A guide for assessing optically-imaged physically unclonable functions for authentication

Ella Mann-Andrews,[1,2] Thomas McGrath,[1,2] Blake Halliday,[2] and Robert James Young[1,2]
[1)] *Physics Department, Lancaster University, LA1 4YB, UK*
[2)] *Quantum Base Limited, Lancaster University, LA1 4YB, UK*

(*Electronic mail: mannandr@lancaster.ac.uk)

(Dated: 8 May 2025)

Physically unclonable functions (PUFs) are a physical security primitive with important applications in authentication, such as in anti-counterfeiting technologies. They can be used to generate unique identities, linked to their structure, by measuring features associated with them. Optically-read PUFs (O-PUFs) are a subset that utilises optical imaging techniques to create these IDs. They offer a promising solution to the perpetual challenge of counterfeiting by providing a robust authentication solution. The metrics chosen for evaluation are varied across the field; there is a lack of consensus in the figures of merit used for evaluation, as well as the protocols and standards used for assessing this specific subset of PUFs. This work reviews the progress in the development of evaluation techniques to date, and it highlights important differences in the statistical assessment of O-PUFs. A summary of the most popular metrics used in the literature in the past decade is presented, and the core metrics are isolated and mathematically defined. These are then distilled into recommendations of best practice for assessing and comparing different technologies. An open-source package, providing a full testing suite, is presented to standardise testing in this field. Finally, novel methods for evaluating the performance of O-PUFs over time are also proposed. A unified approach to assessment is essential for advancing anti-counterfeiting technologies, especially as these systems are now being used in commercial applications.

I. INTRODUCTION

Physically unclonable functions (PUFs) leverage inherently non-reproducible properties to generate unique IDs making them a technology with high potential for enhancing security and preventing counterfeiting. When an Optical-PUF (O-PUF) is read, or *challenged*, it produces a *response* which can be used for authentication. A challenge refers to a controlled stimulus applied to the PUF, while the response is the resulting output uniquely determined by the PUF's physical randomness. O-PUFs use optical imaging techniques as their method of measurement, for example microscopic photography. In terms of a large-scale use case for anti-counterfeiting, O-PUFs can be used in the following way. A database of identities (IDs) can be populated in a registration stage, for example directly after production. Later, this database is cross-referenced for authentication when the IDs are re-measured in the field. Authentication solutions that are: low cost, physical, and mass-producible are in high-demand. Recent advances in digital technologies have developed to a point where ID database storage, connectivity and scanning for an O-PUF implementation are now easily accessible and can be produced by a mobile phone. The random nature of manufacturing processes means it can be physically impractical to replicate. This is unlike a traditional anti-counterfeit technology, such as a hologram, which is easier to reproduce with access to the original manufacturing technology. PUFs, by their nature, rely on inherent randomness and are termed 'unclonable'.

Whilst a variety of O-PUF technologies have been created and explored in the past decade, the methods for analysing their effectiveness vary substantially across studies. As shown in FIG. 1, evaluation of an O-PUF begins with selection of a candidate system. The sample (or 'tag') must be tested for its ability to contain a unique ID. To achieve this, many repeat images are taken of the same tag (intra-images) and of different tags (inter-images). The intra-images are compared, to determine how repeatable imaging the tag is, and the inter-images are compared to demonstrate the randomness and uniqueness of tags. To achieve this, the images are converted to a set of 2D binary arrays using algorithms that down-sample the image data to extract fingerprints (patterns), with an aim of reducing noise in the process. After this process, the similarity of the binary arrays is compared by counting the number of bits of difference between each other (defined more formally as the Hamming distance, HD). This is the starting point for calculating a variety of figures of merit (or metrics) that can be used to compare different PUF technologies. Metrics, such as the effective number of independent bits (ENIB, defined in section III C 5), or decidability (defined in section III C 6) are used. The intra-HD and inter-HD distributions, which are fits to histograms formed from HD data, have been used in the majority of previous works to date, however the metrics that are taken from these plots vary. These range from intra-HD means (related to reliability) and inter-HD means (related to

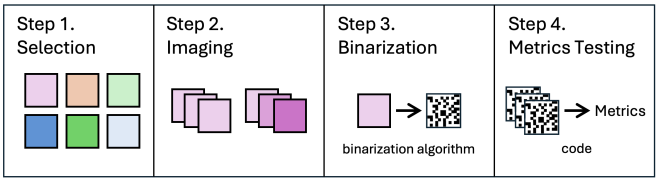


FIG. 1. A generalised overview of the O-PUF assessment process. Starting with the selection of a suitable candidate, then following on to the imaging. The images that are recorded are the 'responses' of the O-PUF. After which the images are binarized using a chosen algorithm or technique. Finally, this process ends by producing the chosen metrics by testing the 2D arrays.

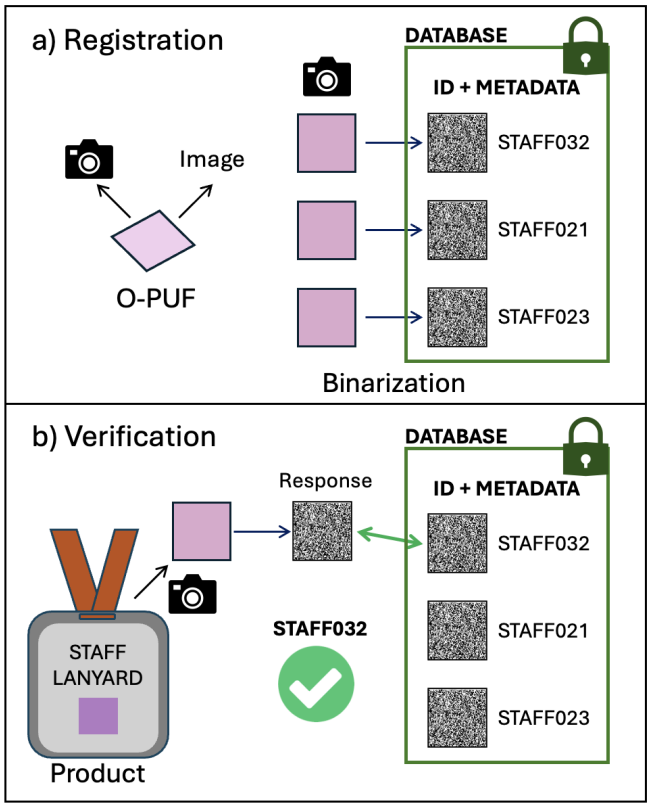


FIG. 2. Illustrating the two key phases involved in O-PUF implementation. (a) Registration; (b) Verification. A challenge is applied to the O-PUF in the form of incident light (e.g., a camera flash), and an image is captured. After binarization, the image forms a response in the challenge-response pair (CRP) architecture used to understand PUFs. The registration process involves capturing these images for individual tags and registering their IDs with associated metadata in a secure database. The binarization process is applied to the different tags (pink squares), each individual ID leading to a registration. (b) Verification: The registered database is used for anti-counterfeiting. A unique tag is attached to the product, imaged, and the resulting binary array is compared to the database. If the array is similar enough, the ID verifies the product's authenticity.

uniqueness) to more complex metrics such as ENIB or false positive rates. Many papers analyse the HD distributions and also provide secondary statistics such as the results of random number predictability analysis. Additionally, there are many other metrics, some only used in a few papers, that need standardised definitions and evaluation. There have been many attempts at producing comprehensive testing suites for PUFs in general, however, for O-PUFs, these are limited. As the main figures of merit vary in the papers published in the last decade on the field of O-PUFs, further work is needed to increase accessibility to the necessary code. The registration and verification steps for O-PUF implementation are shown in FIG. 2.

Contribution: This paper is divided into four main sections. Firstly, a review of the literature from 2013 to 2024, highlighting the key metrics used to evaluate O-PUFs. Secondly, a complete set of mathematical definitions and explanations for these metrics. Thirdly, recommendations and considerations for improving the use of certain metrics. Lastly, an introduction to a custom Python package, 'pyopticalpuf', designed to facilitate easy access to analysis code for future research.

The first section evaluates a selection of O-PUF research papers from the past decade, summarising and analysing their primary figures of merit. Key findings from this review include the widespread use of HD-based measurements and the diverse range of metrics employed across different studies. A list of the most frequently used metrics is presented in TABLE I, providing insights into the most popular evaluation methods. The pros and cons of these metrics are also assessed. While previous review papers have provided metric recommendations for general PUFs, O-PUFs are a distinct subcategory that requires unique considerations. This is primarily due to O-PUFs typically deviating from the "Independent and Identically Distributed" (IID) assumption[1]. While a useful assumption and typically true for more conventional electronic PUFs, any PUF can display a certain level of correlation (i.e. to not be IID) and still maintain security, provided the ID produced is complex enough to produce a large database for authentication. In these cases, popular metrics such as reliability and uniqueness, which depend on the means of the HD distributions, may not fully capture the behaviour of O-PUFs. In this paper, we argue that, due to their non-IID nature, the standard deviations of these distributions can deviate from expected values for simple distributions, necessitating the introduction of additional figures of merit when comparing O-PUFs. This paper also provides clear mathematical definitions for the key metrics used in the last decade of research. Following on from this, recommendations are made on the metrics to use for non-IID O-PUFs. Finally, a Python package is introduced, enabling time-of-manufacture and time-dependent testing for various O-PUFs across a range of metrics. The package provides a range of options for binarization, followed by full metric evaluations for the input data. A novel method for time-dependent assessment is also proposed and integrated into the package, alongside other metrics, in the form of a distribution evolution test.

Outline: This review is structured into the following sections. In section II, the metrics used in the literature from 2013 to 2024 are detailed, with popular figures of merit highlighted. Next, section III, titled "Mathematics and Methodology," clearly defines the metrics and summarises the methods required for data collection. The primary methods used for analysing O-PUFs are detailed in subsection III B, which also covers the foundational concepts needed to understand the remaining sections. The main metrics are then presented mathematically, and the methods for obtaining these metrics are discussed. Section IV contains an evaluation of the different metrics, including key findings and important recommendations, leading to the construction of a full set of testing metrics. Following this, section V introduces a Python package that aims to simplify the process of performing a thorough testing procedure for future work. Finally, the conclusion is presented in section VI.

