profile [10–16]. The implantation of a Trojan horse affects normal circuit operation, and changes in the circuit are reflected in the side-channel information. The IC fingerprint can be extracted from the side-channel information and used to detect the Trojan, but this method has been validated only in the detection of several specific Trojans [9]. The IC fingerprints are then extracted from the side-channel signals of all path delays in the netlist, and they are proposed to be used to classify the explicit payload and implicit payload Trojans [10]. The delay caused by the implicit load Trojan is quite small, so small that its use renders it difficult to detect Trojans. Since the delay path is proportional to the number of chip nodes, this method is seldom used to test large circuits. For the large noise caused by process variation, the multidimensional side-channel analysis is proposed for Trojan detection [11]. The dynamic current measurement using vector generation can improve the sensitivity of Trojan detection. To improve detection accuracy, logic testing and side-channel analysis can be combined [12]. The probability of Trojan activation is increased by statistically generated test vectors, while the accuracy of the detection method based on side-channel information is improved by functional and structural analyses. Nonetheless, this method is rarely applied for real-time detection. For Trojan detection in large circuits, the inspection can take up to ten hours. The Trojan circuit can be detected by analyzing transient power-supply signals [13]. Although the adverse effect of process variations on Trojan resolution is considered, the detection method is not scalable in circuit size. Methods using temperature side-channel information of the board effectively detect the HT in the off-the-shelf ICs, and these methods can detect a Trojan before it is triggered [14].

The golden reference of the chip is necessary for the side-channel detection methods, but the golden references are not always available. Sometimes the EM side-channel information in the early stage of the IC life cycle can replace the golden reference [15]. However, there is a deviation between the golden reference and the EM side-channel signal used as the reference. This method is only proven for the detection of several types of Trojan, and the Trojan has to be restricted to specific areas to achieve a more concentrated EM radiation. In order to detect the HT without golden reference, an unsupervised clustering algorithm is proposed, which is achieved by classifying the data based on controllability and observability of the gate-level netlists information [17]. This method only performs static analysis of the netlist without test vectors to activate the Trojan. The gate-level netlists are not actually available in their entirety, so this method

does not always work. The use of a self-referencing approach to detect Trojan horses eliminates the need for the golden chip. Since the method relies on the device under test for training, process variations do not affect its detection mechanism [18]. However, the method requires highly precise side-channel signal acquisition, and monitoring consumes considerable energy.

For wireless encrypted ICs, there exists a portion of Trojans that can evade the detection of some existing methods, and the information is leaked over the wireless channel. The side-channel information is proposed to detect whether there is information leakage through the wireless channel [19]. The side-channel analysis is proved in the detection of HTs in the wireless communication circuit [20]. It is experimentally derived that the impact of a Trojan is hidden in the transmission tolerances of legitimate process variations, and is coupled to the communication channel and noise. In addition, this method also needs the complete design for the accurate identification of HTs.

In order to detect the HT in the wireless encryption transmitter, a non-destructive detection method is proposed, which is based on radio frequency fingerprinting (RFF). By analyzing the received signal in the receiver, the presence and status of HTs can be recognized. The proposed method can work well without additional devices or components. Moreover, for detecting HTs, updating of the system requires less cost and energy.

The remainder of this paper is organized as follows. Section 2 describes the generalized structure of Trojan circuits and the basic methods of RFF. Section 3 introduces the HT identification method based on RFF in detail. Section 4 describes the wireless communication experimental platform. The experimental results evaluating the proposed detection method are presented and discussed in Section 5. Finally, Section 6 concludes the paper.

2. Background

An HT is a malicious modification of the original circuit, and represents a hidden autonomous circuit. When the HT is triggered, the functionality of the system will be altered. The effects of HTs can be classified as the change of functionality, the degradation of performances, the leakage of information, and the denial of service [21]. The data-leakage Trojans can transmit the internal signal of the circuit to the output port, and hide the information in the RF signal. Then the security information will be leaked to the attacker.

Typically, an HT consists of two parts, namely the trigger and the payload [22]. Figure 1 shows a general HT. The Trojan payload is triggered only when an enabling signal is sent from the trigger. Thus, HTs are difficult to detect during the circuit verification phase. Some HTs can be activated by specific conditions, and others can be activated after a pre-set delay. Depending on the payload type, the Trojan can be considered as explicit payload Trojans and implicit payload Trojans [10]. The explicit payload modifies the internal control signals or data signals, which then changes the original function of the circuit. The implicit payload uses the internal signals for excitation of the trigger. According to its working mechanism, the implicit payload Trojan does not usually destroy the value of the internal signals after triggering, but it may reduce the overall chip life or leak information by activating additional modules. The circuit shown in Figure 1 presents an implicit payload Trojan. When its payload is activated, the internal signals will be transmitted to the specific module. Most of the known data-leakage Trojans are implicit payload Trojans, and can leak confidential system information through hidden channels. This presents a major security risk to system security and communication security.

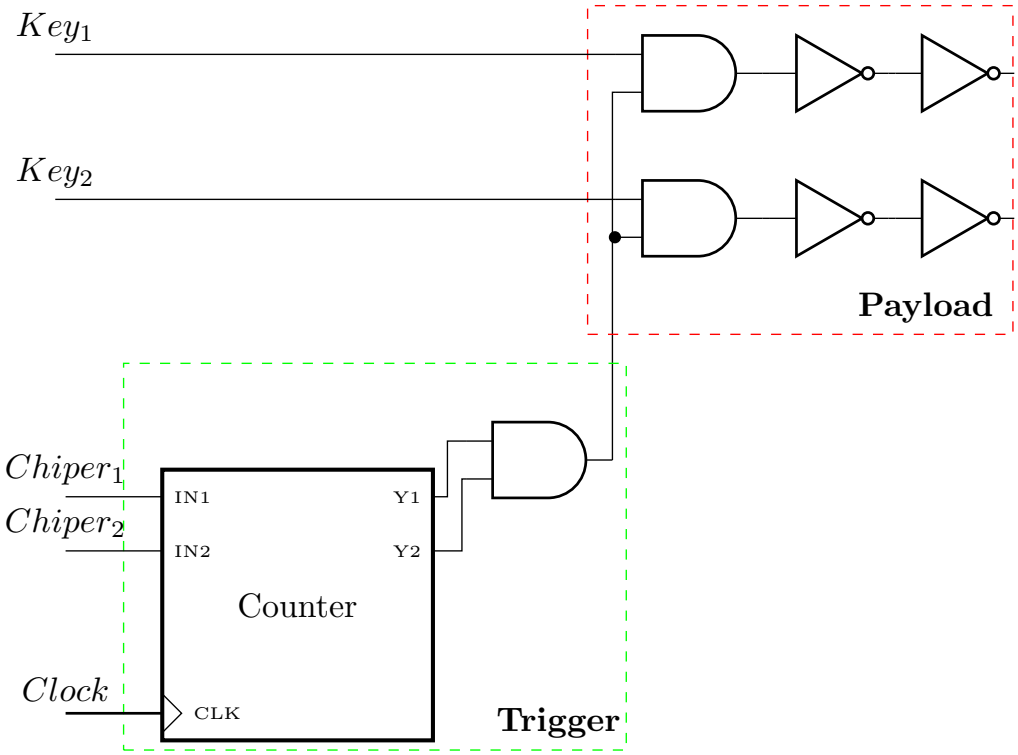


Figure 1. Sequential Trojan circuit architecture.

