Received October 31, 2020, accepted January 17, 2021, date of publication January 29, 2021, date of current version February 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055493

Behavioral and Physical Unclonable Functions (BPUFs): SRAM Example

MIGUEL A. PRADA-DELGADO[1] **AND ILUMINADA BATURONE**[2]
[1]IBM Research, 8803 Zurich, Switzerland
[2]Instituto de Microelectrónica de Sevilla (IMSE-CNM), Universidad de Sevilla, CSIC, 41092 Seville, Spain

Corresponding author: Miguel A. Prada-Delgado (prm@zurich.ibm.com)

This work was supported in part by FEDER/Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación/_TEC2017-83557-R and _RTC-2017-6595-7, and in part by Consejería de Economía, Conocimiento, Empresas y Universidad de la Junta de Andalucía under Project AT17_5926_USE and Project US-1265146.

ABSTRACT Physical Unclonable Functions (PUFs) have gained a great interest for their capability to identify devices uniquely and to be a lightweight primitive in cryptographic protocols. However, several reported attacks have shown that virtual copies (mathematical clones) as well as physical clones of PUFs are possible, so that they cannot be considered as tamper-resistant or tamper-evident, as claimed. The solution presented in this article is to extend the PUFs reported until now, which are only physical, to make them Behavioral and Physical Unclonable Functions (BPUFs). Given a challenge, BPUFs provide not only a physical but also a behavioral distinctive response caused by manufacturing process variations. Hence, BPUFs are more difficult to attack than PUFs since physical and behavioral responses associated to challenges have to be predicted or cloned. Behavioral responses that are obtained from several measurements of the physical responses taken at several sample times are proposed. In this way, the behavioral responses can detect if the physical responses are manipulated. The analysis done for current PUFs is extended to allow for more versatility in the responses that can be considered in BPUFs. Particularly, Jaccard instead of Hamming distances are proposed to evaluate the similarity of behavioral responses. As example to validate the proposed solution, BPUFs based on Static Random-Access Memories (SRAM BPUFs), with one physical and one behavioral responses to given challenges, were analyzed experimentally using integrated circuits fabricated in a 90-nm CMOS technology. If an attacker succeeds in cloning the physical responses as reported, but does not attack the way to obtain the behavioral responses, the attacker fails on SRAM BPUFs. The highest probability to succeed in cloning the behavioral responses with a brute-force attack was estimated from experimental results as $1.5 \cdot 10^{-34}$, considering the influence of changes in the operating conditions (power supply voltage, temperature, and aging).

INDEX TERMS Hardware security, multimodal biometrics, physical unclonable functions, SRAM.

I. INTRODUCTION

The protection of information is of crucial importance, especially when dealing with sensitive data. To achieve a considerable degree of protection, information security has to be conceived from the design of the cryptographic algorithms until its implementation into cryptographic circuits [1]. It is at this time when the creation of a secret key, its storage, and use are especially critical. Military communications are an example of critical applications in which the highest level of security is required. Tampering, which consists in permanently manipulating an entity with the objective of carrying out an unauthorized operation, should be particularly avoided in the case of cryptographic circuits [2]. Multiple solutions to improve anti-tampering were proposed, most of them focused on specific watermarking designs that prove the intellectual property rights of the producers and owners of the chips [3]. Other solutions focus on generating tamper evidences and tamper resistances against attacks [4].

In 2002, Pappu *et al.* introduced a new type of tamper-resistant one-way functions called *physical one-way functions* [5], which were later named *physical unclonable functions* (PUFs) after the article of Gassend *et al.* in the same year [6]. A physical unclonable function (PUF) is a physical construction that exploits the variations produced during the manufacturing process to generate unique responses

The associate editor coordinating the review of this manuscript and approving it for publication was Tiago Cruz.

(or outputs) to given challenges (or inputs). Due to the uncontrollable nature of manufacturing process variations, each manufactured instance of a PUF can be identified conveniently by the unique challenge-response pairs of the PUF. Therefore, if the variations are not controllable, physical unclonability results from the impossibility to create two instances that, given the same challenges, provide similar responses. The tamper resistance provided by PUFs are based on the impossibility to modify a manufactured PUF so that it could continue working and providing responses in a different way to its intrinsic nature.

Some of the PUFs that have been studied in greater depth are the electronic PUFs, and among them, those which predominate in the state of the art are memory-based (such as PUFs based on static random access memories, SRAM PUFs) [7] and delay-based (such as the so-called arbiters [6] and PUFs based on ring oscillators, RO PUFs [8]). In all of them, challenges and responses are binary. The challenges are the binary vectors that address the set of two theoretically identical constructions that are the basic unit of the PUF. For example, the challenges of SRAM PUFs address the bit memory cells considered; the challenges of RO PUFs address the pairs of ring oscillators to compare; the challenges of arbiter PUFs address the pairs of paths to evaluate, and so on. The achievement or not of a physical condition is evaluated to obtain a binary response.

If the relationship between the number of challenge-response pairs with the number of PUF basic units (for example, the delay components in arbiter PUFs or the memory cells in SRAM PUFs) is small, the PUFs are called weak PUFs, while those that provide a large relationship are called strong PUFs. Similar to the definition of a one-way function in cryptography, which can be seen in [9], it is considered that a PUF is infeasible to invert (in an average-case sense). That is, any feasible algorithm that tries to find the challenges associated to the responses may succeed only with negligible probability (where the probability is taken uniformly over the choices of the challenges and the algorithm's coin tosses). Due to this fact, mathematical unclonability is another property of PUFs because a virtual copy of the PUF (its challenge-response set) is infeasible to find.

The properties of PUFs have been exploited to identify devices uniquely (like a biometry for devices), which in turn have been employed in several lightweight authentication protocols [7], [8], [10], [11]. Concerning challenges and responses, human biometrics is similar to weak PUFs, since, for example, humans have at most 10 different fingerprints in their hands.

However, several attacks have been reported on PUFs, which call into question the security of the proposals made to date [12]. The work in [13] shows that machine learning techniques, based on Artificial Neural Networks and Evolution Strategies, applied to challenge-response pairs and additive linear models are successful to generate virtual copies of several arbiter and ring oscillator PUFs. More recently, the work in [14] proposes a general framework for machine learning attacks on strong PUFs and presents two Artificial Neural Network structures to approximate their challenge-response pairs, particularly of multiplexer-based PUFs and XOR arbiter PUFs. Regarding memory-based PUFs, there have been shown that bias [15] as well as spatial correlation [16] exist in many SRAM PUF conventional architectures, which makes them predictable (thus mathematically clonable) to some extent. In addition, using optical semi-invasive attacks from the chip backside (photonic emission analysis, laser fault injection, and optical contactless probing), the work in [17] demonstrates that the responses generated by a PUF can be predicted, manipulated and directly probed without affecting the behavior of the PUF, so that they cannot be considered as tamper-evident or tamper-resistant. This is demonstrated also in [18] by using laser stimulation for semi-invasive, backside, single-trace readout of logic states in SRAMs. Moreover, the works in [19], [20] show that SRAM PUFs can be not only fully characterized and emulated but also cloned physically. They used a Focused Ion Beam circuit edit and produced a fully-functional second instance with identical responses of a first instance SRAM PUF. Moreover, the works in [21] and [22] show that hybrid strategies that combine side-channel attacks with machine learning techniques based on Evolution Strategies are successful even if the attacker does not have direct access to the challenge-response pairs.

The above mentioned attacks reveal the need to improve the security of reported PUFs. The solution proposed in this article is to add another layer of security by making PUFs multimodal in the same way as multimodal biometric systems improve the security of unimodal biometric systems against attacks. Multimodal biometrics employs two or more biometric characteristics of the same individual instead of a single one. In the same line, the PUFs proposed herein allows for multimodal authentication, thus increasing the security of the PUFs reported till now, which only allow for unimodal authentication.

Unimodal biometric systems usually employ physical innate human features such as fingerprints, vein patterns etc. Recently, multimodal biometric systems also use behavioral features such as the electric activity generated by the heart and measured by electrocardiograms or the blood volume change in veins measured by photoplethysmography. Behavioral biometrics is interesting because it provides proofs of liveness. That is, they can detect from several measurements taken at several sample times if the subject is alive or lifeless (the electric activity generated by the heart changes or not over time, the blood volume in veins changes or not over time, and so on). This avoids typical attacks at unimodal physical systems like fake silicone or gelatin fingerprints in the case of fingerprint recognition. In the same line, the multimodal PUFs proposed herein exploit inherent behavioral and physical features of the manufactured devices that are caused by manufacturing process variations. Hence, they will be referred to as BPUFs (Behavioral and Physical Unclonable Functions).

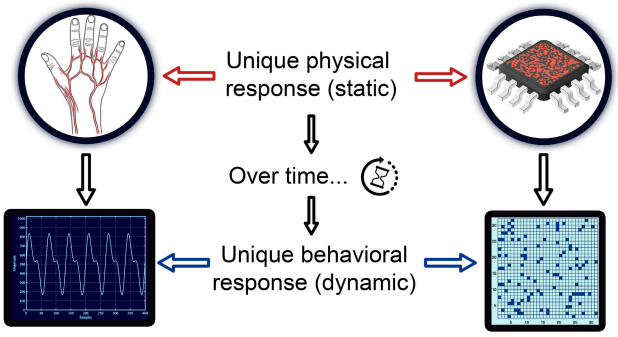


FIGURE 1. Parallelism between electronic BPUFs (on the right) and multimodal biometric systems based on veins (on the left).