II. PREVIOUS WORK

Since the conceptualisation of PUFs in 2002 by Pappu *et al.*[2], numerous O-PUFs have been developed, ranging from imaging random fibres in paper to analysing quantum dot emission patterns[3,4]. Notably, the first PUF proposed by Pappu *et al.* was an optical PUF, which utilised laser speckles to generate responses. While general testing frameworks for PUFs have been widely explored[5–7] , much of the field's focus has been on electronic PUFs (E-PUFs), operating under the assumption that their metrics can be applied universally to all PUF types.

However, this assumption oversimplifies the challenges specific to optical imaging, particularly when comparing IID and non-IID behaviour. The unique properties of optically-read tags often mean that standard figures of merit used for E-PUFs may not function as intended for O-PUFs. This paper aims to add nuance to this established methodology by evaluating the specific needs of O-PUF testing. To that end, it will first review the primary figures of merit used in O-PUF research over the last decade. Following this, it will critique the most commonly applied metrics and introduce a Python package alongside a set of recommended tests designed to address the distinct requirements of O-PUF evaluation.

Hamming distance means (μ_1 and μ_2)	μ_1 and μ_2 are means of distributions fitted to the intra- and inter-HD histograms, respectively.	III B
Reliability	A measure of how similar repeat measurements of the same tag are.	III C 1
Uniqueness	A measure of how distinct the measured responses from different tags are.	III C 2
Uniformity/Bias	The overall proportion of '0' bits to '1' bits in the binary representation of the response.	III C 3
False Positive Rate (FPR)	How often the measurement of an incorrect tag is mistaken for the correct one.	III C 7
NIST SP-800-22	A suite of pass/fail tests to check the determinability (that is, any hidden patterns) of a random bit-string.	III C 10
Bit Error Rate (BER)	The amount of difference between the measurement of a tag and a measurement of the same tag taken later, typically in the study of response deterioration.	III C 11

TABLE I. A summary of the most popular figures of merit used in O-PUF literature over the past decade, excluding some of the less-cited metrics. Each figure of merit, such as uniqueness, is accompanied by a simplified definition to clarify the metric's general meaning. The final column lists the relevant sections in this document where each metric is discussed in detail.

A brief introduction to the main figures of merit used in the last decade to evaluate O-PUFs are shown in TABLE I.

The metrics found in the literature can be divided into three main categories:

- μ-based metrics - these are defined as metrics which mathematically use only the mean of the HD values for the intra-image and inter-image sets.
- μ & σ-based metrics - similar to the previous category, however these metrics also rely on the standard deviation of the Normal fits placed on the HD histogram for both the intra-HD and inter-HD data (σ_1 and σ_2, respectively).
- Other metrics - This operates as a catch-all category for metrics which do not rely on the HD means or standard deviations.

Sample Size Range (N)	Binomial Scaling $\binom{N}{2}$ (Number of Comparisons)	References
$100 \leq N < 1000$	4950 - 49500	[8](2018);[9](2024);[10](2022); [11](2023);[12](2018)
$50 \leq N < 100$	1125 - 4851	[13](2014)
$20 \leq N < 50$	190 - 1176	[3](2022);[14](2022);[15](2021); [16](2023);[17](2023);[18](2023); [19](2022)
$N < 20$	0 - 171	[20](2020);[21](2021);[22](2018)

TABLE II. Sample sizes (N) used in intra-image and inter-image sets for O-PUF analysis. The sample size directly affects the accuracy of the main figures of merit by determining the number of comparisons used to construct the fractional Hamming distance (HD) distributions. The table also provides the corresponding binomial scaling, $\binom{N}{2}$, which indicates the number of data points on the HD plots, as the HD is derived from pairwise image comparisons.

Another insight from the literature comes from the variation in the quantity of readings taken for the HD fits. The ranges of N values are displayed in TABLE II. As shown in TABLE II, the number of images differs widely between studies, often influenced by factors like the use of simulated data, which allows for higher N values. Since the data points in the HD plots scale binomially, as described in equation 2, the number of pairwise comparisons increases rapidly with larger N. This means that larger sample sizes generally improve the accuracy of the Normal fits applied to the HD data, potentially leading to more reliable metrics. It is worth noting that this table assumes all combinations of comparison within and across PUF instances were performed to calculate the HDs.

Now that the main metrics have been generally defined, tables III shows the more popular metrics - those with four or more instances of use in the literature. Table IV then shows other, less popular metrics. This work shows the prevalence of μ-based metrics for the HD analysis, with less-common use of μ & σ-based metrics. With an additional clear presence of 'other' metrics for longevity and randomness testing. There is a clear, varied and wide range of metrics used in different combinations in different papers in the literature.

13			[9](2024)				
12			[11](2023)				
11			[16](2023)				
10			[3](2022)				
9			[23](2022)				
8			[21](2021)			[16](2023)	[18](2023)
7			[15](2021)			[3](2022)	[16](2023)
6	[18](2023)		[24](2020)		[18](2023)	[23](2022)	[17](2023)
5	[25](2022)	[14](2022)	[26](2019)		[23](2022)	[24](2020)	[19](2022)
4	[27](2020)	[24](2020)	[22](2018)	[16](2023)	[3](2022)	[27](2020)	[25](2022)
3	[26](2019)	[26](2019)	[28](2017)	[11](2023)	[15](2021)	[22](2018)	[3](2022)
2	[29](2016)	[30](2015)	[30](2015)	[31](2022)	[12](2018)	[29](2016)	[22](2018)
1	[30](2015)	[13](2014)	[10](2022)	[13](2014)	[22](2018)	[30](2015)	[28](2017)
	Uniqueness	Reliability	HDmeans	FPR*	NIST	BER*	Uniformity*
	μ-based metrics			$\mu\&\sigma$ metrics	Other metrics		

TABLE III. Main figures of merit tested are shown for O-PUF papers reviewed from the last decade of research. This shows the most popular metrics by listing the papers, including references. 'μ-based' metrics are based on the means of the HD distributions, while '$\mu\&\sigma$-based' are based on the mean and standard deviations. The category 'other' describes the range of metrics that do not fit the first two categories. Metrics with an asterisk have been used for time-dependent metrics.

2		[10](2022)		[9](2024)	[25](2022)	[10](2022)	
1	[3](2022)	[8](2018)	[31] (2022)	[28](2017)	[29](2016)	[8](2018)	[30](2015)
	ENIB	Prob of CI*	Accuracy	Bit-aliasing*	Randomness	Robustness	CRP
	$\mu\&\sigma$-based metrics			Other metrics			

TABLE IV. Uncommon figures of merit. '$\mu\&\sigma$-based' are metrics based on the mean and standard deviation of the HD distributions. The category 'other' describes all other kinds of figures of merit. The metrics with an asterisk have been used in the literature for analysis of time-dependent performance.

1. *Key findings*

- **Wide variation in metrics**: There was no clear consensus in the literature on the key metrics used for time-of-manufacture or time-dependent testing. A wide range of metrics were employed in the assessment of O-PUFs.
- **Ubiquitous use of HDs**: HDs form the basis for a subset of metrics used to investigate the properties of the ID. Of the papers surveyed, all calculated sets of HDs using intra-image and inter-image sets as part of their analysis.
- **Use of HD for fitting distributions**: HD distributions are fit to the histograms of the two sets of HD data. These Normal distributions have means and standard deviations that are used to calculate some metrics. While μ-based metrics do not require distribution fitting, μ-and-σ metrics do. Therefore, not all studies will fit distributions or display HD histograms if their metrics do not involve σ.
- **Other metrics**: Some metrics are not based on HD; these metrics assess a wider range of properties of the O-PUF. An example is bit-aliasing, which checks for repeating image patterns in the inter-image dataset. These metrics are usually used in addition to the HD methods.
- **Use of NIST-800-22**: A commonly used metric that does not rely on HD is the NIST-800-22 testing suite. Developed by the National Institute of Standards and Technology (NIST), this suite of 15 tests is used to assess the level of randomness in a 1D binary bit-string.
- **Variation in methods**: The literature highlights variation in methods used to test O-PUF data. A key difference lies in image binarization techniques, ranging from simple hashes to complex algorithms, which significantly influence data quality and metric results. Additionally, the number of intra-image and inter-image samples varies, with simulated studies typically using larger image sets (N) than experimental ones.
- **Impact of previous reviews**: Although a review specific to O-PUFs is lacking, general PUF reviews have influenced metric selection in the literature. Commonly used metrics include reliability, uniqueness, bit-aliasing, and uniformity, though studies often focus on one or two rather than the full set. Previous reviews are generalised for all types of PUF, a specific review is required.

A. Previous review and recommendations papers

Over the past decade, several significant reviews have shaped the understanding of PUF metrics, but a gap remains in addressing the unique requirements of optical PUFs, which often diverge from IID assumptions. In 2013, Maiti et al.

consolidated four primary metrics-reliability, uniqueness, bit-aliasing, and uniformity- designed to assess PUFs across dimensions such as time and bit position[5]. A 2014 tutorial emphasised intra-image and inter-image PUF evaluations using Normal fits and false positive/negative rate measurements, establishing a framework for PUF testing[7]. In 2016, Vijayakumar et al. refined uniqueness testing through inter-fHD metrics and advanced analysis procedures[32].

A 2020 paper on quantifying PUFs proposed additional metrics, including steadiness and diffuseness, while referencing definitions from earlier reviews[33]. In the same year, a conference talk examined the evaluation of electronic PUFs as IID systems, focusing on reliability, uniqueness, and randomness[6]. Finally, Wilde's 2021 PhD thesis offered an extensive discussion of potential metrics, though it primarily concentrated on electronic PUFs adhering to IID assumptions[34].

For open-source code solutions, notable progress includes the creation of the 'Pypuf' Python package in 2021 for general PUF analysis[35], and a MATLAB GUI designed to evaluate the four primary metrics highlighted by Maiti *et al.* in 2013[36]. Building on these efforts, this work introduces the 'pyopticalpuf' Python package, specifically tailored for the analysis of O-PUFs. As most PUF research is rooted in computer science and focuses on E-PUFs, this work aims to support scientists in material science, chemistry, and related fields by providing tools to quantify the cryptographic potential of O-PUFs with minimal coding effort.