The detection method of HTs mainly involves identifying the difference between the Trojan-infected circuit and the golden reference. Compared to Trojan-free circuits, the Trojan circuits cause changes in electromagnetic parameters such as leakage current and parasitic capacitance. These changes can be observed via side-channel detection. Meanwhile, alterations caused by the Trojans will also be indicated in the far-field wireless signal. The proposed method uses the test modality of the wireless transmission channel. It detects Trojan circuits using RFF which is obtained by feature extraction of the collected wireless signals. The impact of activated Trojans in the transmitter circuit will be studied in this paper, and the presence and operating status of Trojans will be identified via RFF. At the receiver side, the wireless signal $r(t)$ is mathematically expressed as

$$r(t) = h(t) * \mathcal{T}(s_I(t) + js_Q(t)) + n(t), \tag{1}$$

$$\mathcal{T}(\cdot) = \begin{cases} \mathcal{T}_{dor}(\cdot), & Trigger = 0 \\ \mathcal{T}_{act}(\cdot), & Trigger = 1 \end{cases} \tag{2}$$

where $h(t)$ is the channel impulse response. *Trigger* is the activation signal of Trojan circuit. $\mathcal{T}_{dor}(\cdot)$ is the Trojan distortion function in dormant state. $\mathcal{T}_{act}(\cdot)$ is the Trojan distortion function in active state. $s_I(t)$ and $s_Q(t)$ represent the in-phase and quadrature part of the transmitted signals, respectively. $n(t)$ is the Gaussian white noise signal. Figure 2 shows the time-domain baseband signal of Quadrature Phase Shift Keying (QPSK) data. The valid signal part contains five sample frames.

The RFF is caused by device-processing errors. Transmitter-processing inevitably introduces reasonable errors, and these errors will be reflected in the transmitted RF signal. RFF has already been proven to be stable and unique [23]. Both explicit payload and implicit payload Trojans can change the internal logic of the circuit, which in turn affects the electromagnetic near field and far field. When the HT is activated, the circuit will also become altered, and the change will be indicated in the RFF. Thus, the RFF can be used to detect the HT in the circuit of the transmitter.

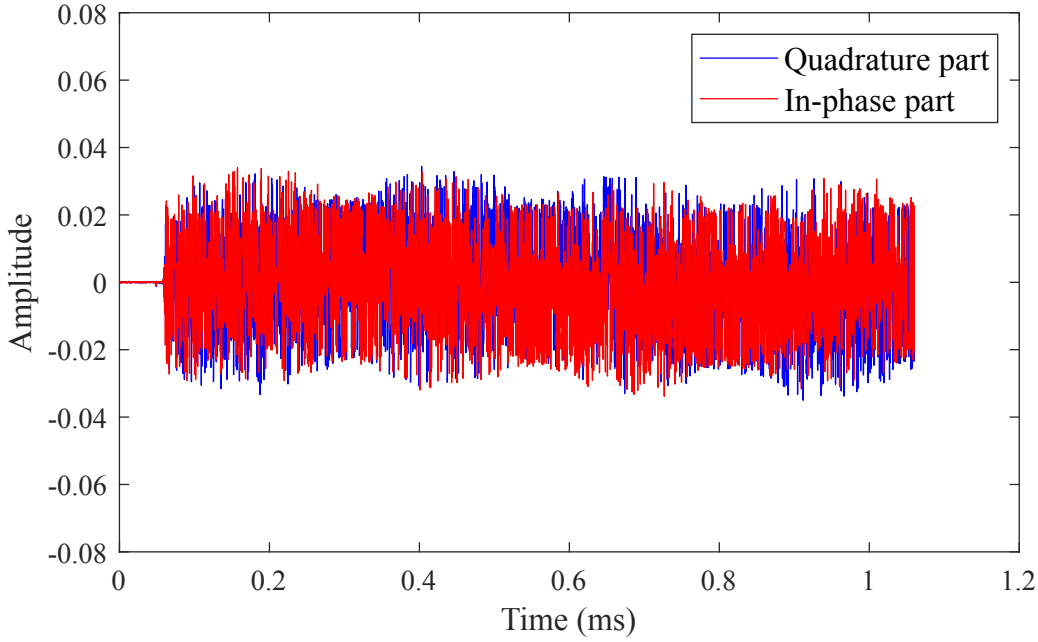


Figure 2. IQ-captured time domain signal samples.

The RFF can be divided into transient RFF and steady-state RFF. The transient RFF are the signal features contained in the instantaneous signal when the transmitter is turned on or off. The extraction of transient RFF features requires a highly precise received signal at the receiver. The steady-state fingerprint feature is the fingerprint feature contained in the stable operation of the transmitter. The extraction of steady-state fingerprint features does not require a relatively high-performance receiver. The RFF features that have been proven effective are carrier frequency offset, synchronization signal correlation value, baseband I/Q offset, and amplitude and phase offset of the demodulated signal [24]. Compared to the RFF extraction methods based on expert knowledge, the application of machine learning in RFF recognition improves the identification in terms of accuracy without expert knowledge [25]. In this case, $RFF = DNN(IQsamples)$. It is well known that recurrent neural networks (RNNs) can process correlations of series data, but the residuals returned by recurrent neural networks decrease exponentially as the running time continues, which results in slow updating of network weights. To solve the problem of residual descent, long short-term memory (LSTM) uses forget gate and input gate to control the residual information. In this paper, the steady-state RFF will be used for the detection of hardware Trojans concerning their impact on the internal transmitter circuit.

3. Hardware Trojan Detection Based on Radio Frequency Fingerprinting

The received signals are sequential, and the features of the sequential signals should be extracted to detect the presence and status of the Trojan. The revived RF signal is often measured in frames, and each frame carries a specific message. Thus, the signal itself is not only related to past-sequence information but also future-sequence information. Since LSTM can only use past-signal information, the bidirectional LSTM (Bi-LSTM) will be employed in this paper. The different signal modulation methods result in various dominant features carried by different sections of the frame. The self-attention method will be used to extract the relationship between temporal features to achieve accurate recognition. The overall flow of the detection method is shown in Figure 3. The proposed method is trained to learn a map $\mathcal{F}(\cdot)$ to obtain the RFF feature of the wireless signal. The

detection system is then used as a classifier to precisely identify whether the IC is infected with a Trojan circuit based on RFF. This is formulated as

$$RFFfeature = \mathcal{F}(r[n]), \tag{3}$$

where $\mathcal{F}$ is the model-mapping function.