The main features of the proposed BPUFs are:

- They exploit the variations produced during the manufacturing process to generate not only one unique response but two or more unique responses to one given challenge. Hence, while a PUF is defined by its challenge-response pairs, a BPUF is defined by its challenge-response tuples (triplets in the simplest case).
- A BPUF provides not only a physical response that results from one measurement taken at a given sample time, but also a behavioral response that results from several measurements taken at several sample times.
- The attacks to BPUFs are more complex than to PUFs since, on the one side, more responses have to be predicted or cloned for an arbitrary challenge and, on the other side, the behavioral responses proposed can detect if the physical responses are alive (they show changes over time) or lifeless (they are always the same). In this sense, we will say that the proposed BPUFs are able to detect liveness.
- The constructions currently employed for electronic PUFs can be used in general for BPUFs. For example, BPUFs based on SRAMs are illustrated in detail in this article.

Figure 1 illustrates the parallelism between electronic BPUFs and multimodal biometric systems that analyze the veins as challenges and provide, as a static response, a result from analyzing the vein patterns, and, as a dynamic response, a result from analyzing the change over time in the blood volume of the veins.

The article is structured as follows. Section II summarizes the main properties and metrics widely employed in current unimodal PUFs. Section III introduces BPUFs, describing their features and the most adequate metrics to evaluate them. Section IV describes how BPUFs increase security compared with unimodal PUFs. Section V provides experimental results obtained from SRAM BPUFs that confirm the models and proposals given in Sections III and IV. Finally, conclusions are given in Section VI.

II. BACKGROUND

An instance of a unimodal PUF provides a unique response vector u_x to a given challenge vector x. Binary challenges and responses are used in the most well-known PUFs so that u_x and x are considered herein as vectors with 1's and 0's. Since several measurements can be made of the responses, the i-th measurement of the response u_x will be denoted herein as u_x^i.

A PUF should feature high reproducibility. It is achieved if the response vectors of a given PUF instance measured multiple times for the same challenge vector are very similar. A PUF must also feature high uniqueness, which means that for the same challenge vector, the response vectors of two different instances must be very dissimilar regardless the measurement. In addition, the response vectors of a PUF must be unpredictable because there should be no method to predict the response provided by an instance to a new challenge.

A. RESPONSE SIMILARITY: HAMMING DISTANCE

Similarity between PUF responses is measured with the Hamming distance. The Hamming distance (HD) between two binary vectors of the same length, N, is calculated as the number of changes that are needed to convert one vector into the other. It is calculated by XORing the bits of the vectors as follows:

$$HD(u, v) = \sum_{b=0}^{N-1} (u[b] \oplus v[b]) \tag{1}$$

where $u[b]$ represents the b-th bit of the vector u.

The fractional Hamming distance (FHD) is calculated as $FHD = HD/N$.

Hamming distances evaluated between responses of the same instance to the same challenges at different measurements (genuine population) are called intra Hamming distances. Intra Hamming distances of a PUF instance with perfect reproducibility are zero. However, perfect reproducibility does not happen since there are always some response bits that change from one measurement to another (known as flipping bits), which generates noise.

Hamming distances evaluated between responses of different instances to the same challenges (impostor population) are called inter Hamming distances. If the number of 1's and 0's in the PUF responses are the same (unbiased responses) and their positions are perfectly random, the average inter FHD is 0.5, as will be seen in Section III.

B. AUTHENTICATION BASED ON PUFs

Authentication of PUF-based systems requires two operation phases: registration and verification. During registration, a measurement of the challenge-response set is taken. Let us denote the registered response to the challenge x as u_x^0. During verification, another measurement of the response, u_x^i is taken. Ideally, the responses of genuine PUF instances are similar to the registered response due to reproducibility property and impostor responses are very different due to uniqueness property. Hence, the instance is verified if the distance (considered as the number of errors) between the responses is below a threshold, $HD(u_x^i, u_x^0) \leq HD^{max}$.

In [23], Bösch *et al.* modeled errors in the PUF response as a random variable E of N independent bits that behaves as a binary symmetric channel. This is demonstrated experimentally in [24] and [25] for the case of SRAM PUFs. The probability distribution selected to model the occurrence of exactly t bit errors in the N bits of E is a binomial distribution, as follows:

$$P(E=t)=\binom{N}{t}p_e^t(1-p_e)^{N-t} \tag{2}$$

where p_e is the bit error probability, which is the same for a 0 that changes to 1, and for a 1 that changes to 0 (because it is assumed a symmetric model).

Under these circumstances, the probability that the genuine response contains more than HD^{max} errors (false rejection due to maximum errors threshold) is given by:

$$P_G\left(E>HD^{max}\right)=1-\sum_{i=0}^{HD^{max}}\binom{N}{i}p_{eG}^i(1-p_{eG})^{N-i} \tag{3}$$

The bit error probability in the genuine responses, p_{eG}, can be estimated experimentally as the bit error rate (*BER*) considering genuine responses or as the maximum intra *FHD* (if the worst case is considered). The bit error rate (*BER*) in a set of R responses, $\{u_x^0,\ldots,u_x^{R-1}\}$, is calculated as the average ratio of the number of bit errors in the total number of bits, N. The *BER* represents the average *FHD* ($\overline{FHD}$) as follows:

$$\begin{aligned}BER(u_x^0,\ldots,u_x^{R-1})&=\overline{FHD}(u_x^0,\ldots,u_x^{R-1})\\&=\frac{2}{R(R-1)}\sum_{i=0}^{R-2}\sum_{j=i+1}^{R-1}\frac{HD(u_x^i,u_x^j)}{N}\end{aligned} \tag{4}$$

BER is normalized by the number of *FHDs* that can be calculated using R responses, which is equal to $R\cdot(R-1)/2$.

Once p_{eG} is estimated, the threshold HD^{max} can be selected, using Equation 3, so as to ensure a small false rejection rate, $P_G(E>HD^{max})<\epsilon$, with $\epsilon=10^{-6}$, for example.

III. PROPOSAL OF BPUFs

A BPUF generates two or more reproducible, unique and unpredictable responses to given challenges, exploiting the variations produced during the manufacturing process. Hence, BPUFs evaluate more distinctive features than unimodal PUFs. In addition, the distinctive features evaluated are not only physical but also behavioral.

As summarized in the Introduction, electronic PUFs contain a set of units made of two theoretically identical constructions. SRAM PUFs contain a set of memory cells (each cell with two theoretically identical inverters), RO PUFs contain a set of RO pairs (each pair with two theoretically identical ROs), and so on. The physical condition evaluated in the memory cell of an SRAM PUF is if one of the two theoretically identical inverters wins or not at power up so as to impose a logic 1 or 0 in the corresponding bit of the response. In the case of an RO PUF, the bit of the response is 1 or 0 if the difference between the oscillation frequencies of the first and the second ring oscillators in the pair considered is positive or negative.

In general, we can say that the b-th bit of the BPUF physical response at measurement i to a challenge x that addresses the unit b is:

$$u_x^i[b]=\begin{cases}1 & \text{if unit } b \text{ meets a physical condition}\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

In general, the b-th bit of the BPUF behavioral response at measurement i to a challenge x that addresses the unit b is:

$$\vartheta_x^i[b]=\begin{cases}1 & \text{if unit } b \text{ meets a behavioral condition}\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

Concerning behavioral features, let us focus on features that are evaluated with several measurements of a given physical response. The b-th bit of the response to a challenge x associated with a behavioral feature of responses u_x at measurement i is as follows:

$$\vartheta_x^i[b]=\begin{cases}1 & \textit{if } \{u_x^1[b],\ldots,u_x^R[b]\} \text{ meet a condition}\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

For example, BPUFs based on ROs can evaluate if the frequency difference of a pair of ring oscillators is always positive in several measurements or not. Similarly, BPUFs based on SRAMs can evaluate if the start-up value of a cell is always the same in several measurements or not (the start-up value shows or not bit flipping). In these cases, Equation 7 is particularized as follows:

$$\vartheta_x^i[b]=\begin{cases}1 & \textit{if } \sum_{j=1}^{R}\left(u_x^0[b]\oplus u_x^j[b]\right)>0\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

Instead of the challenge-response pairs $\{x,u_x\}$ of unimodal PUFs, BPUFs have triplets of challenge-responses $\{x,u_x,\vartheta_x\}$. BPUFs are strong or weak if the relationship between the number of challenge-response triplets with the number of basic elements are large or small.