III. MATHEMATICS AND METHODOLOGIES

This section begins by looking at the IID assumption of a PUF's response, which is often a baseline assumption for metric choice in PUFs. Next, the mathematical building blocks needed to understand these metrics are introduced. Then, key mathematical definitions for the most commonly used O-PUF evaluation metrics are stated. These metrics are categorised into μ-based or $\mu\&\sigma$-based metrics. These are derived using the means and standard deviations of a Normal fit to the HD curves for intra-HD and inter-HD values (section III B). Additional metrics, such as NIST-800-22 randomness testing, are then described. Finally, we analyse metrics used for time-dependent evaluation in the literature (section III E).

A. Why O-PUFs should be treated differently from E-PUFs

Firstly, let us looks at the IID assumption. In the context of a binary sequence, also known as a binary bit-string, composed of 1s and 0s - it means:

- **Identically distributed**: Each bit in the sequence follows the same probability distribution. This ensures that no specific bit position is inherently more likely to be 0 or 1 than another.
- **Independently distributed**: The value of any given bit does not depend on the values of other bits. In other words, knowing the previous bits does not help predict the next bit[37].

When the IID assumption holds, pairwise comparisons of binary sequences, such as fHD calculations, follow a predictable binomial distribution. This can be approximated by a normal distribution, where the mean (μ) and standard deviation (σ) have a fixed mathematical relationship: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, where n is the number of comparisons and p is the probability of success[1]. However, without the IID assumption, as in optical PUFs, this relationship breaks down.

Another key definition to clarify is the boundaries of the different classifications of types of PUFs:

- **E-PUFs:** These PUFs rely on variations in semiconductor manufacturing processes, such as differences in transistor threshold voltages or circuit delays, to generate unique and unclonable responses. Common examples include: Ring Oscillator PUFs[38], SRAM PUFs[39] and Arbiter PUFs[40].
- **O-PUFs:** Also known as 'optically-read' or 'optically-imaged'. These PUFs leverage optical properties of a material to gain a imaged response. The complexity of light propagation from emission or variation in the manufacture processes - which is picked up in the imaging process, allows for an unclonable response. Common examples include: Quantum Dot PUFs[4,41] and laser speckle[2,42].

Both E-PUFs and O-PUFs enhance security by generating unique responses to challenges. One of the distinctions is that while E-PUFs rely on electronic responses and therefore can be implemented at the infrastructure level for large-scale security, O-PUFs produce optical responses and can therefore be integrated at a consumer-level. This enables new consumer-driven O-PUF authentication techniques, particularly through imaging technology such as smartphones, which have been explored in recent studies[43–49].

O-PUF characteristics affect compliance with the IID assumption. Locally correlated features present in the images captured to produce O-PUF responses, post-binarization, can result in runs of 1s or 0s - as shown in FIG. 3.

This paper also suggests that testing with the NIST-800-22 suite for randomness may yield a higher failure rate for O-PUFs compared to E-PUFs. This is because NIST-800-22 tests require unbiased and IID bit-strings. IID is also not essential for a valid PUF, as long as a functional fingerprint of sufficient uniqueness can be maintained and identified then the core function of the PUF in this context, which is authentication, can be performed. O-PUF studies use differing binarization and post-processing techniques that help to reduce noise and improve ID between repeated images. This can vary the level of correlation between adjacent pixels in the array, in turn affecting the IID nature of the data[50]. It is noteworthy that some O-PUFs in the literature pass NIST-800-22[3,12,15,18,22,23]. We can hypothesise that with specific setups or feature-size-to-pixel-size ratios these studies can achieve IID and demonstrate favourable NIST-800-22 results. However, the majority

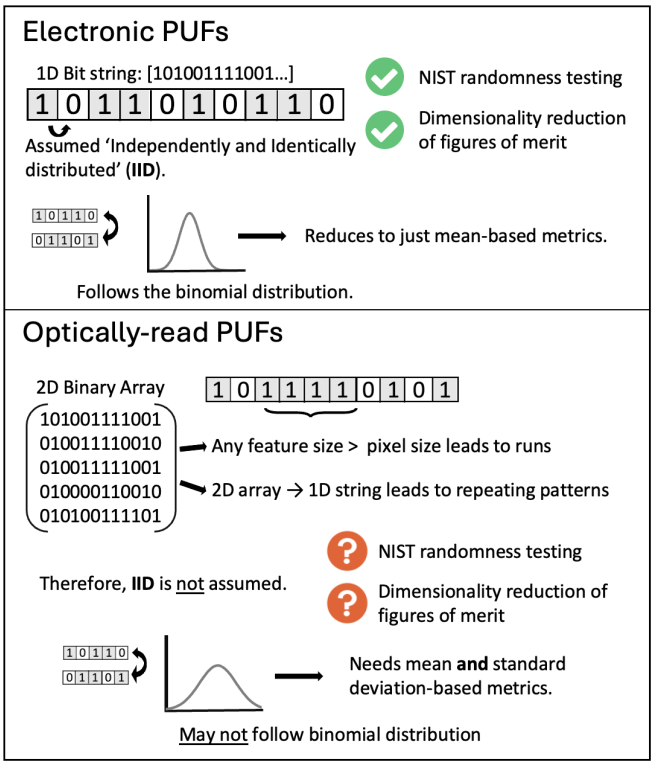


FIG. 3. Comparison of binary responses from Electronic PUFs and Optically-read PUFs. The figure illustrates how the IID assumption holds for E-PUFs, resulting in predictable Binomial distributions and passing NIST-800-22 test results. In contrast, O-PUFs do not normally adhere to the IID assumption, this is due to 2D structure and features in the binary arrays from their responses. This necessitates considering both the mean and standard deviation in comparison metrics. In this context, 'dimensionality reduction of figures of merit' refers to how a full curve fit will be reduced to a single summary statistic such as the mean (reliability).

of the examined O-PUF papers avoid NIST-800-22 testing, as sufficient uniqueness and reliability can be demonstrated without the randomness indicated by NIST-800-22 tests.

Since the standard deviation is not directly linked to the mean mathematically, the traditionally used mean-only metrics - such as reliability and uniqueness, often employed as primary summary statistics in PUF analysis - may overlook crucial information, leading to potentially flawed conclusions. Incorporating metrics that account for both the mean and standard deviation is therefore essential for non-IID PUFs.

B. Mathematical building blocks

This section outlines the foundational methods and concepts required for O-PUF analysis. These building blocks serve as a basis for the mathematical definitions of the main figures of merit, which are detailed in the following section. The discussion begins with an explanation of the method used for calculating HD.

Firstly, let's outline the general process outlined in FIG. 4 and FIG. 5. To start, consider two sets of images: 'Intra-images', a set of images which are repeated images of individual tags and 'Inter-images', a set of images of different tags. These two sets are binarized using a binarization algorithm, of which there are many options. This ensures a clearer ID and minimises noise. Following this, each set of arrays is then used to plot a HD histogram, to which a Normal distribution can be fit. Not all metrics use this approach, but the final step can be replaced with alternative analysis. However, HD calculations were found in most of the papers reviewed over the past decade. Additional metrics that do not rely on HD values are often incorporated alongside these calculations.

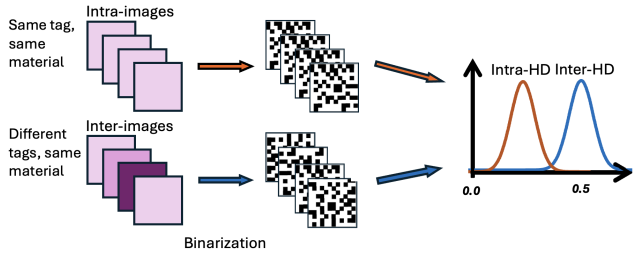


FIG. 4. Two image sets are collected for analysis: the intra-images set, consisting of repeated readings of a specific ID, and the inter-images set, containing images of different tags. After binarization, which converts the images into 2D binary arrays, the intra-HDs and inter-HDs are plotted on a graph. Normal fits are then applied to these plots to facilitate the extraction of key metrics for evaluating O-PUF performance.

For optimal results, one should choose a large N for the inter-image set and a set of intra-images such that both HD histograms contain a similar number of data points. Methods for collecting intra-images vary; some studies use small sets of repeat measurements from each ID to capture a wider range of data variation, while others use a larger set of images from a single ID for the intra-image set. The first approach benefits from its breadth, offering a more comprehensive view of the dataset's ability to provide a unique ID for each piece. However, this method can be more labour-intensive, depending on the acquisition process. In contrast, the second method allows

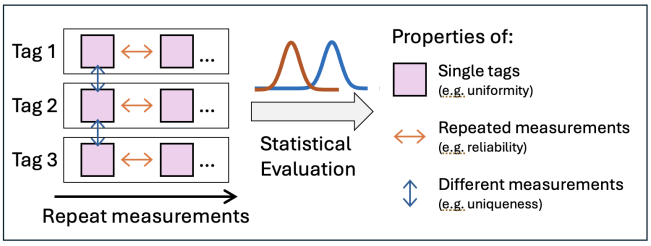


FIG. 5. General schematic illustrating the differences in data collection for each O-PUF candidate. Separate tags are fabricated and tested, with repeated imaging of each tag enabling comparisons within the same tag and between different tags. This approach provides a comprehensive statistical understanding of the O-PUF candidates' behaviour.

for a more in-depth analysis of a specific example of an ID.

FIG. 6 illustrates the definitions of the intra-array and inter-array sets, C and D, which are derived from the binarized intra-image and inter-image data. These arrays, represented as binary matrices, require precise mathematical characterisation to ensure clarity in the subsequent definitions of the metrics.