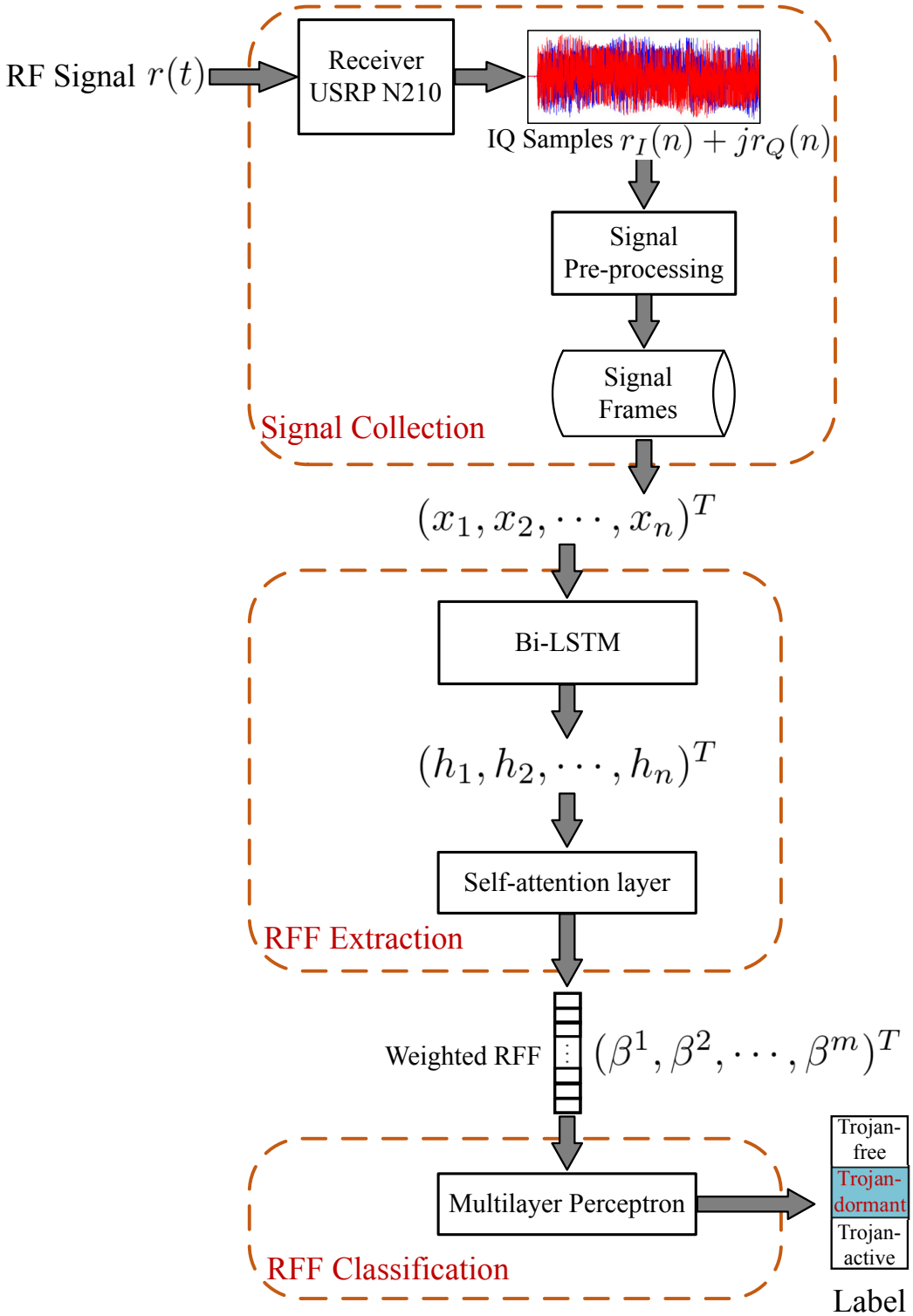


Figure 3. Structure diagram of the proposed hardware Trojan-detection method.

Bi-LSTM consists of a forward LSTM and a backward LSTM as shown in Figure 4, which is a special neural network model designed to solve problems with dependence on temporal relationships, consisting of forget gates, input gates, and output gates. Bi-LSTM is used to extract the feature of RFF for accurate identification.

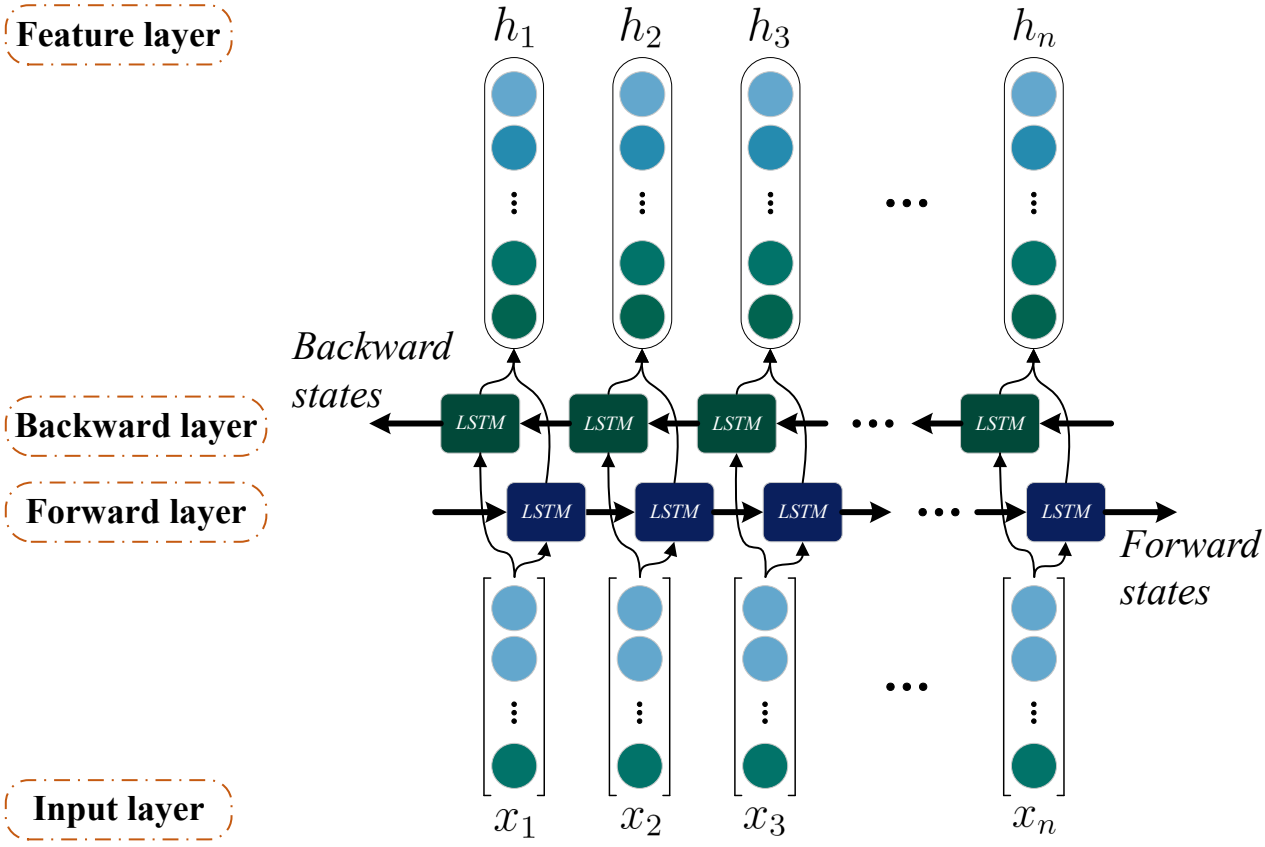


Figure 4. Bi-LSTM structure.