The physical conditions usually evaluated in unimodal PUFs give unbiased responses, that is, responses with the same number of 1's and 0's. The fractional Hamming weight (*FHW*) of a binary vector measures the normalized percentage of 1's and 0's as follows:

$$FHW(r)=\frac{1}{N}\sum_{b=0}^{N-1}r[b] \tag{9}$$

The average *FHW* of unbiased responses is 0.5 ($\overline{FHW}=0.5$). In order to allow for more versatility, the conditions considered herein for BPUFs can provide biased responses, that is, with different numbers of 1's and 0's. Without loss of generality, let us assume that the number of 1's, M, is equal or smaller than half of the bits, $M\leq N/2$, that is $\overline{FHW}=M/N\leq 0.5$. This happens to the above commented example

of flipping bits in SRAM PUFs, which usually are around 10% of the start-up values, thus making $\overline{FHW} \simeq 0.1$ in those BPUF behavioral responses.

It is known in binary codes that the average *FHD* between two independent binary vectors, y and z, is related to the average *FHW*, $\overline{FHW}$, as follows [26]:

$$\overline{FHD}(y,z) = \overline{FHW}(y) + \overline{FHW}(z) - 2 \cdot \overline{FHW}(y) \cdot \overline{FHW}(z) \quad (10)$$

If the average *FHW* of both binary vectors is M/N, then:

$$\overline{FHD}(y,z) = \frac{2M(N-M)}{N^2} \quad (11)$$

Applying that to the usual unimodal PUF responses with $\overline{FHW} = 0.5$, $M = N/2$, it results in a $\overline{FHD}_{inter} = 0.5$. In the other side, since biased responses of BPUFs can have, for example, $\overline{FHW} = 0.1$, it can result in a $\overline{FHD}_{inter} = 0.18$. Hence, genuine and impostor distributions analyzed with Hamming distances, are closer if the responses are biased than if they are unbiased.

In order to better distinguish between genuine and impostor distributions using BPUF biased responses, a metric based on Jaccard instead of Hamming distance is proposed in the following.

A. RESPONSE SIMILARITY: JACCARD DISTANCE

Similarity between BPUF responses is measured with Jaccard distance (*JD*). The Jaccard distance between two binary vectors is calculated as follows:

$$JD(y,z) = \frac{M_{01}(y,z) + M_{10}(y,z)}{M_{01}(y,z) + M_{10}(y,z) + M_{11}(y,z)} = \frac{\sum_{b=0}^{N-1} (y[b] \oplus z[b])}{\sum_{b=0}^{N-1} (y[b] \oplus z[b]) + \sum_{b=0}^{N-1} (y[b] \wedge z[b])} = \frac{\sum_{b=0}^{N-1} (y[b] \oplus z[b])}{\sum_{b=0}^{N-1} (y[b] \vee z[b])} \quad (12)$$

where $\vee$ represents the OR operation, $\wedge$ the AND operation, $M_{01}(y,z)$ the number of bits that are 0 in y and 1 in z, $M_{10}(y,z)$ the number of bits that are 1 in y and 0 in z, and $M_{11}(y,z)$ represents the number of bits that are 1 in both y and z.

The difference between Jaccard and fractional Hamming distance (*FHD*) is that *JD* does not add $M_{00}(y,z)$ in the denominator, which represents the number of bits that are 0 in both y and z.

The numerator in Equation 12 is the Hamming distance between the two vectors. The denominator in Equation 12 counts the number of 1's that are only in one of the vectors or in both. This is equivalent to sum the number of 1's in each vector and subtract the number of 1's in both:

$$M_{01}(y,z) + M_{10}(y,z) + M_{11}(y,z) = HD(y,z) + M_{11}(y,z) = HW(y) + HW(z) - M_{11}(y,z) \quad (13)$$

Hence, $M_{11}(y,z)$ can be expressed as:

$$M_{11}(y,z) = \frac{HW(y) + HW(z) - HD(y,z)}{2} \quad (14)$$

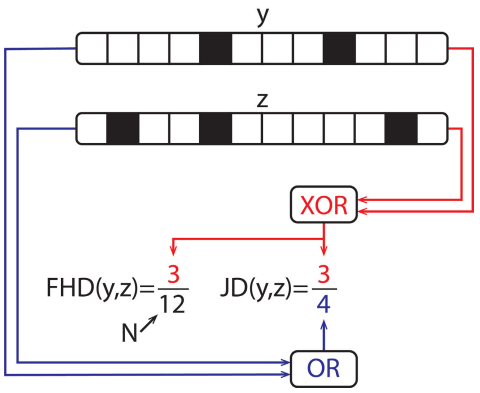


FIGURE 2. Example to illustrate the *JD* and the *FHD* of two vectors.

The mathematical relation between *JD* and *FHD* can be found by substituting $M_{11}(y,z)$ as expressed in Equation 14 into Equation 12 and reordering:

$$JD(y,z) = \frac{HD(y,z)}{HD(y,z) + \frac{HW(y)+HW(z)-HD(y,z)}{2}}$$
$$JD(y,z) = \frac{2 \cdot FHD(y,z)}{FHW(y) + FHW(z) + FHD(y,z)} \quad (15)$$

Hence, while the range of *FHD* is $0 \leq FHD(y,z) \leq minimum(1, FHW(y) + FHW(z))$, which can be small if the *FHWs* are small, the range of *JD* is always $0 \leq JD(y,z) \leq 1$, independently of the *FHWs*.

Figure 2 shows an example to illustrate the *FHD* and *JD* of two vectors.

Jaccard distance evaluated between responses of the same instance to the same challenges at different measurements (genuine population) is called intra Jaccard distance. Intra Jaccard distances of a BPUF response with perfect reproducibility are zero. The relation between average intra *JD* and *BER* can be obtained from Equation 15. For example, if $\overline{FHW} = 0.1$ and assuming $\overline{FHD}_{intra} = 0.01$, $\overline{JD}_{intra} = 0.095$.

Jaccard distance evaluated between responses of different instances to the same challenges (impostor population) is called inter Jaccard distance. The relation between average inter *JD* and average inter *FHD* can be obtained from Equation 15 and Equation 11. Considering two independent binary vectors y and z with average *FHW* equal to M/N, whose average *FHD* is given by Equation 11, their average *JD*, $\overline{JD}_{inter}$, can be approximated by $2(N-M)/(2N-M)$, which is closer to 1 as N is greater than M.[1] For the same example above, if $\overline{FHW} = M/N = 0.1$, which makes $\overline{FHD}_{inter} = 0.18$ (according to Equation 10), $\overline{JD}_{inter} = 0.947$. It can be seen in this example that the distance between genuine and impostor distributions is longer using *JDs* ($\overline{JD}_{inter} - \overline{JD}_{intra} = 0.852$) than *FHDs* ($\overline{FHD}_{inter} - \overline{FHD}_{intra} = 0.17$), which is better for authentication purposes.

[1] The same result to $\overline{JD}_{inter}$ is obtained applying the expected value of the Jaccard similarity, $M/(2N-M)$.

B. AUTHENTICATION BASED ON BPUFs

Like a multimodal biometric system, which registers physical and behavioral responses for a given individual, the registered responses of a BPUF for a challenge vector x are physical and behavioral, $\{u_x^0, \vartheta_x^0\}$. During verification, the distances between the measured responses $\{u_x^i, \vartheta_x^i\}$ and the registered ones are calculated. As in multimodal biometric systems, distances can be combined with several operators to compute a global distance or individual distances can be considered. The latter option is considered herein to extend the work already done in unimodal PUFs that use Hamming distances.

As commented above, Hamming distances are adequate for unbiased physical responses. Hence, let us maintain the same condition summarized in Subsection II-B to verify the BPUF physical responses, that is, if $HD(u_x^i, u_x^0) \leq HD^{max}$, the BPUF physical responses are authentic. The security improvement with BPUFs is that another condition should be met by the BPUF behavioral responses. Since BPUF behavioral responses can be biased, our proposal is to use Jaccard distances for that condition. Thus, a BPUF instance is verified if the distances of its responses are below selected thresholds, HD^{max} and JD^{max}, that is, if $HD(u_x^i, u_x^0) \leq HD^{max}$ and $JD(\vartheta_x^i, \vartheta_x^0) \leq JD^{max}$. Since the authentication with physical responses is well known and was summarized in Subsection II-B, this subsection is focused on the authentication with behavioral responses, analyzing the peculiarities of using Jaccard distances.

According to Equation 12, the numerator of $JD(\vartheta_x^i, \vartheta_x^0)$, $M_{01}(\vartheta_x^i, \vartheta_x^0) + M_{10}(\vartheta_x^i, \vartheta_x^0)$, measures the number of errors, e_ϑ, between ϑ_x^i and ϑ_x^0. In the denominator, $M_{11}(\vartheta_x^i, \vartheta_x^0)$ measures the number of bits that are 1 in both responses, which are considered as the number of successes, s_ϑ. Hence, $JD(\vartheta_x^i, \vartheta_x^0)$ can be expressed as:

$$JD(\vartheta_x^i, \vartheta_x^0) = \frac{e_\vartheta}{e_\vartheta + s_\vartheta} = \frac{1}{1 + s_\vartheta / e_\vartheta} \tag{16}$$

The threshold JD^{max} is associated with a minimum number of successes, s^{min}, and a maximum number of errors, e^{max}, as follows:

$$JD^{max} = \frac{1}{1 + s^{min}/e^{max}} \tag{17}$$

Therefore, JD^{max} is selected depending on both constraints, s^{min} and e^{max}.