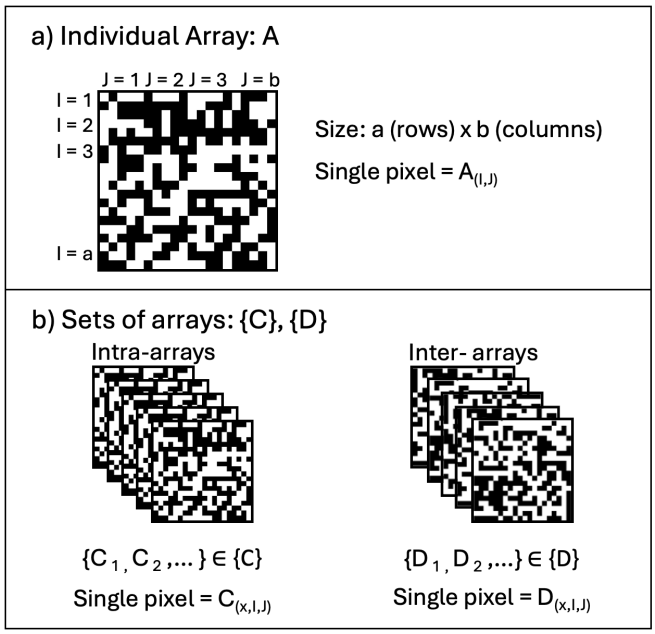


FIG. 6. a) A generic array A is shown in, in which a single pixel is denoted as $A_{(I,J)}$. Arrays of size $a \times b$ are shown, with elements indexed by I and J corresponding to positions of the pixels in the array. b) The intra-arrays form set $\{C\}$, consisting of repeat binarized images of the same tag. The inter-arrays are represented in set $\{D\}$, these are the arrays produced from the images of different tags. These arrays, composed of 1s and 0s, represent the binary IDs of the samples.

The key mathematical comparison performed on the arrays in set $\{C\}$ and $\{D\}$ is the fractional hamming distances, fHD. This mathematical operation can be understood in one of two ways. As the proportion of bits that would need to be flipped to go from one array to another or as a 2D version of the pixel-by-pixel application of XOR gate rules. fHD is 'fractional' due to its normalisation to the bit string length.

As shown in the FIG. 7, the fHD is calculated between two arrays and produces a single metric of similarity/difference. The two arrays are compared using XOR rules, pixel by pixel. Then the average is found of this bit-string.

Mathematically, this fHD operation is performed on a pair of images in this form:

$$fHD(A,B) = \frac{1}{ab}\sum_{I=1}^{a}\sum_{J=1}^{b}(A \oplus B)_{IJ} \tag{1}$$

where A and B are the binary arrays representing two images being compared, a and b are the dimensions of these arrays, and $\oplus$ denotes the XOR operation. The summation is carried out over all pixel positions I (1, 2, ..., a) and J (1, 2, ..., b) in the images. The result, represents the similarity between the two images. This calculation is later applied to full sets of arrays, such as the $\{C\}$ intra-array set and then the mean value of the fHD is found for the full set.

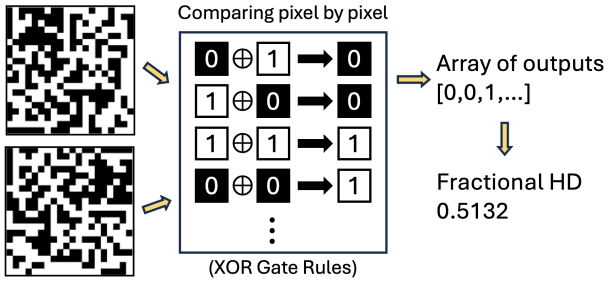


FIG. 7. Calculation of the fHD value using two pixel arrays of identical dimensions. Each corresponding pixel is compared using the XOR operation, resulting in a binary bit-string of 0s and 1s. The mean of this bit-string represents the fHD, which ranges from 0 (identical arrays) to 1 (completely opposite arrays).

The meaning of the different values for HD are shown below in TABLE V.

Hamming Distance	Meaning
0	All pixels are the same; the two 2D arrays are identical.
0.5	50% of the pixels are different.
1	All pixels are the opposite; the two 2D arrays are opposite.

TABLE V. Fractional Hamming distance calculations between two arrays produce results in the range of 0-1. The value for the calculations determines the relationship between the two different arrays.

Another building block to understand is how the quantity of data points for the fHD plots scales proportionally with the number of images captured, denoted as N, for each set of tags. This is because each fHD is derived from comparing a pair of arrays. Hence, the total number of combinations for N images is represented by the binomial coefficient.

$$C(N,2) = \binom{N}{2} = \frac{N!}{2!(N-2)!} \tag{2}$$

where $C(N,2)$ denotes N choose 2. This exemplifies a binomial distribution where N signifies the quantity of images in both intra-images and inter-images sets. The escalating count of combinations with increasing N allows for substantial data scaling with manageable experimental time frames.

These fHD values are represented in the fHD distribution fits shown below in FIG. 8. These distributions are then used to determine the means and standard deviations of the two sets $\{C\}$ and $\{D\}$ which are then used as the mathematical base of future metrics.

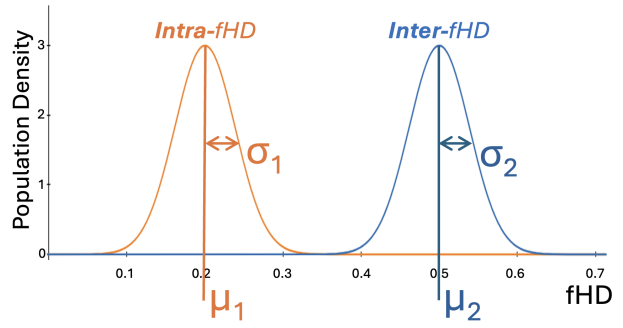


FIG. 8. Distributions of intra-arrays and inter-arrays fHD. Intra-fHD values typically range from 0 to 0.4, while inter-fHD values cluster around 0.5. The histograms are generally approximated by a Normal distribution, which facilitates the calculation of mean and standard deviation values utilised in subsequent metrics. This is due to 0 fHD signifying identical arrays and tending towards 0.5 signifying randomly different arrays. For the intra-fHD distribution, μ_1 is the mean and σ_1 is the standard deviation. For the inter-fHD distribution, μ_2 is the mean and σ_2 is the standard deviation.

C. Metrics

Now that the core building blocks have been discussed, they can be put together to explain the key metrics for O-PUF analysis. Firstly, fHD is used to directly make a set of metrics from the means and standard deviations of the distributions that are fit from the fHD histograms.

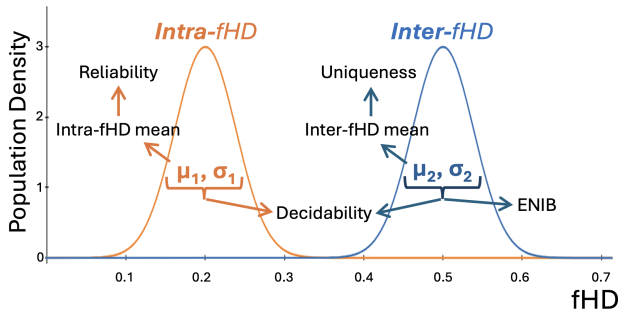


FIG. 9. A summary of the mean ($\mu-based$) and means & standard deviation ($\mu\&\sigma-based$) metrics mentioned in this paper and how they are taken from the Normal fits of the intra-fHDs and inter-fHDs plots. Reliability is based on the intra-fHDs mean, while uniqueness is based on the inter-fHD mean. Decidability is based on both means and standard deviations. ENIB is based on the inter-fHD mean and standard deviation. For the intra-fHD distribution, μ_1 is the mean and σ_1 is the standard deviation. For the inter-fHD distribution, μ_2 is the mean and σ_2 is the standard deviation.

FIG. 9 shows how the fHD-based metrics are related to each other and the ways in which they are interdependent. With the intra-fHDs curve producing the reliability and part of the decidability and the inter-fHDs producing ENIB, uniqueness and part of the decidability. Then the whole plot is used for FPR calculations.

Note that in this set of mathematical definitions, the 2D version - where the input for the calculations consists of binary arrays - has been chosen over the 1D bit-string version. If the 2D arrays are collapsed into a 1D bit-string, the 1D formalism becomes equivalent.

1. Uniqueness

Uniqueness is a metric based on the inter-fHD mean, measuring the difference in IDs produced from measurements of the different tags. For two tags, in the set of inter-arrays $\{D\}$ which contains N images total, sets of two tags D_i and D_j are compared using the fHD function, described in equations /ref. These fHD are accomplished in all unique pairs in set $\{D\}$. The value μ_2 is found as the mean of these fHD measurements. To find uniqueness, the value of μ_2 must be determined:

$$\mu_2 = \frac{1}{\binom{N}{2}} \sum_{x=1}^{N-1} \sum_{y=x+1}^{N} fHD(D_x, D_y) \tag{3}$$

where μ_2 is the inter-HD mean, N is the number of inter-images. D_x and D_y are arrays in the inter-array set $\{D\}$. N is therefore also the total number of arrays in the set $\{D\}$.

$$\text{Uniqueness} = \mu_2 \text{ x } 100\% \tag{4}$$

where the ideal value for this parameter, μ_2 is 0.5. This is referred to as 'uniqueness' as if the inter-fHD mean is 0.5 that means that the N different tags produce fHDs around 0.5. This shows that the tags are unique and the pixels are 50% different, showing random variation[51].

Metric	Simple Definition	Ideal Value
Uniqueness	(μ_2) inter-fHD mean x 100%	50%

2. Reliability

Reliability is based on the intra-fHD mean and is therefore a measure of the difference between the readings of the same ID. This metric uses μ_1, which is calculated using the fHD between all combinations of pairs of binarized 2D arrays in the set of intra-images. The set $\{C\}$ is composed of N images, with x and y representing the arrays in the set that are paired. Reliability is considered a time-dependence-negligible metric, while the time-dependent evaluation of μ_1 discussed when considering BER, in section III C 11, as these readings are taken over a set and non-negligible time intervals. The mean of the fHD between the arrays in set $\{C\}$ is μ_1:

$$\mu_1 = \frac{1}{\binom{N}{2}} \sum_{x=1}^{N-1} \sum_{y=x+1}^{N} fHD(C_x, C_y) \tag{5}$$

$$\text{Reliability} = (1-\mu_1) \times 100\% \tag{6}$$

where μ_1 is the mean intra-fHD. Ideally, μ_1 would be 0.0, indicating identical IDs from repeated measurements of the same tag. However, achieving an intra-fHD mean of 0.0 is challenging for reasonably sized IDs due to factors such as noise, alignment errors, and other mitigating influences, which increase the intra-fHD between IDs.