The LSTM input comprises the output of the previous LSTM cell, h_{t-1}, and the input of the current cell, s_t. The information required to remove is controlled by the forget gate, and is given as follows

$$f_t = \sigma(W_f \cdot [h_{t-1}, s_t] + b_f), \tag{4}$$

where σ is the activation function. The value of σ is between 0 and 1, and is used to adjust the nonlinear transformation. W_f is the parameter of this cell. b_f indicates the offset. W_f and b_f are used to relate f_t to the state of the previous cell. The input gates are computed in parallel to control the content of the next LSTM cells. The input gate is

$$i_f = \sigma(W_i \cdot [h_{t-1}, s_t] + b_i), \tag{5}$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, s_t] + b_C), \tag{6}$$

where $\tilde{C}_t$ is the updated value of the previous state. i_f is the input. By adding the input gate and the forget gate, the state of the current LSTM cell, C_t, can be obtained. The output gate needs to consider the forget gate and the input gate to determine the output value of the current LSTM cell, h_t. The output gate is calculated as

$$o_t = \sigma(W_o \cdot [h_{t-1}, s_t] + b_o), \tag{7}$$

$$h_t = o_t * tanh(C_t). \tag{8}$$

After the received signal has been processed by Bi-LSTM, the features extracted by each cell are output in temporal order. Since the self-attention mechanism can extract the long-distance dependencies of features, it is used to learn the temporal relations of features extracted by Bi-LSTM. The self-attention takes a set of Bi-LSTM sequential features as the input, and uses the Query and Key mapped by the linear layer to determine the weights of each feature. The output of self-attention is a vector of weighted feature values as shown in the following

$$\beta^i = \sum_{m=1}^{n} Softmax(\frac{q^i k^{mT}}{\sqrt{d_k}})v^m, \tag{9}$$

where q^i is the Query mapped by the i-th output of Bi-LSTM. k^m is the Key mapped by m number of outputs of Bi-LSTM. v^m is the value mapped by m number of Bi-LSTM outputs. d_k is the dimension of Query and Key, and it is used to scale. β^i is the result of self-attention for the i-th output of Bi-LSTM. Based on the Bi-LSTM model, this paper proposes an algorithm to extract the RFF as shown in Algorithm 1.

Algorithm 1 RFF Extraction Algorithm Based on Bi-LSTM

Input: wireless signal $r[n]$
Output: RFF feature β
1: $epoch$ = number of training epochs.
2: m = number of signals
3: Generate a set of training samples $D_m = \{(r_m[n], label_m\}$
4: **for** j =1 to $epoch$ **do**
5: **for** i =1 to m **do**
6: Update Forget Gate f_t based on Equation (4)
7: Update Input Gate i_f based on Equation (5)
8: Update LSTM Cell C_t based on Equation (6)
9: Update Output Gate h_t based on Equation (8)
10: Compute self-attention vector β^i based on Equation (9)
11: **end for**
12: **end for**
13: **return** bi-LSTM model

After obtaining the RFF features, a multilayer perceptron (MLP) is used to classify and identify the received signal. ReLU is used to perform the nonlinear transfer to accelerated convergence and prevent overfitting and gradient disappearance.

4. Experimental Section

In order to validate the proposed method, the programmable universal software radio peripheral (USRP) is used to simulate the Trojan-free transmitter and the Trojan-infected transmitter. Our experiment comprises two parts, namely the platform construction and the dataset generation. For platform construction, the FPGA development kit we used for experimentation is ISE14.7. The USRP Hardware Driver (UHD) is 3.11.1 and the Current Hardware Revision of N210 is 4. Narrowband-IoT (NB-IoT), a standard for Low-Power Wide-Area Networks (LPWAN), offers longer transmission range and lower energy consumption. QPSK (Downlink transmission scheme) and $\pi/4$-QPSK (Uplink transmission scheme) modulation schemes are used in the physical channels of NB-IoT. In the experiment, QPSK modulation is used. The RF receiver is set to receive a sampling rate of 5 MHz to oversample the wireless signal. The carrier is set at a frequency of 2.43 GHz.

4.1. Experimental Platform

A wireless communication experimental platform is set up to evaluate the proposed approach. The wireless encryption IC used in the experimental platform is USRP N210 from Ettus, which consists of a digital part and an analog part.

The digital part is a Xilinx Spartan 3A-DSP 3400 FPGA integrated into the board, and is used as the controller circuit. The digital circuit works at a master clock frequency of 100 MHz. From a complexity perspective, the Spartan-3A DSP 3400 chip provides 3.4 million system gates. This FPGA chip is comparable to the millions of gates of SoC tested in real production. The RF daughterboard, as the analog part, is CBX-40. It operates in the RF frequency range from DC to 6 GHz, and supports a maximum of 25 MHz RF bandwidth with 16-bit samples. USRP N210 uses a Dual 100 MS/s, 14-bit ADC, and a Dual 400 MS/s, 16-bit DAC to complete the complex sampling of the signal. This increases the maximum processing capacity of the full-duplex communication system to 100 MS/s, while potentially improving processing latency. We performed additional functions based on the original functions of the N210 digital part in order to simulate the working operation of a wireless transmitter with subtle differences. With the programmable FPGA, we implemented a Trojan-free transmitter with encryption and a Trojan-infected transmitter on the Spartan-3A DSP 3400 chip.

4.1.1. Trojan-Free Transmitter

The Trojan-free transmitter with the function of encrypting the transmit signal is implemented. The circuit benchmark used in this paper is AES. AES is a widely used key encryption algorithm, and it is popular in the field of wireless communication encryption. It should be noted that the proposed method is not limited to detecting Trojans in the AES cryptographic circuits. The detailed implementation of the encryption circuit is based on the AES IP core. The IP core is a sequential design with an input plaintext length of 128 bits. A total of 10 iterations are employed, with each iteration comprising 4 phases, namely byte substitution, row shifting, column mixing, and round key addition.

In the experiment, AES-128 with OFB mode is used. A 128-bit register is set for storing the encrypted output. The default data sample on the USRP N210 board is a 32-bit fixed-point number. The AES encryption module receives the plaintext in 32-bit length, and the 128-bit key is stored on the chip to encrypt the plaintext. To meet the data throughput rate of the encrypted communication circuit N210, the AES-128 module uses pipelined parallel computing. The setup time of the pipelined circuit requires 20 clock cycles. The top-level module of the AES IP core includes a sub-module for performing key expansion and a sub-module for implementing the single-round AES encryption algorithm.

4.1.2. Transmitter with Trojan

Most HT-detection methods are evaluated by the Trojan circuit benchmarks of the trust HUB online repository [26]. The data leakage Trojan AES-T2100 is used to attack the AES encryption module in the wireless encryption IC. This Trojan is designed to leak data via a leakage current channel, and will be triggered after a specific number of encryptions. It is also representative of the implicit load circuit.

To test the sensitivity of the proposed method regarding the detection of the implicit load-numbing circuit, the triggering process of the Trojan horse is simulated. After processing a pre-set amount of data, the Trojan horse will be activated. The signals emitted by the transmitter with both dormant Trojan and active Trojan are received and analyzed.