Concerning e^{max}, the procedure is like the selection of HD^{max} for the physical responses. The value of e^{max} is selected so as to ensure a small false rejection rate of the genuine behavioral responses, $P_G(E > e^{max}) < \epsilon$, with $\epsilon = 10^{-6}$, for example (like in Equation 3). The bit error probability of the genuine behavioral responses (like p_{eG} in Equation 3) can be estimated experimentally as the bit error rate, BER (in Equation 4), considering genuine behavioral responses, that is, $\overline{FHD}_{intra}$ of behavioral responses.

Concerning s^{min}, the procedure requires modeling successes. The successes of a genuine behavioral response can be modeled as a random variable S of N independent bits in which each bit has a bit success probability of p_{sG}. Under these circumstances, the probability that the genuine behavioral response contains less than s^{min} successes (false rejection due to minimum successes threshold) is given by:

$$P_G(S < s_{min}) = \sum_{i=0}^{s_{min}-1} \binom{N}{i} p_{sG}^i (1 - p_{sG})^{N-i} \tag{18}$$

The bit success probability in the genuine behavioral responses, p_{sG}, can be estimated experimentally as the bit success rate (BSR) considering genuine behavioral responses. Similarly to the BER, the bit success rate (BSR) in a set of R responses, $\{\vartheta_x^0, \ldots, \vartheta_x^{R-1}\}$, is calculated as the average ratio of the number of bit successes in the total number of N bits, as follows:

$$\begin{aligned} BSR(\vartheta_x^0, \ldots, \vartheta_x^{R-1}) &= \frac{2}{R(R-1)} \sum_{i=0}^{R-2} \sum_{j=i+1}^{R-1} \frac{M_{11}(\vartheta_x^i, \vartheta_x^j)}{N} \\ &= \frac{2}{R(R-1)} \sum_{i=0}^{R-2} \sum_{j=i+1}^{R-1} \frac{\sum_{b=0}^{N-1} \vartheta_x^i[b] \wedge \vartheta_x^j[b]}{N} \end{aligned} \tag{19}$$

Once p_{sG} is estimated, the threshold s^{min} is selected so as to ensure a small false rejection rate of the genuine behavioral responses, $P_G\left(S < s^{min}\right) < \epsilon$, with $\epsilon = 10^{-6}$, for example. From s^{min} and e^{max}, JD^{max} is calculated.

IV. SECURITY IMPROVEMENT WITH BPUFs

In general, BPUFs are more difficult to attack than PUFs since given a challenge, not only the physical response but also the behavioral response associated to that challenge have to be predicted or cloned. In addition, if the behavioral response is obtained from measurements of the physical response, as shown in Equation 7, the behavioral response can detect if the physical response has been manipulated. This is detailed in the following.

A. RESISTANCE TO REPORTED ATTACKS

The reported attacks based on machine learning techniques to generate virtual clones of PUFs employ static challenge-response functional relationships to approximate the challenge-response pairs [13], [14]. In fact, Artificial Neural Networks with feed-forward architectures and static neural units are static approximators [27]. If static approximators are used to replace PUFs, the physical responses provided for the challenges do not change over time. In this sense, we can say that these so cloned PUFs are lifeless.

The physical attacks reported to generate physical clones of PUFs fix the physical responses to the registered ones [17], [18]. Hence, the physical responses provided for the challenges do not change over time, as in the virtual clones above. Again, we can say that these so cloned PUFs are lifeless.

If the behavioral responses are obtained from several measurements of the physical responses taken at several

sample times, as shown in Equation 7, the behavioral response can detect if the physical responses are alive (they change over time) or lifeless (they are always the same). In particular, if the behavioral responses are obtained as shown in Equation 8, and the physical responses are provided by the above mentioned virtual or physical clones, all the bits of the resulting behavioral responses are zero because fake physical responses never flip. As will be explained in the following subsection and confirmed experimentally in Section V, behavioral responses must have a minimum number of bits equal to 1 to be accepted as genuine. A behavioral response with all the bits zero is rejected as impostor. In this sense, we can say that the behavioral responses of BPUFs are able to detect liveness.

In the other side, the physical cloning attacks reported in [19], [20] would have a low success rate in the proposed BPUFs since cloning physical responses that are not static but change over time (flip) is much more difficult.

The hybrid attack proposed in [21] and [22] uses the information provided by the response bits that sometimes flip as an indirect way to evaluate which virtual copies of the PUF under attack are the most accurate, i.e., provide the best static challenge-response pairs. They apply Evolution Strategies to select the best models as parents for the next generation and repeat this process to improve the accuracy of the children models. The work in [22] demonstrates how these attacks, named as reliability-based machine learning attacks, are successful to break several protocols where the challenge-response pairs are not available to the attacker, but the obfuscated data managed leak information about the flipping or unreliable bits in the physical responses. In the proposed BPUFs, the attacker should not know the challenge-response triplets that define the BPUF. This means the attacker should not have information not only about the physical responses but also about their reliability, which is associated with the behavioral response. Reliability-based machine learning attacks cannot be applied to BPUFs since the attacker should not know reliability information.

The protocols attacked in [22] cannot be used with BPUFs since they leak information about behavioral responses. The behavioral responses cannot be used like the physical responses to generate helper data that obfuscate an encoded secret by XORing it with the behavioral response, since information about the secret is leaked from those helper data (due to the bias of behavioral responses). Other protocols and algorithms should be used with BPUFs but they are outside the scope of this work. Prior to develop particular protocols, the following subsection analyzes the resistance of BPUFs to false acceptance attacks or brute-force attacks, since these attacks are generic for any protocol.

B. RESISTANCE TO FALSE ACCEPTANCE ATTACKS

As commented in Subsection III-B, the authentication thresholds HD^{max} and JD^{max} are selected to guarantee that the BPUF will not suffer from false rejection except with a negligible probability. However, another consideration to take into account is the false acceptance, that is, the probability that an attacker could be authenticated as genuine. Let us assume a scenario in which the attacker had been successful in discovering the physical response of the BPUF instance, so that he/she is able to meet $HD(u_x^i, u_x^0) \leq HD^{max}$. Let us also assume that it is not checked if the behavioral response meets or not Equations 7 or 8. That is, the attacker can try a behavioral response without relation with the physical responses. Even in that case, the attacker should also have to provide a behavioral response, ϑ_x^i, able to meet $JD(\vartheta_x^i, \vartheta_x^0) \leq JD^{max}$.

Let us assume that the registered behavioral response of N bits, ϑ_x^0, had M bits equal to 1, i.e., $HW(\vartheta_x^0) = M$, and that the attacker knows N, M, and JD^{max} (otherwise the attack is further complex). The attacker tries to authenticate with an N-bit response, ϑ_x^i, that have n bits equal to 1, $HW(\vartheta_x^i) = n$. In this case, let us assume that the attacker succeeds in choosing s_ϑ bits equal to 1, that is $M_{11}(\vartheta_x^i, \vartheta_x^0) = s_\vartheta$. It can be deduced that the total number of errors, e_ϑ, of this authentication trial is given by the number of 1's wrongly selected, that is $M_{10}(\vartheta_x^i, \vartheta_x^0) = n - s_\vartheta$, plus the number of unselected 1's, $M_{01}(\vartheta_x^i, \vartheta_x^0) = M - s_\vartheta$ (assuming $M \geq s_\vartheta$). Hence, the total number of errors is:

$$e_\vartheta = M + n - 2 \cdot s_\vartheta \tag{20}$$

The attack is successful if:

$$\frac{1}{1 + \frac{s_\vartheta}{M+n-2 \cdot s_\vartheta}} \leq JD^{max}$$

$$s_\vartheta \geq (M+n) \cdot \frac{1 - JD^{max}}{2 - JD^{max}} = s_{attack} \tag{21}$$

Therefore, the minimum number of bits equal to 1, n^{min}, that the attacker should try to succeed is given by:

$$n^{min} \geq s_\vartheta \geq (M + n^{min}) \cdot \frac{1 - JD^{max}}{2 - JD^{max}}$$

$$n^{min} = M \cdot (1 - JD^{max}) \tag{22}$$

In the other side, taking into account that the maximum number of successes cannot be greater than M, it follows that the maximum number of n, n^{max}, that an attacker should try to succeed is given by:

$$M \geq s_\vartheta \geq (M + n^{max}) \cdot \frac{1 - JD^{max}}{2 - JD^{max}}$$

$$n^{max} = \frac{M}{1 - JD^{max}} \tag{23}$$

In summary, the attacker will try with a response that have n bits equal to 1, with $n^{min} \leq n \leq n^{max}$. The probability of success is the probability that an attacker generates an N-bit response with n 1's of which s_{attack} (in Equation 21) or more bits are successfully selected. The feature of behavioral responses that minimizes this probability is that any of the N units in the BPUF (SRAM for example) can meet the behavioral condition, that is, the probability of success is minimum if the attacker does not know which units to select because all of them can be valid. This is the desired unpredictability

feature of the behavioral response: given a challenge (the address of a SRAM cell), the behavioral response cannot be known (the cell cannot be known to be reliable or not). If the behavioral response is so unpredictable, the probability of success is given by the cumulative hypergeometric distribution function as follows:

$$P_I\left(S \geq s_{attack}\right) = \sum_{s=s_{attack}}^{n} \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \tag{24}$$

This probability can be quite small, as shown in the following section with experimental results of SRAM BPUFs.