Metric	Simple Definition	Ideal Value
Reliability	(1 - (μ_1) intra-fHD mean) x 100%	100%

3. Uniformity

Unlike the previous two metrics, uniformity is not μ-based. It is simply a measure of the binary bias of the 2D array. Generally, it is used to indirectly represent if the array contains enough information to be considered a good fingerprint. Binary IDs with a uniformity that is much higher or lower than 50% are likely too uniform, mostly white or black, to contain sufficient information. Uniformity for a single array D_n is defined below as U_n:

$$\mathrm{U}_n = \frac{1}{a \text{ x } b} \sum_{I=1}^{a} \sum_{J=1}^{b} D_{n,I,J} \tag{7}$$

$$\mathrm{U} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{U}_n \tag{8}$$

where the binary array D_n is composed of a rows and b columns and a pixel in this array is $D_{n,I,J}$ where (I,J) are the pixel coordinates. U_n will be a number between 0-1 and represent the average pixel value of the array. For the use in the categorisation of O-PUFs, the set $\{D\}$ is often used. Also known as 'Hamming weight' or 'bias,' uniformity quantifies the proportion of 0s and 1s in the binary fingerprint produced by the O-PUF. A value of 50% indicates maximum information density, with an equal probability of obtaining a 0 or 1 at each pixel location, resulting in the highest entropy. To further assess uniformity, it may be useful to evaluate subsections of the image to identify local deviations from the expected distribution. This approach provides additional insight into the consistency of the fingerprint's information density. As shown in FIG. 10, a white fingerprint has a uniformity of 0%, while a fully white square yields 100%.

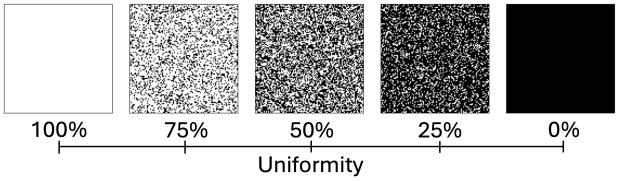


FIG. 10. Illustration of pixel array uniformity. A uniformity value of 100% corresponds to an array where all pixels are white, while a value of 0% represents a fully black square. An array with 50% white and 50% black pixels has a uniformity value of 50%.

Metric	Simple Definition	Ideal Value
Uniformity	0/1 pixel bias for individual tags	50%

4. Bit-aliasing

Bit-aliasing is a metric used to check for features that repeat over arrays which should be randomly varying. It can be useful for checking for errors in the O-PUF production process. Bit-aliasing, in the context of O-PUFs, is used to check that specific pixels in the inter-array set $\{D\}$ are not repeating or biased towards 0 or 1. The bit-aliasing at specific pixel (I,J) is the percentage mean value of the specific binarized pixel, which can hold the value 1 or 0:

$$(\text{Bit - aliasing})_{I,J} = (\frac{1}{X} \sum_{x=1}^{X} D_{x,I,J}) \times 100\% \tag{9}$$

In the time-dependent case, X=T representing the total number of time measurements and x=1 is replaced with t=1. In the other case, where bit-aliasing is used to determine patterns in the inter-image dataset, X represents the number of inter-arrays. $D_{x,I,J}$ represents the pixels in the inter-array set. Note that even in the time-dependent case, the $\{D\}$ set is used and bit-aliasing is performed over different tags so that the ID should not be the same in each instance.

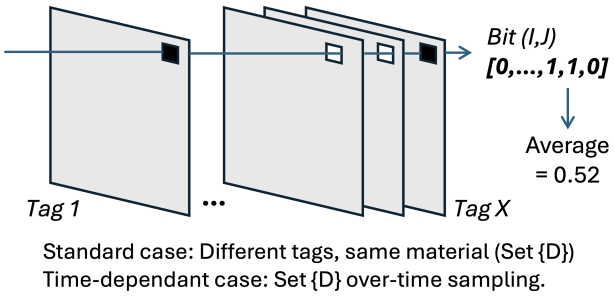


FIG. 11. Bit-aliasing is assessed across a specific pixel (I,J) in different O-PUF responses. An average is then found, the example value of 0.52 implies no pattern between the tags. This would demonstrate the IDs are good unique objects. This is often done over set $\{D\}$, the inter-arrays, which are different tags. An additional case involves sampling of the D set over time to check for anomalies.

Estimating the bias of a particular response bit across several tags provides insight into any systematic or spatial effects present in the fingerprint. Bit-aliasing is a simple correlation test that does not require considering the order of the tags. Ideally, there should be no positional relationship with bias, meaning each specific area of the fingerprint will have an equal likelihood of having a 0 or 1. As shown in FIG. 11, a specific bit is tested over the inter-array set to check for any inconsistencies or repeating patterns in the data.

While bit-aliasing tests are straightforward, they can reveal underlying issues and systematic errors. In the context of O-PUFs, repeating bit-specific errors may result from factors like dust on the lens, faulty pixels in the measuring device, or manufacturing defects during fabrication. However, since the

manufacturing process involves a sequential or ordered production of tags, more complex correlation testing might be necessary to diagnose issues. In such cases, checking for correlations based on the order of manufacture could help detect patterns that simple bit-aliasing analysis may overlook.

This is a good example of a metric primarily applicable to E-PUF testing and less relevant to O-PUFs. In the case of randomly generated electronic signals an issue with the system can cause a stuck bit in the bit-string that will always fall one way or another over different tags - which should vary randomly, this was the reason for including this metric in general PUF analysis. However, this is much less likely with imaging setups and any issues would be visible by eye in the image data sets - therefore this metric is not often used in O-PUF data.

Metric	Simple Definition	Ideal Value
Bit-aliasing	0/1 pixel weighting for a specific pixel between different tags	50%

5. *Effective number of independent bits*

ENIB, initially created for evaluation of the human iris by J. Daugman[52], is a measure of usable information content stored in the tag and is calculated based on the inter-fHD distribution mean and standard deviation. To calculate ENIB the formula is as follows:

$$N = \frac{\mu_2(1-\mu_2)}{{\sigma_2}^2} \tag{10}$$

where N is the ENIB, μ_2 is the mean of the inter-fHD distribution and σ_2 is its standard deviation. By extension, $(\sigma_2)^2$ is the variance. Note that in some literature, such as the 2022 paper by Kim *et al.*[3], this metric is referred to as *'Degrees of Freedom (DoF)'*.

ENIB provides a measure of the randomness and independence of bits within a tag, with higher values indicating that a single PUF is effective in providing a good ID for security applications. ENIB is maximised when the mean of the inter-fHD distribution (μ_2) is near 0.5, as this reflects maximum random variation between tags, ensuring the bits are highly independent. Conversely, as the mean deviates from 0.5, ENIB decreases, signalling less randomness and a reduction in the effective number of independent bits. Additionally, a narrower inter-fHD distribution, characterised by a lower standard deviation (σ_2), increases ENIB. This narrower distribution indicates that most values are clustered closer to 0.5, further supporting the randomness and independence of the bits within the tag.

Metric	Simple Definition	Ideal Value
ENIB	Number of effective independent bits in the IDs, calculated based on μ_2 and σ_2	Higher

In contrast to the directly binary nature of electronic PUF logic levels, Optical PUFs often encode from a much larger amount of information per unit, or pixel, of unprocessed response. To capture and compare this level of unique input availability to the PUF response, the metric of encoding capacity can be used. This is specifically relevant for multi-factor O-PUFs as showcased in recent studies[16,53–55]. This metric quantifies the total number of possible unique PUF responses, and can be calculated as the size of the permutation space of the response of a PUF. A common formula for this calculation is as follows:

$$\text{Encoding capacity} = C \times L^N \tag{11}$$

Where L is the number of possible response states for each unit or pixel of the response, N is the total number of units or pixels that make up that response, and C is the number of separate channels over which this response is being measured[56]. For a digitised binary response L=2, with 0 and 1 being the two outcomes, however for a multi-factor PUF that produces three separate output states per pixel L=3 and so on[16]. If these 2 or more response states are being extracted from a single channel, for instance in greyscale, then C=1, but if the same process is being applied over 3 separate channels, such as each of some RGB channels, then C=3. For a 100x100 response array N=10'000, for instance. It is worth noting here that this metric does not address the probability of each response permutation outcome and assumes each response is equally likely. Deviations from equiprobability can be addressed either after some form of binarization using the other metrics in this work, or directly using the state-likelihood based entropy estimation metric found in section III C 12. In accounting for correlations, the entropy estimation (or ENIB where $L = 2$) for the number of bits (or state units) in practice can then be represented in the encoding capacity equation by modifying the number of response units N into the *effective* number of response units. In doing so, the number of channels and response unit states can remain in consideration alongside any deviation from ideal unpredictability (or IID distribution) in states or responses.

6. *Decidability*

Used in the field of human iris recognition[57], decidability is assessed using curve fits on fHD histograms for two image sets. Decidability is calculated as follows:

$$d' = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{{\sigma_1}^2 + {\sigma_2}^2}{2}}} \tag{12}$$

where d' represents decidability, μ_1 is the mean intra-fHDs, μ_2 is the mean inter-fHD, σ_1 is the standard deviation of the intra-fHDs, and σ_2 is the standard deviation of the inter-fHD. A decidability of zero indicates no distinguishable features between tags, rendering them ineffective as O-PUFs. In the iris study, decidabilities as high as 14.7 were observed, with a threshold as low as 7.2 for practical use[57].

Decidability provides a concise summary statistic that encapsulates the relationship between the intra-fHD and inter-fHD distributions in a single value. Unlike ENIB, which focuses solely on the inter-fHD distribution, decidability evaluates both distributions simultaneously to determine how distinguishable they are. Decidability is optimised when the difference between the means (μ_1 and μ_2) is large, indicating greater separation between the distributions. Additionally, lower standard deviations (σ_1 and σ_2) improve decidability by narrowing the distributions, making it easier to discern between the two. This aligns with the purpose of decidability: to quantify how effectively the two distributions can be separated for reliable identification or authentication.