The circuit of the Trojan is shown in Figure 5. A shift register holding sensitive information, and multiple sets of inverters comprise the load circuit of the AES-T2100 Trojan. The least significant bit of the shift register is connected to the first inverter, and the output of this inverter is connected to another inverter. When the least significant bit of the shift register remains low, the PMOS of the first inverter conducts. Meanwhile, the path between the input pin of the other inverter and the ground is composed,

and the NMOS conducts to set the output pin low. A direct path between power and ground in each couple of inverters is composed in a limited time. This temporary leakage current causes a direct increase in power consumption of the circuit. Sensitive information is leaked out stealthily. The Trojan can apparently be detected by measuring the electromagnetic information in the near field [16]. However, variations in EM energy in the Trojan circuit will also have an impact on the emitted RF signal.

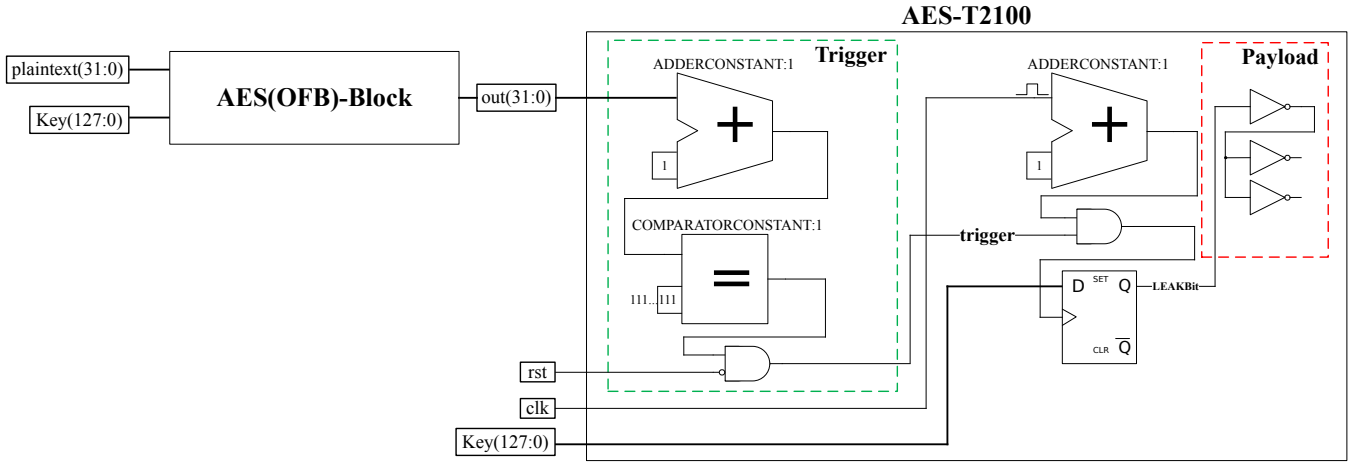


Figure 5. Circuit diagram of AES-T2100 Trojan.

The lookup table (LUT) is the smallest unit-storing circuit in an FPGA circuit. The number and percentage of LUTs utilized in the FPGA system infected by AES-T2100 Trojan are shown in Table 1. The second, third, and fourth columns of the table show the number of LUTs utilized in the FPGA system, the LUTs utilized in the encryption module and the LUTs utilized in the Trojan module, respectively. It can be deduced from the table that the AES-T2100 Trojan is a subtle variation compared to the N210's FPGA system.

Table 1. Resource utilization of the AES circuit and AES-T2100 Trojan circuit.

FPGA Syetem	LUTs	LUTs for AES	LUTs for Trojan
Transmitter with AES-T2100	32883	129 (0.392%)	103 (0.313%)

4.1.3. Wireless Communication System

The communication flow block diagram of the experimental platform is shown in Figure 6. The AES-T2100 Trojan is inserted into the FPGA of USRP in the functional design phase, and this USRP is used as the RF transmitter. The USRP is connected to the host computer via a Gigabit Ethernet cable. With the software radio operating on the host computer, we can easily change multiple communication standards on this experimental platform to simulate the transmitter equipment in real communication networks.

Another Trojan-free USRP is used to receive the signal. The receiver is also connected to the host computer. The computer is used to sort and identify the received signal to detect the Trojan and evaluate the detection method.

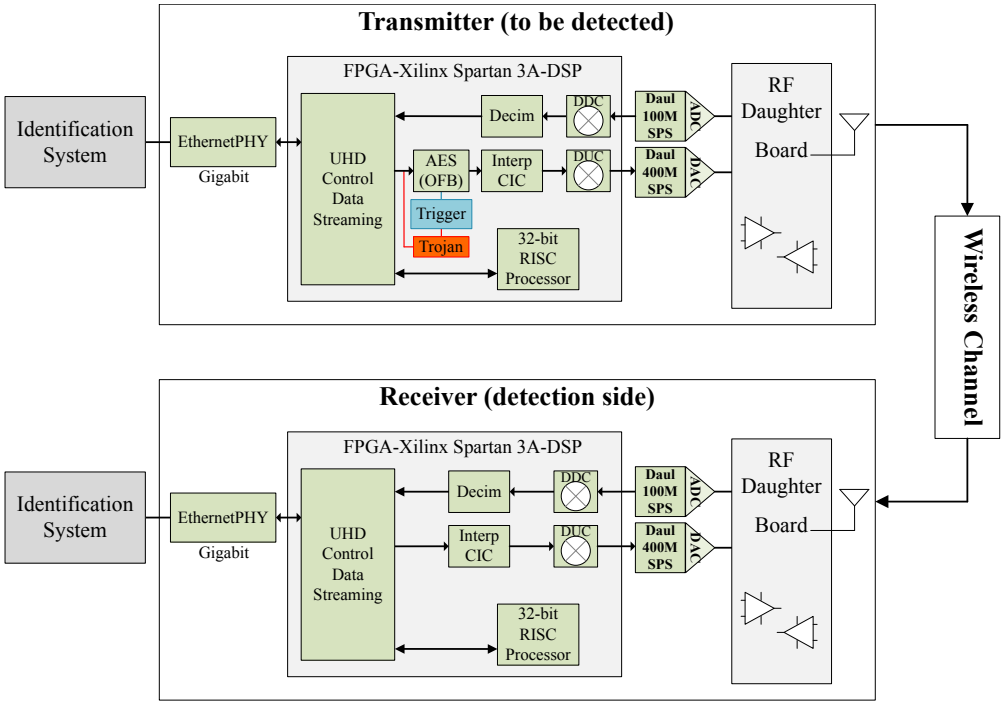


Figure 6. Trojan detection procedure along wireless communication flow block diagram.

4.2. Dataset Generation

The received signal needs to be preprocessed before RFF recognition. On the detection side, the collected RF signal is first preprocessed by synchronization and framing to obtain the signal with the same frame structure as the transmitter. Then, the RFF is performed on the framed signal. The preamble sequence in the data frame structure is searched by the correlation operation in the following

$$R_{pr}(m) = \sum_{n=-\infty}^{+\infty} p(n) * r(n+m) \tag{10}$$

where $p(n)$ is the preamble sequence, and $r(n)$ is the received time-domain sequence. m is taken as the range where there is an overlap between the two sequences after time shifting. The calculation result $R_{pr}(m)$ is the correlation sequence.