A way to prove the unpredictability of the behavioral response is to consider for simplicity that N is a power of 2, that is, $N = 2^Q$. Hence, each of the M units (SRAM cells for example) that meets the behavioral condition is identified by a code with Q bits, so that all the possible Q-bit codes are employed to identify all the units. If any unit can meet the behavioral condition then any Q-bit code can appear in the registered responses of the genuine BPUF instances. From the point of view of the attacker, discovering a registered behavioral response with M bits equal to 1 is equivalent to discovering a sequence with M codes of Q bits, that is, a sequence of $M \cdot Q$ bits. Of course, the Q-bit codes should be assigned randomly to each unit so as not to reveal anything about the unit. If they follow an order, for example the position of the SRAM cell in the SRAM, the attacker would know that the first cells would have many 0's and the last ones many 1's, thus being somewhat predictable. Also, once the codes are assigned, the same codification is used for all the instances.

Considering a set of BPUF instances, the registered sequences of $M \cdot Q$ bits should be unpredictable to the attacker. Standard tests that evaluate the unpredictability of a set of sequences are included in the NIST Statistical Test Suite [28]. The basic condition that the sequences should meet is to have the same appearance probability of 1's and 0's, that is, their average fractional Hamming weight should be 0.5, which is evaluated by the frequency test. In addition, subsequences of an unpredictable sequence should also be unpredictable, which is evaluated by the block, cumulative sums and runs tests. These will be the tests employed in the following Section to evaluate the unpredictability of behavioral responses.

V. EXPERIMENTAL RESULTS OF SRAM BPUFs

BPUFs based on SRAMs were analyzed experimentally in order to validate the proposals described in Section III and evaluate quantitatively the security improvement analyzed in Section IV. The SRAMs analyzed were low-power dual-port 8-transistor TSMC (Taiwan Semiconductor Manufacturing Company) IP (Intellectual Property) SRAMs that were included in ASICs (Application Specific Integrated Circuits) fabricated in the 90-nm CMOS technology from TSMC. Results corresponding to 8 IP blocks are shown herein, each block with capacity for 4096 words of 60 bits.

The well-known start-up values of SRAM cells were measured as physical response. The physical responses considered have a maximum of 7296 bits (128 words of 57 bits). Behavioral responses were measured as described by Equation 8, evaluating the behavior of 20 measurements ($R = 20$) of the physical responses. Hence, behavioral responses have a maximum of 7296 bits (128 words of 57 bits). Since start-up values of SRAM cells of the same and different IP blocks show uniqueness, as shown in [24], 128 different physical and behavioral responses of the same size were analyzed (16 responses/IP in 8 IPs). The distribution of genuine responses was analyzed using 1280 comparisons (for each of the 128 responses, a measurement is compared against 10 measurements of the same response). The distribution of impostor responses was analyzed using 8128 comparisons (the different pairs between the 128 responses).

The results shown in the following focus on behavioral responses since physical responses have been already illustrated in many other works.

A. HAMMING WEIGHT OF RESPONSES

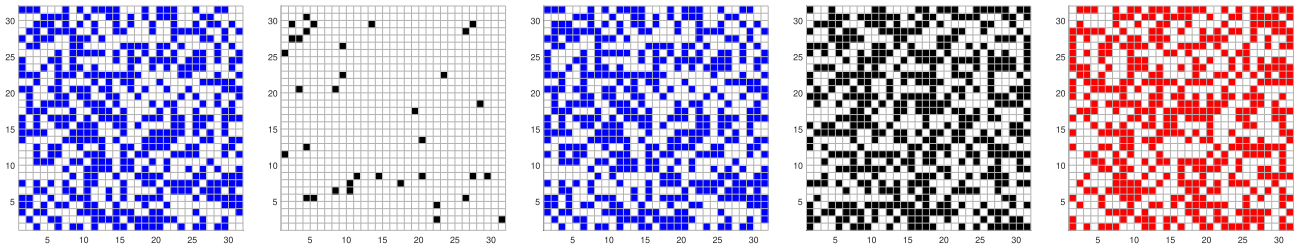
It was verified that the Hamming weight of the behavioral responses is smaller than that of the physical responses. This difference is apparent in the graphic representations of Figure 3 and 4, which illustrate both responses organized as squared maps with 32×32 bits. The response bits equal to 0 are white and those equal to 1 are not white (blue if they are associated with a genuine instance and red if they correspond to an impostor instance). All the responses shown in Figure 3 and 4 were taken under nominal operating conditions.

The Hamming distances between responses of genuine instances and between genuine and impostor instances are shown, respectively, in Figure 3(b) and 4(b), and in Figure 3(d) and 4(d). It can be observed that when the responses were generated by genuine instances the number of bits that changed or had errors (colored in black) is much smaller.

Although in both kind of responses it is possible to distinguish if the comparison is made between genuine and impostor instances, there are significant differences between the Hamming distance corresponding to physical responses (Figure 3(d)) and behavioral responses (Figure 4(d)) because, as discussed above (Equation 10), fractional Hamming weight is around 0.5 for physical responses and around 0.1 for behavioral responses.

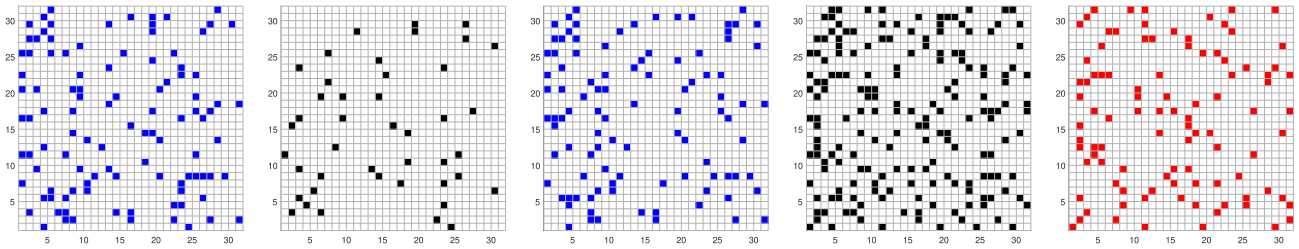
B. JACCARD VERSUS HAMMING DISTANCE

In order to make a more exhaustive analysis of the genuine and impostor instances, the metric of the Jaccard distance introduced in Subsection III-A was used in comparison with the fractional Hamming distance and responses with 7296 bits were considered. For the physical responses,



(a) u_x^0 (b) $u_x^0 \oplus u_x^1$ (c) u_x^1 (d) $u_x^1 \oplus v_x^0$ (e) v_x^0

FIGURE 3. Physical responses of a genuine instance (a), (c), physical response of an impostor instance (e), and bitmap of their Hamming distances (b), (d).



(a) ϑ_x^0 (b) $\vartheta_x^0 \oplus \vartheta_x^1$ (c) ϑ_x^1 (d) $\vartheta_x^1 \oplus \zeta_x^0$ (e) ζ_x^0

FIGURE 4. Behavioral responses of a genuine instance (a), (c), behavioral response of an impostor instance (e), and bitmap of their Hamming distances (b), (d).

Figure 5(a) shows, on the left, in blue, the distribution of the *FHDs* between genuine instances (using 1280 comparisons) and, on the right, in red, the distribution of the *FHDs* between genuine and impostor instances (using 8128 comparisons). Analogously, Figure 5(b) shows the same probability distributions but using the *JD*.

As claimed in Section III, it can be seen that although the *FHD*-based metric shows a good separation between both populations, the use of *JD* allows further distancing the genuine population (which is attracted towards the ideal value of 0) from the impostor (which is attracted to the ideal value of 1, instead of 0.5 as in the *FHD*).

For the behavioral responses, similar results are shown in Figure 6 with the genuine (on the left, in blue) and impostor populations (on the right, in red), also using 1280 and 8128 comparisons, respectively. In this case, it is more evident that the use of *JDs* represents in a much more significant way the distance between the genuine and impostor populations (Figure 6(b)) compared to the *FHDs* (Figure 6(a)).

C. REPRODUCIBILITY AND UNIQUENESS OF BEHAVIORAL RESPONSES

Once the advantages of the Jaccard versus the Hamming distance were verified, *JDs* were used to analyze the reproducibility and uniqueness of the behavioral responses as proposed in Section III. To carry out an exhaustive analysis, measurements of the 7296-bit responses were taken in 6 different operating conditions (see Table 1) that include the cases that generate the most significant changes in the physical responses, such as voltage changes in the power supply (C2 and C3), changes in the ambient temperature (C4 and C5), and aging (C6). The ASICs were working continuously during 96 hours under accelerating aging at a temperature of 75°C. A climatic chamber ACS-EOS 200TC was used to carry out aging. Details about how the experiments were carried out can be seen in [24].