Metric	Simple Definition	Ideal Value
Decidability	Quantifies how easy it is to decide between one ID and another	Higher

7. *False positive rate*

The false positive rate (FPR) quantifies the proportion of incorrect positive identifications, representing instances where the system incorrectly classifies an incorrect PUF evaluation as a correct one, and is a key metric for evaluating system misclassification likelihood over time. Favourable FPR+ values will mean effective authentication during the verification stage if the O-PUF is used in industry. Depending on the type of FPR, the value can be derived from the Normal distribution of intra-fHD and inter-fHD areas or from True/False readings from testing. Note that in this section FPR+ is used to refer to the full set of 4 metrics found in the confusion table.

When calculating FPR it is important to consider the following:

- **Experimental FPR**: A threshold point must be calculated: an fHD value used to classify results. If an incoming fHD calculation is less than this threshold, it is categorised as the same tag; if it is greater, it is categorised as a different tag. Test this with a large sample set of intra-arrays (if only doing FPR) and inter-arrays (if expanding to fill the rest of the confusion table). As shown in FIG. 12 and TABLE VI.
- **Theoretical FPR**: Using a pre-measured set of fHD values the means and standard deviations are determined. Then using the fitted curves, new fHD measurements are converted into FPR+ using the areas under the curves. As shown in FIG. 12, part **(b)**.
- **Readings over time**: Using either method the FPR+ over time can be determined. Therefore, the performance of the fingerprint over time can be plotted and evaluated.

Mathematically, the experimental FPR+, as shown in TABLE VI, is calculated using proportions of areas. Depending on the specific research or commercial application, different metrics may have different importance, but in general lower FPRs are most useful for reliable security.

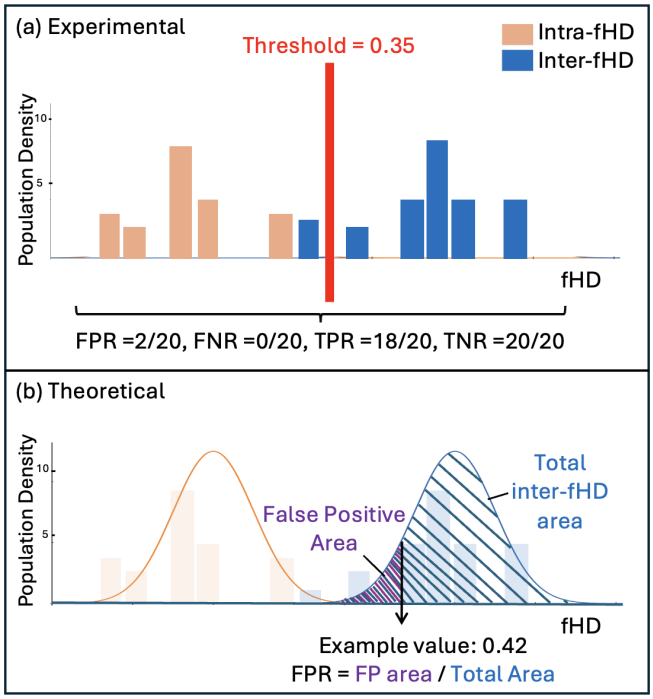


FIG. 12. In panel (a), *Experimental FPR+* are calculated on a set of data, whereas in panel (b) *Theoretical FPR+* are calculated. In this example, 30 tests are considered. The values for both sets of rates are different. The Normal fits in (b) are previously generated before the testing phase and are used to calculate theoretically FPR+ as this is based on the areas of the fitted curves to the left and right of the fHD value.

TABLE VI. Table displaying the formulas used to calculate the experimental rates for False Positive Rates (FPR), True Positive Rates (TPR), True Negative Rates (TNR), and False Negative Rates (FNR). This table summarises key performance metrics delivered from the confusion matrix. FPR represents the proportions of negatives incorrectly identified as positives, TPR measures the proportions of correctly identified positives. TNR captures the correctly identified negatives and FNR indicates the proportion of positives that were missed.

	Predicted Positive	Predicted Negative
Actual Positive	$\text{TPR} = \frac{N_{tp}}{N_{tp}+N_{fn}}$	$\text{FNR} = \frac{N_{fn}}{N_{tp}+N_{fn}}$
Actual Negative	$\text{FPR} = \frac{N_{fp}}{N_{fp}+N_{tn}}$	$\text{TNR} = \frac{N_{tn}}{N_{fp}+N_{tn}}$

Where N is the total number of tests of the tags and N_{fp} is the number of inter-images tested that were incorrectly identified as intra-images, while N_{tp} is the number of inter-images correctly identified as inter-images. N_{tn} is the number of intra-images correctly identified as intra-images and N_{fn} is the number of intra-images incorrectly identified as inter-images. For context, N_t is the total number of intra-images and N_f is the total number of inter-images. This method requires large quantities of testing for high levels of accuracy and precision in the FPR, but can be considered a more authentic reflection of the performance of the O-PUF than theoretical FPR. This could be considered *experimental* FPR as it is tested directly

through experiment.

In order to calculate experimental FPR, a threshold point must be calculated or decided. This value will provide the points for the threshold for which the True/False decision is made. Mathematically, this can be calculated as the x value of the intercept between the two distributions. The threshold point can also be adjusted to adapt an O-PUF to a specific application that requires different proportions of FPR+.

The *Theoretical FPR* can be calculated from as little as one reading or comparison, and is computed by comparing the result to previously fitted curves, as shown in FIG. 12. Since this metric relies on the curve fits of the data, it benefits from a large sample set to ensure the accuracy of these fits. The FPR is considered theoretical because it is computed using the following formulas:

However, it should be noted that the *Theoretical FPR* is not simply the height at a point x on one curve over the sum of the heights of the two curves. Instead, it is the area under the "false" curve (the inter-array distribution) to the left of point x, normalised by the total area under both the "true" and "false" curves. Mathematically, this is expressed as:

$$f_{\text{inter}}(x) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \tag{13}$$

where μ_2 is the mean of the inter-fHD distribution and σ_2 is the standard deviation. Equation 13 shows a normal distribution. Taking this area the following proportion is calculated:

$$\text{Theoretical FPR}(x) = \frac{\int_{-\infty}^{x} f_{\text{inter}}(x)\,dx}{\int_{-\infty}^{\infty} f_{\text{inter}}(x)\,dx} \tag{14}$$

where this equation computes the proportion of the "false" curve's area (to the left of x) relative to the total area under both the "inter" (false) and "intra" (true) curves.

Finally, the mean theoretical FPR across all samples can be computed as:

$$\text{Mean Theoretical FPR} = \frac{1}{n}\sum_{i=1}^{n} \text{TheoreticalFPR}(x_i) \tag{15}$$

Where x is the value of the measurement fHD and '$False_{inter}$' represents the inter-fHD distributions. n is the number of readings that the mean is taken over. The Normal distributions are integrated over to gain area under the curve. The integral on the top of the fraction represents the parts of the false curve that lies to the left of the fHD measurement, x, over the area of the total false curve. Theoretical FPR is calculated as the mean of multiple readings. In the summary below, x is the fHD reading inputted into the equation 15.

Metric	Simple Definition	Ideal Value
Experimental FPR	The fraction of $\frac{N_{fp}}{N_f}$	Lower
Theoretical FPR	Area <x in false distribution divided by total area of false distribution	Lower

8. Accuracy, precision & recall

Used in a small number of the papers surveyed[31], these metrics are based on the probability rates from the mechanisms shown in the above section. It should be made clear when these are used if they are calculated with *theoretical* FPR etc. or *experimental* FPR. These metrics are calculated using these forms:

$$\text{accuracy} = \frac{TPR+TNR}{TPR+TNR+FNR+FPR} \times 100\% \tag{16}$$

$$\text{precision} = \frac{TPR}{TPR+FPR} \times 100\% \tag{17}$$

$$\text{recall} = \frac{TPR}{TPR+FNR} \times 100\% \tag{18}$$

where TPR are the true positive rates, TNR are true negative rates, FPR are false positive rates and FNR are false negative rates. By converting the rates into easy-to-understand percentage metrics on a scale of 0-100, this allows for more comparable and easy-to-understand communication of performance.

9. Probability of cloning

Used in a limited number of the papers reviewed, this niche metric quantifies the area of overlap between the two fitted curves[10]. In this case, Normal fits are presented as the distribution shapes.

$$f_{\text{intra}}(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \tag{19}$$

$$f_{\text{inter}}(x) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \tag{20}$$

$$P_{\text{cloning}} = \int_{-\infty}^{\infty} \min\left(f_{\text{intra}}(x), f_{\text{inter}}(x)\right) dx \tag{21}$$

f_{intra} is the probability density function of the intra-fHDs which contains μ_1, the mean of the intra-fHD distribution. σ_1 is the standard deviation of the intra-fHD distribution. f_{inter} is the probability density function of the inter-distribution which is based on μ_2 and σ_2. x represents the variable over which the integration is performed. The integral is taken over the entire real line $(-\infty, \infty)$, representing the overlap region of the two distributions. Smaller overlaps, and therefore smaller 'probability of cloning' statistics, indicate a robust security system where individual IDs are unlikely to be mistaken for other IDs in the database.

Metric	Simple Definition	Ideal Value
Prob. of cloning	The area of the overlap region between the two fHD distributions	Lower

10. NIST-800-22 randomness testing

The NIST-800-22 test suite, developed by the National Institute of Standards and Technology (NIST), is a well-established method for evaluating the randomness of binary sequences, primarily in the context of random number generator (RNG) testing. The suite consists of 15 statistical tests that check for predictable elements in a binary sequence to identify overrepresented patterns that might compromise randomness. The first test in the suite is the 'Frequency (Monobit) Test', which checks for the proportion of 1s and 0s in the bit-string and the deviation from the expected 50/50 split. Another example would be the 'Runs test', which counts the length and frequency of same-parity runs of bits. These are then compared to what is expected from an ideally random source. Full documentation of the suite can be found in the 2010 revision of the documentation by Bassham et al.[58] and further explanations of the tests can be found in the Appendix IX. While passing these tests suggests a high degree of randomness, it cannot guarantee unpredictability, as certain patterns may evade detection, or repeat over longer periods than is tested for. Further to this, even simulated random data can fail some NIST-800-22 tests when input lengths increase significantly[59].