The maximum points in the correlation sequence are the candidate positions for the preamble. Furthermore, it should be confirmed whether the interval between the maximum points is equal to the frame length of the transmit signal. After filtering by the previous two criteria, the framing position of the original signal is obtained from the correlation sequence. However, for the modified target benchmark circuit, the preamble sequences in the transmit signal are encrypted using AES in OFB mode. The segmentation between frames becomes confused, and the Trojan-free signal and the Trojan-infested signal are then framed according to the reliable framing position of the original signal at the same time reference. The final obtained RF signal after framing is defined as $x(n)$.

As mentioned in Section 2, the implicit load Trojan does not change the original function, but adds extra logic to compromise the circuit integrity. In this experimental platform, detection of the implicit Trojan AES-T2100 is used as an example to verify the proposed method based on RFF.

In order to identify the status of Trojans, the proper RFF should be extracted to indicate the distinction between circuits with subtle deviations. A dataset containing different transmit signals is generated. Both the QPSK and triangular-wave signals are measured. In total, more than 30,000 frames are collected for the dataset to identify the presence and status of Trojans in the transmitter. In the experiment, the Trojan-free FPGA system and the Trojan-infected FPGA system are burned into two USRP N210 devices, and the transmitters

are then controlled to start communication. After that, the RF signals are received at the detection side, and the pre-processed golden reference signal frames and the signal frames infected with Trojan horse are stored in the dataset.

According to the Trojan-triggering process, the Trojan circuit works in the dormant and active periods sequentially during the signal reception time. The signals are then divided into signals from the transmitters with inactive Trojan and signals from the transmitter with active Trojan. In this paper, these two types of signals are labeled as the HT-dormant signal and the HT-active signal, respectively.

5. Results and Discussion

In order to evaluate the proposed method, two kinds of modulated signals commonly applied in wireless communication are used for sorting and identification. The signals transmitted from the Trojan-free USRP and the USRP with dormant and active Trojans are identified using the proposed method based on Bi-LSTM.

5.1. Evaluation

The proposed detection method was evaluated in terms of both effectiveness and time consumption. Accuracy, true-positive rate (TPR), false-positive rate (FPR), precision, and recall are used as evaluation metrics. There are several indices for the detection method performance: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In the case of Trojan presence detection, TP refers to the number of correctly classified Trojan-free signals. TN refers to the number of signals with Trojan correctly classified. TPR is defined by TP/(TP+FN). FPR is defined by FP/(FP+TN).

The performance of the proposed method is compared with some of the latest techniques based on RNN, LSTM, and MLP in Table 2. All experiments are performed on a computer equipped with 32G memory, an Intel i7-12900 CPU and an NVIDIA GTX 3070 GPU. The received signals are fed directly into the identification network, which is an end-to-end identification system. All networks are trained using batch-gradient descent with a batch size of 128.

Table 2 shows that the detection accuracy of MLP model is extremely poor. It indicates that MLP is not suitable for distinguishing Trojan signals by learning the features from the complete signals. The general RNN and LSTM can significantly improve the classification performance of signals with Trojan by learning the features of the time-series signals. Compared with MLP, the improvements in RNN and LSTM likely stem from the large number of parameters they have learned. The accuracy of Bi-LSTM is relatively high while increasing the number of parameters. However, the 0.35M parameters of Bi-LSTM are acceptable when compared with MobileNet(4.2M) [27] and VGG16(138M) [28] models implemented on the FPGA platform. Although the number of Bi-LSTM model parameters is 4.5 times greater than LSTM, it is still competitive.

The Bi-LSTM provides the highest accuracy in Table 2. The average accuracy of Bi-LSTM is 98.9%, and the detection error rate of the Trojan is 1.625%. Compared with LSTM, RNN, and MLP, the accuracy of Bi-LSTM is 1.1–53% more accurate. As Bi-LSTM can learn both the forward and backward features within each frame, it performs quite well for distinguishing devices with dormant Trojans and active Trojans.

Table 2. Comparison between different methods.

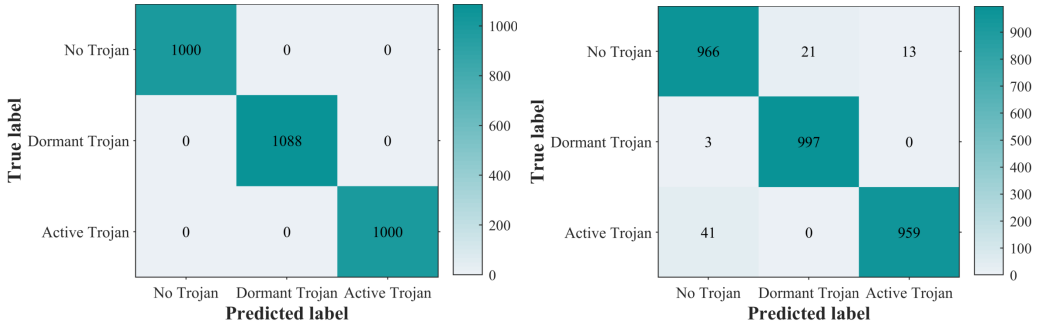
Model	Params $\times 10^6$	Transmission Signal	Accuracy	TPR	FPR
MLP	0.003075	Triangular wave	42.8%	11.8%	8.95%
		QPSK	46.2%	76.5%	52.3%
LSTM	0.078723	Triangular wave	95.4%	94.3%	4.1%
		QPSK	97.5%	99.9%	0.25%
RNN	0.028035	Triangular wave	84.8%	80.9%	8.25%
		QPSK	85.7%	78.6%	54.45%
Bi-LSTM	0.351875	Triangular wave	95.9%	94.1%	3.25%
		QPSK	99.2%	100%	0%

The proposed method can be used to detect the absence of the Trojan circuit in order to focus on high-risk devices. The identification results are shown in Table 3. It can be seen that the identification accuracy for QPSK signals is higher than triangular-wave signals, but both perform quite well. In addition, the detection FPR of 0 for QPSK signals indicates that the proposed method correctly detects all Trojan samples including HT-dormant signal and HT-active signal. Thus, the proposed method can also be used to recognize low-risk devices and high-risk devices. For a test set with a size of 3000, the proposed method completes the detection of triangular-wave transmission signals and QPSK transmission signals in 0.030 and 0.417 seconds, respectively. The detection of QPSK signals with a frame length of 1000 points requires substantially more time than the detection of triangular-wave signals with a frame length of 128 points due to the high complexity of the Bi-LSTM algorithm. The detection time-consumption indicates that the proposed method can achieve faster Trojan detection for short-length input signals.

Table 3. Detection performance of the absent Trojan.

Transmission Signal	TPR	FPR	Precision	Recall	Time(s)
Triangular wave	96.6%	2.2%	97.4%	97.4%	0.030
QPSK	100%	0%	100%	100%	0.417
Total	98.3%	1.1%	98.7%	98.7%	0.224

In order to evaluate the accuracy of the proposed method for detecting the active status of the Trojan circuit, confusion matrices for classifying the Trojan-free signal, HT-dormant signal, and HT-active signal are shown in Figure 7. Figure 7a shows the classification error for QPSK signals. It can be seen that the proposed method can recognize the status of Trojan in transmitters of QPSK signals quite well, and the accuracy can be achieved as high as 100%. The classification error of the triangular wave signal is shown in Figure 7b, and the diagonal data indicate the number of correctly classified samples. For the Trojan-free signal, there are 21 samples in the classification that are misclassified as the HT-dormant signal. For HT-dormant signals and HT-active signals, there are 3 and 41 frames are misclassified as Trojan-free signals, respectively. It is noticed that there is no confusion regarding the HT-dormant signal and the HT-active signal. Therefore, the proposed method can be utilized to precisely monitor the activity status of the Trojan for both QPSK signals transmitters and triangular-wave transmitters. If the recognition network model is embedded in the FPGA in the receiver, real-time detection for HT can be achieved.