TABLE 1. Conditions evaluated.

Condition	Voltage (V)	Temperature (ºC)	Aging
C1	1.20	25	After
C2	1.08	25	After
C3	1.32	25	After
C4	1.20	5	After
C5	1.20	75	After
C6	1.20	25	Before

In order to illustrate the reproducibility of behavioral responses, 1280 comparisons using the Jaccard distance were used per operating condition to generate the distribution of the genuine population associated with each condition (from C1 to C6). Figure 7 shows these distributions, as well as the distribution representing the impostor population calculated using 8128 Jaccard distance comparisons in all the conditions (from C1 to C6).

The average value of each *JD* distribution is shown in Table 2, as well as the average value of the same distributions if the metric used to evaluate the comparisons were the *FHD*. As can be seen, all the values are similar for the impostor distributions ($\overline{JD}_{inter}$ and $\overline{FHD}_{inter}$) and the

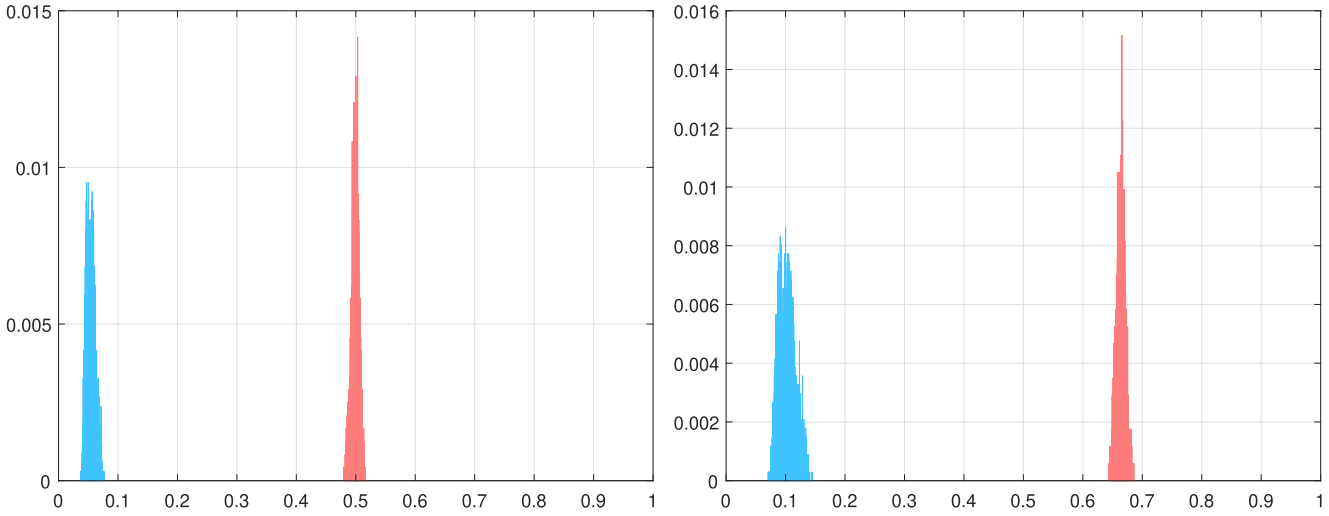


FIGURE 5. Distributions of *FHDs* (a) and *JDs* (b) of genuine and impostor populations of physical responses. The vertical axes represent the fractional number of comparisons.

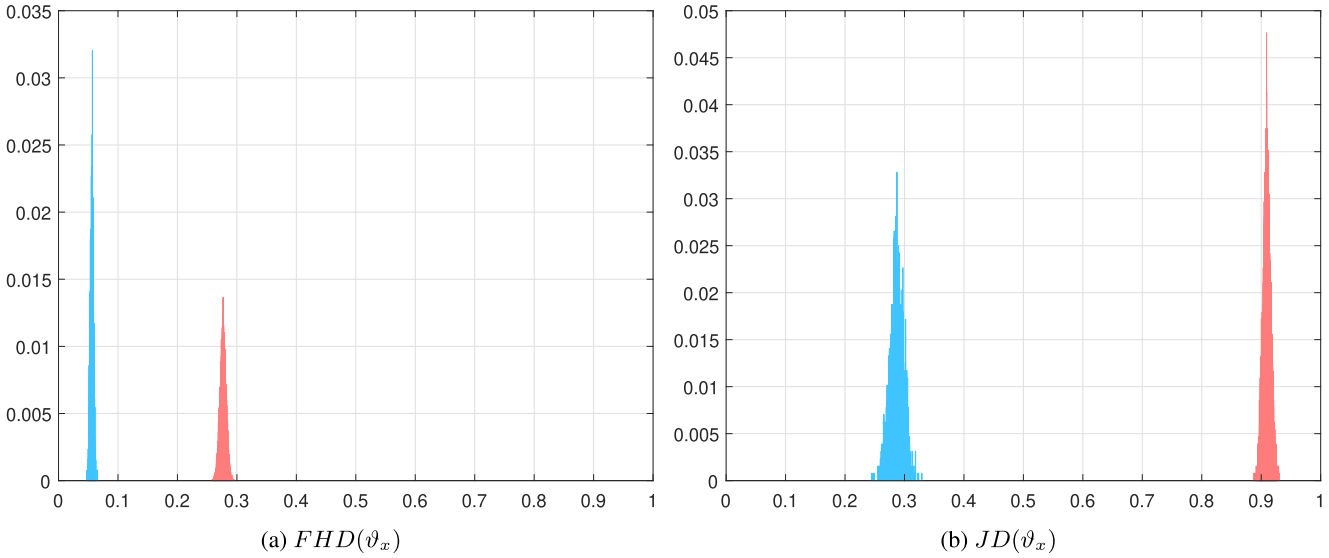


FIGURE 6. Distributions of *FHDs* (a) and *JDs* (b) of genuine and impostor populations of behavioral responses. The vertical axes represent the fractional number of comparisons.

distances between genuine and impostor distributions are smaller using *FHD* than in the case of using *JD* metric. The *JD* metric completely separates the impostor from the genuine distributions in all the operating conditions.

D. RESULTS ON SECURITY IMPROVEMENT

Let us analyze quantitatively the resistance of these SRAM BPUFs to false acceptance attacks in the behavioral responses. As explained in Subsection III-B (Equation 17), the authentication threshold JD^{max} is selected from the maximum number of errors in the genuine responses, e^{max}, and the minimum number of successes in the genuine responses, s^{min}.

TABLE 2. *JD* and *FHD* of the genuine and impostor populations of behavioral responses at different operating conditions.

Condition	$\overline{JD}_{intra}$	$\overline{JD}_{inter}$	$\overline{FHD}_{intra}$	$\overline{FHD}_{inter}$
C1	0.2875	0.9082	0.0561	0.2797
C2	0.3430	0.9164	0.0663	0.2611
C3	0.2823	0.9046	0.0559	0.2876
C4	0.6008	0.9119	0.1408	0.2714
C5	0.7406	0.9239	0.1807	0.2428
C6	0.3858	0.9051	0.0811	0.2866

The values of e^{max} for each operating condition were selected by imposing a false rejection rate of 10^{-6}, $P_G(E > e^{max}) < 10^{-6}$, estimating the bit error probability

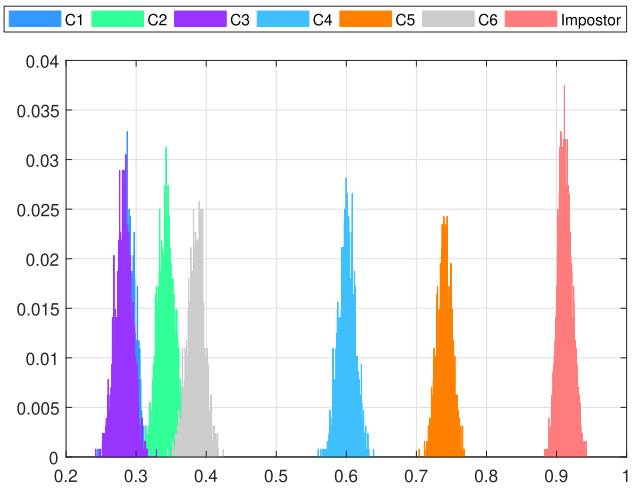


FIGURE 7. Distributions of the *JDs* of the genuine and impostor populations of behavioral responses at different operating conditions. The vertical axes represent the fractional number of comparisons.

TABLE 3. Bit error and success probabilities, maximum errors, minimum successes, and *JD* thresholds.