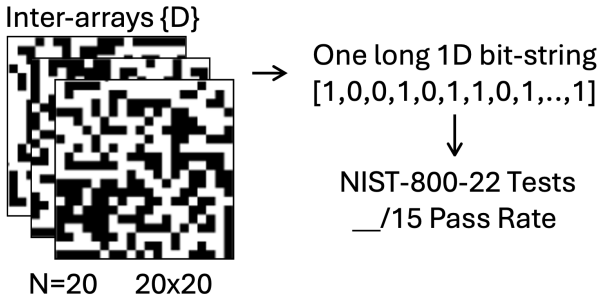


FIG. 13. The process of NIST-800-22 randomness testing. A number of inter tags (N) are imaged, then the 2D arrays are converted into a 1D bit-string. The NIST-800-22 randomness tests are then applied and their pass rate stated.

When applied to O-PUFs, NIST-800-22 is adapted to analyse the 1D bit-string generated by flattening the 2D binary arrays as shown in FIG. 13. This flattening can be performed along the rows or columns, with the resulting 1D bit-string serving as the test input. These tests provide insights into the randomness of the binary bit-string, a property distinct from the ID-centric metrics previously discussed. While randomness testing is relevant for PUFs that hope to have each response bit ideally unpredictable, this is not a requirement for anti-counterfeiting and authentication applications. As such, true mathematical randomness, as assessed by NIST-800-22, is not essential for their performance.

O-PUFs most often fail NIST-800-22 tests due to their non-IID nature, as shown in FIG. 3 and FIG. 14. This limitation reflects the challenges posed by feature sizes in binarized IDs, which affect array randomness. Similar to this test suite, as part of their series of recommendations for random number generators (NIST SP800-90 series, here 90B), NIST suggests a series of entropy (or predictability) estimation tests for non IID data, which may be more applicable. Nevertheless, approximately a quarter of O-PUF studies sampled employ NIST-800-22 testing, with varying results. Passing of these tests may be due to post-processing methods such as hashing, or where the binarization or sampling distances across an image are much larger than any local feature correlation on the physical tag. Despite its limitations, NIST-800-22 remains a useful, if imperfect, tool for exploring certain aspects of O-PUF behaviour.

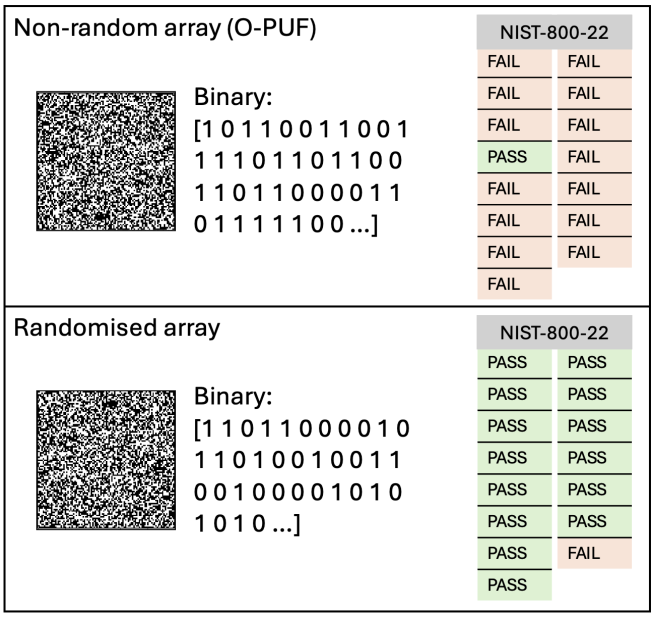


FIG. 14. A non-random array derived from an O-PUF source is shown on the left along with the binary code of the first 50 digits. The same is shown for an array produced with a random number generator. The pass/fail outcomes of the NIST-800-22 tests are listed. Note that O-PUF arrays that are not IID and therefore fail the NIST tests can still provide unique and unpredictable patterns which lead to valid anti-counterfeit properties.

Metric	Simple Definition	Ideal Value
NIST-800-22	A series of 15 tests for randomness in a 1D bit-string	15/15 Pass

11. Bit error rate

The BER quantifies the extent of deviation between captured fingerprint images and their expected array, as enrolled as the 'true' PUF response. This metric is pivotal in assessing the fidelity and consistency of O-PUFs over various evaluations. BER is computed using the fHD metric, as illustrated by Lu *et al.*[22], and is defined as:

$$\text{Bit Error Rate} = \left(\frac{1}{X}\sum_{x=1}^{X} fHD(C_t, C_x)\right) \times 100\% \qquad (22)$$

where X is the number of arrays, corresponding to the same tag, that the calculation is being performed on. fHD is the fractional hamming distance between arrays C_t, which represents the ground truth array and C_x which is the array in the set it is being compared to. C_x can either be a set of readings to a different time interval or different conditions. These arrays are in set $\{C\}$. These comparisons are completed in set of X and averaged to give BER. If the BER is the time case, reliability can be determined. Reliability is a metric used in a few papers which ties directly to bit error rate when measured over time:

$$\text{Reliability} = 1 - \text{BER} \tag{23}$$

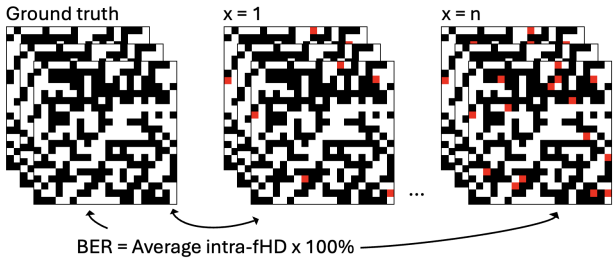


FIG. 15. Bit Error Rate. The 'errors' are shown in red and represent the deviation from the ground truth fingerprint matrix. The BER is calculated by taking the average intra-fHD between the initial image set and the final image.

Using the process illustrated in FIG. 15, the initial fingerprint is compared to subsequent readings. This enables the determination of any deterioration of the ID over time. BER is reported using various metrics, including Hamming values, percentages, the number of bits per 100 pixels, or total bit loss.

Metric	Simple Definition	Ideal Value
Bit Error Rate	The deviation between a response & accepted value	0%

12. *Entropy*

A range of entropy-measuring techniques and systems can be applied to analyse any image, including those that are produced as the 'response' of an O-PUF. Theoretically, higher entropy in images should correlate with better O-PUF performance. However, practical challenges such as noise and exposure often result in inaccurate assessments or comparisons. Despite these challenges, discussing these methods is important.

A small subset of papers use entropy metrics. The methodology for this analysis is as follows:

- **Setup**: Capture a variety of sample images, controlling variables to minimise noise such as: overexposure, underexposure, glare, and other extraneous factors.
- **Analysis**: Ensure uniform pixel size, a critical control variable, and process the images, calculating entropy from the pixel data using a chosen mathematical method. Calculate the mean entropy across the image dataset.

In the work by Cao *et al.*[30], 1D entropy, also known as 'Shannon Entropy' is defined as:

$$H(X) = -\sum_{x \in X} p_x \log p_x \tag{24}$$

where $H(X)$ represents the entropy of a discrete random variable X with a maximum state probability p_x. In the case of an IID binary array, this value is equal to the maximum of the probabilities of either the 0 or 1 state in a binary array, directly relating to bias. This connection highlights how Shannon entropy, which is derived from p_x, captures the global distribution of bit states. However, it does not account for local or dimensional variations in the array, such as alternating 0/1 rows, which can produce localised biases. Another relevant entropy metric for PUFs would be that of min entropy[60]. This value is calculated as the negative logarithm of the most probable state in a distribution, and is considered the most conservative estimate of the entropy of a system.

While not frequently referenced in O-PUF literature, a binary-specific variant of Shannon entropy exists and was first introduced in Shannon's seminal 1948 work[61]. Binary entropy is commonly used to analyse the entropy of binary bitstrings:

$$\text{binary entropy} = -[p \log_2 p + (1-p) \log_2 (1-p)] \tag{25}$$

where p represents the probability of one of the two states (0 or 1) in a binary distribution. Binary entropy provides a direct measure of randomness and bias in binary sequences in terms of bits.

While the standard form of Shannon entropy, seen in equation 24, provides a useful measure of randomness for non-binary sources, its limitations make it less comprehensive for evaluating non-IID sources. For such cases, entropy metrics like the Equivalent Number of Independent Bits (ENIB) or non-IID entropy estimates can offer more meaningful insights. These metrics aim to estimate the number of truly entropic bits per raw bit, rather than providing a simple pass-fail result as with traditional methods.

Advanced standards like NIST SP-800-90B provide entropy estimation techniques tailored to non-IID sources, enabling per-bit entropy analysis for a more nuanced understanding of randomness[62]. Examining entropy and bias in smaller subsections of the array can further reveal regional variations, offering a detailed view of randomness across dimensions.

The definition of "entropy" in metrics is broad, and different tools are suited to different purposes. ENIB is better for assessing the effective randomness of inter-fHD distributions, while NIST-800-22 provides insights into bit-string randomness. Each metric serves distinct applications, highlighting the importance of choosing appropriate tests based on the specific requirements of O-PUF evaluation. While NIST-800-22 testing is effective for evaluating randomness in IID sources, its relevance to non-IID data is limited. Incorporating non-IID entropy estimation methods, such as those in SP-800-90B, could enrich the analysis of O-PUFs.

D. Weak and strong PUFs

In the context of PUFs, the classification of weak or strong is based on the number of cryptographic keys or challenge-response pairs that can be reliably extracted from the hardware. This typically derives from the rate at which the CRP space increases with the physical size of the unique disorder of the tag, and directly affects their applicability in cryptographic security models[63]:

- Weak PUFs provide only a small number of stable responses and are often used as unique tags. Their security depends on the difficulty of reproducing the same response, but once extracted, the key is static. This is primarily used for identification or cryptographic seeds, making them suitable for applications where a static secret key is needed, such as secure storage or authentication.