(**a**) (**b**)

Figure 7. The classification confusion matrices. (**a**) Confusion Matrix of QPSK Signals. (**b**) Confusion Matrix of triangular-wave signals.

The detection accuracy of the Trojan status is also calculated and listed in Table 4. As Table 4 shows, the accuracy of Trojan-state detection is 100% for both QPSK signals and triangular-wave signals. The proposed method can identify the activated and inactivated state of the Trojan accurately, so it can detect the moment when the Trojan is triggered in a timely mamnner, and can accurately trigger subsequent warning and contingency measures. Even though the proposed method has an FPR of 1.1% for detecting Trojan-infected circuits, the method can compensate for previously performed misclassifications by continuously operating to monitor the activation of the Trojan load in the IC.

Table 4. Detection accuracy of the Trojan status.

Status of Trojan	Dormant Trojan	Active Trojan	Total
Accuracy for QPSK signals	100%	100%	100%
Accuracy for triangular wave signals	100%	100%	100%

Some Trojan circuits are triggered after a longer period to assist in concealing their presence. Even some devices with Trojans are incorrectly identified as low-risk devices, because the Trojan is not activated. Moreover, there is usually a lack of golden reference, so the methods for identifying the status of Trojans are highly practical. Fortunately, the proposed model can discriminate between HT-dormant signals and HT-active signals well. Even though the proposed method cannot distinguish the presence of Trojan horses in the circuit perfectly, it can capture the triggering of Trojan perfectly by long-term observation. Consequently, the harm to the Trojans is reduced.

5.2. Discussion

In this paper, the RFF-based HT-detection method is evaluated and verified by identifying the measured signals. The experimental results demonstrate that the proposed method performs effectively in detecting both dormant Trojan circuits and active Trojan circuits. Significantly, the proposed method can detect triggering of the Trojan circuit in time to immediately reduce any damage caused by the Trojan.

The proposed method is compared with the related works. Table 5 shows a comparison of the proposed method, the machine-learning-based detection method at register transfer level (RTL), the Hierarchical Temporal Memory (HTM) architecture, the co-training-based detection method, and the learning-assisted side-channel delay analysis (LASCA) methodology [2,18,29,30]. The detection method at RTL is based on circuit features extracted from the Trojan source code, so its TPR values can depend heavily on the known benchmark Trojan circuits [2]. However, it is highly difficult to learn the operating mechanism of HTs in practical applications. On the other hand, the circuit source code is proportional to its design. Therefore, the detection cost is proportional to the circuit size. Nevertheless, even for large circuits, the proposed method detects the Trojan horse only through features of the

wireless signal. It is proved that the complexity cost of the proposed method is independent of the circuit size.

The HTM method eliminates the need for a gold reference, and has a short-detection-time delay [18]. Table 5 shows that the proposed method outperforms the HTM method in terms of detection accuracy. Compared with the co-training-based detection method, the proposed method has fewer constraints on the requirement of signal acquisition equipment. In Table 5, some compared methods detect the HTs based on the side-channel information of power or path delay measurement. The acquisition of these side channel signals requires high-precision acquisition equipment. Moreover, for delay information, calculating all possible path delays of an IC presents a time-consuming task. Meanwhile, the huge amount of data acquired leads to a model training time of 5 h for some methods, such as the LASCA. On the other hand, the proposed method does not require additional acquisition equipment for detecting wireless encrypted ICs. It is also easy to integrate into existing communication tests because it only uses the RF signals provided by the receiver to detect the HTs. The proposed method can detect Trojans in the operating circuit, and can detect changes in the status of the Trojan circuit in time for shorter transmit signals (such as 128 points).

Table 5. Comparison to the existing methods.

Method	Test Modality	Accuracy	FPR	Recall	Time(s)
Ours	Wireless transmission channel	98.7%	1.1%	98.7%	0.224
[2]	RTL code	-	0%	94.94%	1.107685
[18]	Power	92.2%	-	-	0.0072
[29]	Power	93.4%	-	94.7%	-
[30]	Path delay	87.5%	0.15%	-	18,000

- means missing the value in their work.

Similar to most side-channel-based detection methods, the proposed detection method based on RFF assumes that there is one trusted Trojan-free device. It should be noted that it is difficult to obtain trusted Trojan-free devices in many applications. However, in the continuous observation of the working circuit, the Trojan causes a transient effect on the circuit whenever it is triggered. This effect can be observed without the golden chip. Thus, for Trojans with triggers, it is easy to capture the moment of their activation. In practice, we often have no knowledge of the triggering mechanism of Trojan circuits. Therefore, we cannot guarantee that the signals of active Trojans and dormant Trojans can be accurately collected. Hence, it is an effective and feasible method to identify the presence of Trojans by monitoring the circuit.

6. Conclusions

It is widely known that the use of near-field EM side-channel information is highly successful in detecting HTs, but the near-field information is not always easy to obtain. In this paper, a non-destructive method for Trojan detection in wireless encryption ICs is proposed, which employs far-field wireless signals. By extracting and identifying the RFF of the received signals, the presence and status of Trojans can be identified. The proposed method provides a feasible solution for the remote detection of HT. It does not require the gate-level netlists of the IC design, saving on detection costs. The experimental results show that the proposed method can effectively detect the presence and status of the AES-T2100 Trojan in wireless ICs. This presents a successful attempt at using RFF for HT detection. The proposed method can be easily integrated into a current radio device test flow without the need for additional devices.

Author Contributions: Conceptualization, S.M. and Z.Z.; Data curation, Z.Z.; Formal analysis, S.M. and Z.Z.; Funding acquisition, S.M.; Investigation, Z.Z.; Methodology, S.M. and Z.Z.; Project administration, S.M.; Resources, S.M.; Software, S.M., Z.Z. and Y.Z.; Supervision, S.M.; Validation, S.M., Z.Z., Y.Z. and A.H.; Visualization, S.M. and Z.Z.; Writing—original draft, S.M. and Z.Z.; Writing—review and editing, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2018AAA0100104, 2018AAA0100100), Natural Science Foundation of Jiangsu Province (BK20211164).