Condition	p_e	p_s	e^{max}	s^{min}	JD^{max}
C1	0.0561	0.1390	506	876	0.3661
C2	0.0663	0.1269	588	794	0.4255
C3	0.0559	0.1421	504	898	0.3595
C4	0.1408	0.0935	1171	567	0.6738
C5	0.1807	0.0633	1476	366	0.8013
C6	0.0811	0.1291	706	809	0.4660

as the *BER* of the genuine responses. The second column in Table 3 shows the *BER* obtained experimentally for each operating condition. They are equal to the $\overline{FHD}_{intra}$ shown in the fourth column of Table 2. The values of e^{max} calculated for each operating condition are shown in the fourth column of Table 3. For operating condition C1, Figure 8 illustrates with bars the distribution of the random variable E, that is, the normalized number of errors $\sum_{b=0}^{N-1} \vartheta_x^i \oplus \vartheta_x^0/N$ in the genuine responses. The continuous red line in Figure 8 represents a binomial distribution (Equation 2) with $N = 7296$ bits and $p_{eG} = 0.0561$ (the bit error probability as the value of *BER* in condition C1). It can be seen how experimental results confirm the model employed.

The values of s^{min} for each operating condition were selected by imposing also a false rejection rate of 10^{-6}, $P_G(S < s^{min}) < 10^{-6}$, estimating the bit success probability as the *BSR* of the genuine responses (Equation 19). The third column in Table 3 shows the *BSR* obtained experimentally for each operating condition. The values of s^{min} calculated for each operating condition are shown in the fifth column of Table 3. For operating condition C1, Figure 9 illustrates with bars the distribution of the random variable S, that is, the normalized number of successes $\sum_{b=0}^{N-1} \vartheta_x^i \wedge \vartheta_x^0/N$ in the genuine responses. The continuous blue line in Figure 9 represents a binomial distribution with $N = 7296$ bits and $p_{sG} = 0.1390$ (the bit success probability as the value of

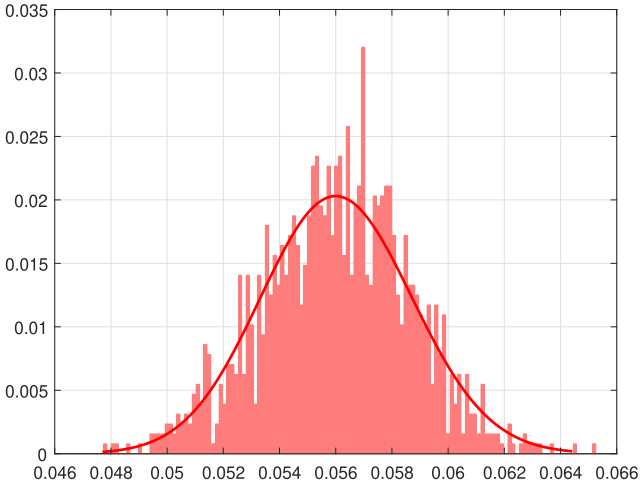


FIGURE 8. Error distribution of the genuine responses.

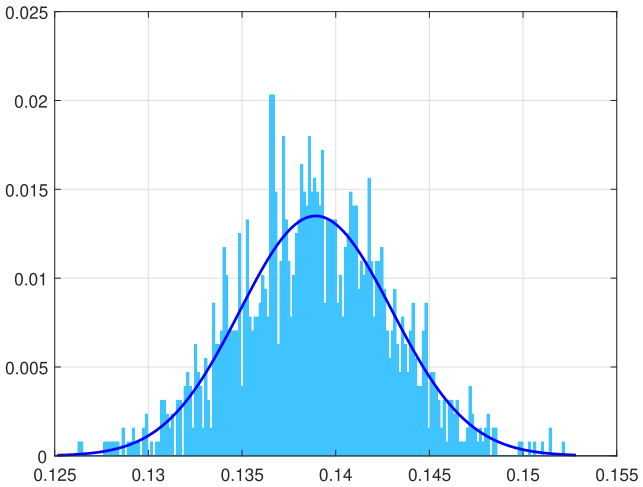


FIGURE 9. Success distribution of the genuine responses.

BSR in condition C1). It can be seen how experimental results confirm the model employed.

Table 3 shows that the minimum number of bits equal to 1 in the behavioral responses to not be rejected, s^{min}, is always greater than zero. Hence, behavioral responses with all their bits zero (resulting, for example, from fake physical responses that do not flip as mentioned in Subsection IV-A) would be rejected.

With the values of e^{max} and s^{min} for each operating condition, the thresholds JD^{max} were calculated with Equation 17. They are shown in column 6 of Table 3. As expected from the distributions in Figure 7, more restrictive thresholds from an impostor point of view are set for the operating conditions with fewer errors (C1, C2, C3 and C6), while the conditions with the highest number of errors (temperature variations, C4 and C5) have the least restrictive thresholds in order to provide a low false rejection rate.

As explained in Subsection IV-B, from the point of view of the attacker, discovering a registered behavioral response with M bits equal to 1 is equivalent to discovering a sequence with M codes of Q bits, that is, a sequence of $M \cdot Q$ bits.

TABLE 4. Results for the uniformity of P-values and the proportion of passing sequences for the 128 registered behavioral responses.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPORTION	STATISTICAL TEST
9	12	9	12	15	14	14	11	17	15	0.756476	128/128	Frequency
10	12	9	10	10	15	18	16	13	15	0.568055	127/128	BlockFrequency
13	9	8	14	12	20	13	10	12	17	0.324180	128/128	CumulativeSums
11	10	11	11	16	16	7	18	10	18	0.232760	128/128	CumulativeSums
13	12	11	10	12	16	19	9	15	11	0.585209	126/128	Runs
6	14	14	17	18	11	13	9	15	11	0.311542	127/128	LongestRun

TABLE 5. Optimal number of n and minimum s to perform a false acceptance attack with the highest probability P_I.

Condition	n_{attack}	s_{attack}	$P_I(S \geq s_{attack})$
C1	1040	879	$4.1{\cdot}10^{-659}$
C2	1049	830	$5.9{\cdot}10^{-562}$
C3	1074	898	$3.9{\cdot}10^{-670}$
C4	1401	646	$1.7{\cdot}10^{-193}$
C5	3439	773	$1.5{\cdot}10^{-34}$
C6	1107	812	$1.4{\cdot}10^{-497}$

To prove that the registered sequences of $M \cdot Q$ bits are unpredictable, we considered for simplicity that N was a power of 2, that is, $N = 2^Q$, with $Q = 12$, and the codification assigned to the cells was fixed. In this experiment, the 128 BPUF instances were registered with sequences of $596 \cdot 12$ bits ($M = 596$), since a minimum of 596 cells out of the total 2^{12} cells provided flipping bits in all the 128 instances, each cell identified by the codification fixed prior to the experiment. The NIST tests were applied to these 128 sequences with 7152 bits. As shown in Table 4, all the tests were passed with a significance level of $\alpha = 0.01$. Hence, Equation 24 can be applied to calculate the probabilities of the attacker to succeed.

If the attacker knows JD^{max} and M for each operating condition, he/she will try with a behavioral response with n bits equal to 1, with $M \cdot (1 - JD^{max}) \leq n \leq M/(1 - JD^{max})$. The highest probabilities to succeed obtained by the attacker for each condition ($P_I(S \geq s_{attack})$) are shown in the fourth column of Table 5 (applying Equation 24). They correspond to n_{attack} bits equal to 1 (second column in Table 5), which in turn correspond to the minimum number of successes required to be authenticated, s_{attack} (Equation 21), shown in the third column of Table 5.

It can be concluded by observing Table 5 that threshold JD^{max} greatly affects security. A designer can choose a more restrictive threshold defined by conditions C1, C2, C3 and C6 in order to make the system more secure at the expense of accepting that the system will be less robust to temperature variations (because higher false rejection rate may occur). In the other side, the designer can choose a less restrictive threshold to make the system robust at any operating condition, such as the one determined by condition C5, at the expense of making the system less secure against false acceptance attacks. Even in that case ($JD^{max} = 0.8013$), the highest probability of false acceptance is $1.5 \cdot 10^{-34}$, which is equivalent to a security of more than 2^{113} bits, quite enough for a BPUF even in the case of its physical response were successfully attacked.

VI. CONCLUSION

Since the proposed behavioral responses of BPUFs are obtained from several measurements of the physical responses taken at several sample times, they can detect if the physical responses are provided by the virtual or physical clones resulting from the machine learning and physical attacks reported to current PUFs. In this sense, the BPUFs proposed in this article are tamper-resistant and tamper-evident to those PUF attacks.

Physical clones of BPUFs are more challenging to obtain since cloning physical responses that are not static but change over time (flip) is much more difficult. Virtual clones are also more challenging to obtain because more responses have to be predicted or cloned for an arbitrary challenge in BPUFs. In addition, hybrid attacks like the reliability-based machine learning attacks cannot be applied to BPUFs since the attacker should not know reliability information, which is associated with the behavioral responses.

While Hamming distance is employed to measure the similarity of current PUF responses, this article shows that Jaccard distance is more suitable for evaluating similarity of behavioral responses. The BPUF behavioral responses analyzed provide enough reproducibility and uniqueness so as to provide negligible rates of false rejection and false acceptance when fixing an authentication threshold based on the Jaccard distance. Moreover, the behavioral responses provide enough unpredictability so as to provide very high resistance to false acceptance or brute-force attacks.

Many constructions currently employed for PUFs can be used for BPUFs. As example, SRAM BPUFs are presented in this article. SRAM BPUFs were characterized experimentally using low-power dual-port 8-transistor SRAMs fabricated in 90-nm CMOS technology, considering nominal and non-nominal operating conditions (changing power supply voltage and temperature as well as aging the circuits). From these experimental results, the highest probability estimated for an attacker to succeed in mathematically cloning the behavioral responses of SRAM BPUFs with a brute-force attack (allowing that fake behavioral and physical responses could have no relation), in any operation condition, is as low as $1.5 \cdot 10^{-34}$.

The use of BPUFs in cryptographic protocols is a research line of our future work.

ACKNOWLEDGMENT

Miguel A. Prada-Delgado was with IMSE-CNM, Universidad de Sevilla, CSIC, 41092 Seville, Spain.

REFERENCES

[1] *FIPS PUB 140-2: Security Requirements for Cryptographic Modules*, NIST, Gaithersburg, MD, USA, May 2001. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.140-2.pdf

[2] Altera Corporation. (2008). *Anti-Tamper Capabilities in FPGA Designs*. [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/lite%rature/wp/wp-01066-anti-tamper-capabilities-fpga.pdf

[3] A. B. Kahng, J. Lach, W. H. Mangione-Smith, S. Mantik, I. L. Markov, M. Potkonjak, P. Tucker, H. Wang, and G. Wolfe, "Constraint-based watermarking techniques for design IP protection," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 20, no. 10, pp. 1236–1252, Dec. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/952740/

[4] O. Kömmerling and M. G. Kuhn, "Design principles for tamper-resistant smartcard processors," in *Proc. USENIX Workshop Smartcard Technol.*, Chicago, IL, USA, 1999, p. 2. [Online]. Available: https://dl.acm.org/citation.cfm?id=1267117

[5] R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions.," *Science*, vol. 297, no. 5589, pp. 2026–2030, Sep. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12242435

[6] B. Gassend, D. Clarke, M. van Dijk, and S. Devadas, "Silicon physical random functions," in *Proc. 9th ACM Conf. Comput. Commun. Secur.*, New York, NY, USA, 2002, p. 148. [Online]. Available: http://portal.acm.org/citation.cfm?doid=586110.586132

[7] D. E. Holcomb, W. P. Burleson, and K. Fu, "Initial SRAM state as a fingerprint and source of true random numbers for RFID tags," in *Proc. Conf. RFID Secur.*, Malaga, Spain, 2007, p. 1. [Online]. Available: https://www.semanticscholar.org/paper/Initial-SRAM-State-as-a-Fingerpri%nt-and-Source-of-Holcomb-Burleson/987b3119f356477ee49834098201745ff2666fcf

[8] A. Maiti, J. Casarona, L. McHale, and P. Schaumont, "A large scale characterization of RO-PUF," in *Proc. IEEE Int. Symp. Hardware-Oriented Secur. Trust*. Anaheim, CA, USA, Jun. 2010, pp. 94–99. [Online]. Available: http://ieeexplore.ieee.org/document/5513108/

[9] O. Goldreich, *Foundations Cryptography—A Primer*, vol. 1. New York, NY, USA: Now, 2005.

[10] J. Delvaux, R. Peeters, D. Gu, and I. Verbauwhede, "A survey on lightweight entity authentication with strong PUFs," *ACM Comput. Surv.*, vol. 48, no. 2, pp. 1–42, Nov. 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2830539.2818186

[11] D. Lim, J. W. Lee, B. Gassend, G. E. Suh, M. van Dijk, and S. Devadas, "Extracting secret keys from integrated circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 10, pp. 1200–1205, Oct. 2005. [Online]. Available: http://ieeexplore.ieee.org/document/1561249/

[12] S. Katzenbeisser, U. Kocabaş, V. Rožić, A.-R. Sadeghi, I. Verbauwhede, and C. Wachsmann, *PUFs: Myth, Fact or Busted? A Security Evaluation of Physically Unclonable Functions (PUFs) Cast in Silicon*. Berlin, Germany: Springer, 2012, pp. 283–301. [Online]. Available: http://link.springer.com/10.1007/978-3-642-33027-8_17

[13] U. R Ührmair, F. Sehnke, J. S Ölter, G. Dror, S. Devadas, and J. Ü. Schmidhuber, "Modeling attacks on physical unclonable functions," in *Proc. 17th ACM Conf. Comput. Commun. Secur.*, Chicago, IL, USA, 2010, pp. 237–249.

[14] J. Shi, Y. Lu, and J. Zhang, "Approximation attacks on strong PUFs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 10, pp. 2138–2151, Oct. 2020.

[15] A. Roelke and M. R. Stan, "Attacking an SRAM-based PUF through wearout," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Pittsburgh, PA, USA, Jul. 2016, pp. 206–211.

[16] F. Wilde, B. M. Gammel, and M. Pehl, "Spatial correlation analysis on physical unclonable functions," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 6, pp. 1468–1480, Jun. 2018.

[17] S. Tajik, *On Phys. Secur. Physically Unclonable Functions* (T-Labs Series in Telecommunication Services). Cham, Switzerland: Springer, 2017. [Online]. Available: http://link.springer.com/10.1007/978-3-319-75820-6

[18] D. Nedospasov, J.-P. Seifert, C. Helfmeier, and C. Boit, "Invasive PUF analysis," in *Proc. Workshop Fault Diagnosis Tolerance Cryptography*. Santa Barbara, CA, USA, Aug. 2013, pp. 30–38. [Online]. Available: http://ieeexplore.ieee.org/document/6623553/

[19] C. Helfmeier, C. Boit, D. Nedospasov, and J.-P. Seifert, "Cloning physically unclonable functions," in *Proc. IEEE Int. Symp. Hardware-Oriented Secur. Trust (HOST)*, Jun. 2013, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6581556/

[20] C. Helfmeier, C. Boit, D. Nedospasov, S. Tajik, and J.-P. Seifert, "Physical vulnerabilities of physically unclonable functions," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2014, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6800564

[21] G. T. Becker, "The gap between promise and reality: On the insecurity of XOR arbiter PUFs," in *Proc. Cryptograph. Hardw. Embedded Syst.* Saint Malo, France: Springer, 2015, pp. 535–555. [Online]. Available: http://link.springer.com/10.1007/978-3-662-48324-4_27

[22] G. T. Becker, "On the pitfalls of using arbiter-PUFs as building blocks," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 8, pp. 1295–1307, Aug. 2015.

[23] C. Bösch, J. Guajardo, A.-R. Sadeghi, J. Shokrollahi, and P. Tuyls, "Efficient Helper Data Key Extractor on FPGAs," in *Cryptograph. Hardw. Embedded System*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 181–197. [Online]. Available: http://link.springer.com/10.1007/978-3-540-85053-3_12

[24] I. Baturone, M. A. Prada-Delgado, and S. Eiroa, "Improved Generation of Identifiers, Secret Keys, and Random Numbers From SRAMs," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2653–2668, Dec. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7217837/

[25] J. Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls, "FPGA Intrinsic PUFs and Their Use for IP Protection," in *Cryptographic Hardware and Embedded Systems*. Berlin, Germany: Springer, 2007, pp. 63–80. [Online]. Available: http://link.springer.com/10.1007/978-3-540-74735-2_5

[26] Z. Zhi Zhang, "A relation between the average Hamming distance and the average Hamming weight of binary codes," *J. Stat. Planning Inference*, vol. 94, no. 2, pp. 413–419, Apr. 2001.

[27] M. M. Gupta, L. Jin, N. Homma, and L. A. Zadeh, *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory*. Hoboken, NJ, USA: Wiley, 2005.

[28] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo. (2010). *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-2%2r1a.pdf

MIGUEL A. PRADA-DELGADO received the 5-year degree in telecommunication engineering (with electronics specialization), the master's degree (Hons.) in microelectronics, and the Ph.D. degree (Hons.) in microelectronics from the University of Seville, Seville, Spain, in 2013, 2014, and 2020, respectively. Since 2019, he is a Postdoctoral Researcher at IBM Research, Zurich. His current research interests include hardware security, authentication protocols, privacy-preserving algorithms, and blockchain technology.

ILUMINADA BATURONE received the 5-year and Ph.D. degrees (Hons.) in physics from the University of Seville, Seville, Spain, in 1991 and 1996, respectively. She has been with the Microelectronics Institute of Seville (IMSE-CNM), CSIC, University of Seville, since 1990. She is currently with the Department of Electronics and Electromagnetism, Universidad de Sevilla, where she is also a Full Professor. She has coauthored the books *Microelectronic Design of Fuzzy Logic-Based Systems* (CRC Press, 2000) and *Fuzzy Logic-Based Algorithms for Video De-Interlacing* (Springer, 2010) and more than 150 scientific articles. She has participated in more than 40 Spanish and European research and industrial projects, leading 12 of them. She holds three patents and is one of the developers of the Xfuzzy environment. Her current research interests include hardware security, microelectronic design of crypto-biometric systems, hardware design for embedded control, and neuro-fuzzy systems.