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by Zhishan Scholar Program of Southeast University, and the Purple Mountain Laboratories, Nanjing, China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sadhu, P.K.; Yanambaka, V.P.; Abdelgawad, A. Internet of Things: Security and Solutions Survey. *Sensors* **2022**, *22*, 7433. [CrossRef] [PubMed]
2. Han, T.; Wang, Y.; Liu, P. Hardware trojans detection at register transfer level based on machine learning. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; pp. 1–5.
3. Hasegawa, K.; Yanagisawa, M.; Togawa, N. Trojan-feature extraction at gate-level netlists and its application to hardware-Trojan detection using random forest classifier. In Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 28–31 May 2017; pp. 1–4.
4. Dong, M.; Pan, W.; Qiu, Z.; Gao, Y.; Qi, X.; Zheng, L. An Efficient Framework with Node Filtering and Load Expansion for Machine-Learning-Based Hardware Trojan Detection. *Electronics* **2022**, *11*, 2054. [CrossRef]
5. Liakos, K.G.; Georgakilas, G.K.; Plessas, F.C.; Kitsos, P. GAINESIS: Generative Artificial Intelligence NEtlists SynthesIS. *Electronics* **2022**, *11*, 245. [CrossRef]
6. Wolff, F.G.; Papachristou, C.A.; Bhunia, S.; Chakraborty, R.S. Towards Trojan-Free Trusted ICs: Problem Analysis and Detection Scheme. In Proceedings of the 2008 Design, Automation and Test in Europe (DATE), Munich, Germany, 10–14 March 2008; pp. 1362–1365.
7. Wang, S.J.; Wei, J.Y.; Huang, S.H.; Li, K.S.M. Test generation for combinational hardware Trojans. In Proceedings of the 2016 IEEE Asian Hardware-Oriented Security and Trust (AsianHOST), Yilan, Taiwan, 19–20 December 2016; pp. 1–6.
8. Deyati, S.; Muldrey, B.J.; Chatterjee, A. Targeting hardware trojans in mixed-signal circuits for security. In Proceedings of the 2016 IEEE 21st International Mixed-Signal Testing Workshop (IMSTW), Sant Feliu de Guixols, Spain, 4–6 July 2016; pp. 1–4.
9. Agrawal, D.; Baktir, S.; Karakoyunlu, D.; Rohatgi, P.; Sunar, B. Trojan detection using IC fingerprinting. In Proceedings of the 2007 IEEE Symposium on Security and Privacy (SP'07), Berkeley, CA, USA, 20–23 May 2007; pp. 296–310.
10. Jin, Y.; Makris, Y. Hardware Trojan detection using path delay fingerprint. In Proceedings of the 2008 IEEE International Workshop on Hardware-Oriented Security and Trust (HOST), Anaheim, CA, USA, 9 June 2008; pp. 51–57.
11. Narasimhan, S.; Du, D.; Chakraborty, R.S.; Paul, S.; Wolff1, F.; Papachristou, C.; Roy, K.; Bhunia, S. Multiple-parameter side-channel analysis: A non-invasive hardware Trojan detection approach. In Proceedings of the 2010 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST), Anaheim, CA, USA, 13–14 June 2010; pp. 13–18.
12. Huang, Y.; Bhunia, S.; Mishra, P. Scalable Test Generation for Trojan Detection Using Side Channel Analysis. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2746–2760. [CrossRef]
13. Rad, R.; Plusquellic, J.; Tehranipoor, M. Sensitivity analysis to hardware Trojans using power supply transient signals. In Proceedings of the 2008 IEEE International Workshop on Hardware-Oriented Security and Trust (HOST), Anaheim, CA, USA, 9 June 2008; pp. 3–7.
14. Rooney, C.; Seeam, A.; Bellekens, X. Creation and detection of hardware trojans using non-invasive off-the-shelf technologies. *Electronics* **2018**, *7*, 124. [CrossRef]
15. He, J.; Zhao, Y.; Guo, X.; Jin, Y. Hardware Trojan Detection Through Chip-Free Electromagnetic Side-Channel Statistical Analysis. *IEEE Trans. Very Large Scale Integr. Syst.* **2017**, *25*, 2939–2948. [CrossRef]
16. He, J.; Ma, H.; Guo, X.; Zhao, Y.; Jin, Y. Design for EM Side-Channel Security through Quantitative Assessment of RTL Implementations. In Proceedings of the 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, China, 13–16 January 2020; pp. 62–67.
17. Salmani, H. COTD: Reference-Free Hardware Trojan Detection and Recovery Based on Controllability and Observability in Gate-Level Netlist. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 338–350. [CrossRef]
18. Faezi, S.; Yasaei, R.; Barua, A.; Faruque, M.A.A. Brain-Inspired Golden Chip Free Hardware Trojan Detection. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 2697–2708. [CrossRef]
19. Jin, Y.; Makris, Y. Hardware Trojans in wireless cryptographic ICs. *IEEE Des. Test Comput.* **2010**, *27*, 26–35. [CrossRef]

20. Liu, Y.; Jin, Y.; Nosratinia, A.; Makris, Y. Silicon Demonstration of Hardware Trojan Design and Detection in Wireless Cryptographic ICs. *IEEE Trans. Very Large Scale Integr. Syst.* **2017**, *25*, 1506–1519. [CrossRef]
21. Shakya, B.; He, T.; Salmani, H.; Forte, D.; Bhunia, S.; Tehranipoor, M. Benchmarking of Hardware Trojans and Maliciously Affected Circuits. *J. Hardw. Syst. Secur.* **2017**, *1*, 85–102. [CrossRef]
22. Chakraborty, R.S.; Narasimhan, S.; Bhunia, S. Hardware Trojan: Threats and emerging solutions. In Proceedings of the 2009 IEEE International High Level Design Validation and Test Workshop (HLDVT), San Francisco, CA, USA, 4–6 November 2009; pp. 166–171.
23. Wang, W.; Sun, Z.; Piao, S.; Zhu, B.; Ren, K. Wireless Physical-Layer Identification: Modeling and Validation. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 2091–2106. [CrossRef]
24. Peng, L.; Hu, A.; Zhang, J.; Jiang, Y.; Yu, J.; Yan, Y. Design of a Hybrid RF Fingerprint Extraction and Device Classification Scheme. *IEEE Internet Things J.* **2019**, *6*, 349–360. [CrossRef]
25. Sun, L.; Ke, D.; Wang, X.; Huang, Z.; Huang, K. Robustness of Deep Learning-Based Specific Emitter Identification under Adversarial Attacks. *Remote Sens.* **2022**, *14*, 4996. [CrossRef]
26. Trust-Hub. Available online: https://www.trust-hub.org/ (accessed on 27 August 2022).
27. Liao, J.; Cai, L.; Xu, Y.; He, M. Design of Accelerator for MobileNet Convolutional Neural Network Based on FPGA. In Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; Volume 1, pp. 1392–1396.
28. Qiu, J.; Wang, J.; Yao, S.; Guo, K.; Li, B.; Zhou, E.; Yu, J.; Tang, T.; Xu, N.; Song, S.; et al. Going Deeper with Embedded FPGA Platform for Convolutional Neural Network. In Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, USA, 21–23 February 2016; ACM: New York, NY, USA, 2016; pp. 26–35.
29. Xue, M.; Bian, R.; Wang, J.; Liu, W. Building an accurate hardware Trojan detection technique from inaccurate simulation models and unlabelled ICs. *IET Comput. Digit. Tech.* **2019**, *13*, 348–359. [CrossRef]
30. Vakil, A.; Behnia, F.; Mirzaeian, A.; Homayoun, H.; Karimi, N.; Sasan, A. LASCA: Learning Assisted Side Channel Delay Analysis for Hardware Trojan Detection. In Proceedings of the 2020 21st International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, USA, 25–26 March 2020; pp. 40–45.