- Strong PUFs have a large CRP space that allows for dynamic, on-demand cryptographic key generation. The key advantage is that individual responses are not reused, making them resistant to replay and modelling attacks. The vast number of CRPs also enables challenge-response authentication schemes, where different challenges can be used to continuously verify authenticity.

The security strength of a strong PUF is typically assessed using the intra-fHD and inter-fHD distributions of challenge responses within each PUF instance, as well as across different PUF instances as with weak PUFs earlier. While the intra-fHD metric is the same as before (repeated measurements of the same response on the same device), the definition of intra-fHD is extended to include the fractional hamming distance across the space of CRPs within each PUF, as separate to the mean bit distances for responses at the same CRP location across multiple different devices. For a PUF to be secure, the challenge responses within a PUF must be unique and unpredictable between themselves, as well as being unique and unpredictable between different PUF instances. A further method of evaluating the security of a strong PUF is to compute the Shannon entropy or min-entropy over the space of observed response distributions within each PUF instance. The total entropy of a strong PUF can be considered as the equivalent number of independent CRPs that can be generated without pattern repetition or correlation.[51].

While most optical PUFs (O-PUFs) have traditionally been weak[7,11,17,64], certain optical scattering-based designs, such as that proposed by Pappu et al.[2], demonstrate the potential for strong PUF behaviour. In such designs a large set of independent CRPs can be extracted. However, achieving practical strong PUF implementations in optical systems remains challenging due to measurement noise, environmental variations, and physical stability constraints. Despite these challenges, designing optical systems with higher rates of CRP scaling remains an active area of development. As an example, recent research has explored multi-factor O-PUFs, which typically involve 2-5 layered responses per challenge[9,16,53,65]. These designs explore a variety of methods to collect a larger set of responses for the same size of tag.

It is important to note here that since strong PUFs tend to have multiple CRPs that derive their response from the same unique physical disorder, there is a risk of higher-order predictability across the CRP set[66]. This means that there is the risk of an attacker deriving, or gaining undue insight into, the nature of a certain challenge response pair from a different subset previously captured - typically via machine learning based attacks[29,67–69]. This is tested for by evaluating against such attacks in the design stage, or computing metrics (such as Pearson's Correlation Coefficient[70,71]) for dependency between responses. However, due to the limited use of strong PUFs in the subdomain of physical authentication, this work does not seek to review these methods in detail. To be secure, the CRPs within the same PUF instance should still be maximally unique on a per-bit basis, and can be broadly considered as a set of separate tags when using the metrics discussed in this work.

E. Metrics for performance over time

Understanding how well the array ID of a specific O-PUF is maintained over time is crucial. The ID or 'digital fingerprint' of the O-PUF should remain readable over extended periods. This is particularly vital for O-PUFs generated from materials prone to decay, such as organic or fluorescent substances. Fluorescence-based O-PUFs must combat photobleaching[72]. Some other material-based O-PUFs must resist harsh environmental conditions such as high temperature and humidity - depending on their intended application[7]. This is especially true for commercial applications, where the long-term fidelity and security of products or important documents are of utmost importance. Among the studies that analysed performance over time, the following metrics were used:

- **FPR**: Following either the *experimental FPR* or the *theoretical FPR* method, which is laid out in section III C 7.

- **BER**: Testing the differences between the initial fingerprint and resulting fingerprint is important. This process is shown in FIG. 15 in section III C 11.

- **Uniformity**: As shown in section III C 3, uniformity is the 0/1 bias proportion for IDs. The average uniformity taken over time shows if the tags are maintaining high information density.

By capturing images at set intervals after sample creation, the fidelity of the O-PUF over time can be evaluated. Metrics such as BER and FPR can be derived from the same dataset, allowing for simultaneous calculations of accuracy, recall, and precision. Uniformity can also be tracked over time by examining changes in the inter-fHD images.

In fact, all metrics can be adapted for time-dependent analysis with the appropriate data collection. A comprehensive evaluation of an O-PUF's performance over time would not

be unfeasible, transforming static tests into dynamic ones to assess long-term reliability.

IV. COMPREHENSIVE EVALUATION AND RECOMMENDATIONS

This section addresses a number of key considerations for evaluating O-PUFs. We discuss the limitations of mean-based metrics, the impact of curve fits with low (N)choose(2) values, and the effects of array size. We also advocate for using decidability and the probability of cloning as crucial figures of merit, examine the reliance on Normal fits for theoretical FPR, and suggest metrics for ensuring accurate time-dependent testing. Additionally, we highlight important considerations for NIST-800-22 testing and recommend an effective testing suite.

A. The problems with only using mean-based metrics for O-PUFs

A significant proportion of papers use only the intra-fHD means (related to reliability) and inter-fHD means (related to uniqueness) as their figures of merit. When the intra-fHD and inter-fHD curves are plotted, an overlap between the two distributions indicates a difficulty to distinguish between different PUF instances. However, without access to the fHD plots, the reader of these papers cannot determine whether the O-PUF is genuinely reliable or unique. This is because the standard deviations of the fits determine the overlap in fHD between the instances of repeat measurements of the O-PUF ID and measurements of different IDs. As detailed in section III A, greater variability in them standard deviations may be present as the data does not follow the binomial dependence between the μ, n and σ.

FIG. 16 highlights the limitations of relying solely on mean-based summary statistics. In the first scenario, with $\mu_1 = 0.1$ and $\mu_2 = 0.5$, the reliability (90%) and uniqueness (50%) metrics appear favourable. However, these metrics can be misleading, as the O-PUF may perform poorly in other critical statistics such as False Positive Rate or ENIB/DoF. The standard deviations of the Normal distributions are crucial for accurately assessing the quality of the O-PUF. An almost ideal PUF is depicted in the second scenario, where all statistical measures are consistent. This oversight is understandable, as these tests were designed to compare PUFs that follow IID, unlike some O-PUFs. As illustrated in FIG. 3, for other PUF types, the standard deviation remains constant as $\sigma(\mu_1, n)$. However, in the case of O-PUFs, the standard deviation can vary significantly, potentially leading to substantial overlap and rendering the O-PUF ineffective.

By failing to account for these variations, researchers may overestimate the performance of O-PUFs. Thus, a more rigorous evaluation framework, which includes metrics beyond mean values, is essential for accurately determining the efficacy of O-PUFs.

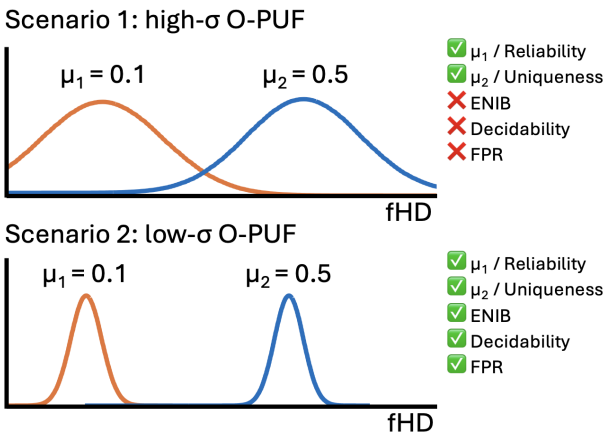


FIG. 16. An example scenario illustrating a curve fit for a lower performing, high σ O-PUF in scenario 1 as contrasted with a better performing, low σ, O-PUF in the second scenario. Due to the identical means of both distributions, mean-based metrics alone erroneously classify both as high-quality IDs. However, the introduction of metrics that consider both mean and standard deviation, such as ENIB, accurately reflects the true nature of the O-PUFs. Standard deviations can vary more as this data does not follow IID.

Key Findings	Using only μ-based metrics for non-IID O-PUFs is insufficient. Include standard-deviation-based metrics or state standard deviation (σ).

B. Curve fits with low quantities of data

The accuracy of metrics derived from Normal fits of fHD depends heavily on the robustness of the fits, which in turn relies on well-determined goodness-of-fit parameters. Achieving this robustness requires a sufficiently large sample size. Since fHD values are computed for all possible pairs of arrays, the number of data points scales combinatorially with the sample size, following the binomial coefficient $\binom{N}{2}$. This scaling provides ample data for reliable curve fitting while keeping the experimental workload manageable.

As illustrated in TABLE II, the number of samples (N) utilised in different experiments varies significantly. Although there is no definitive cut-off for an insufficient number of data points, a larger N yields a more accurately fitted curve. The quality of this fit may be assessed through the use of 'goodness-of-fit parameters'. Notably, deviations in the values of standard deviation σ from theoretical IID predictions may indicate interdependence within the data.

The value of N is often constrained by the practicalities of O-PUF manufacturing. In simulation studies, N can be exceptionally high, providing a more robust dataset for analysis. Conversely, in experimental scenarios where N is low, researchers have sometimes mitigated this limitation by dividing tags into multiple CRPs[21]. Thus, while larger datasets are preferable, smaller datasets can still provide valuable insights, albeit with increased consideration for the fit quality and potential inter-dependencies.

A Chi-squared (χ^2) test can be used to evaluate the quality of the Normal fit by comparing the observed frequency distribution of the fHD values to the expected frequency distribution derived from the fitted Normal curve. The test statistic is given by:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{26}$$

where O_i is the observed frequency and E_i is the expected frequency for the i-th bin. A smaller χ^2 value indicates a closer match between the observed data and the expected distribution, suggesting a robust fit. The p-value associated with the χ^2 statistic can further indicate whether deviations from the fit are statistically significant.

In addition to the Chi-squared test, other goodness-of-fit metrics such as Root Mean Square Error (RMSE), adjusted R^2, and the Kolmogorov-Smirnov (K-S) test can provide valuable insights. RMSE measures the average magnitude of residuals, adjusted R^2 evaluates the proportion of variance explained while accounting for model complexity, and the K-S test compares the observed cumulative distribution function (CDF) to the theoretical CDF. Employing multiple tests can be a method of ensuring a comprehensive evaluation of the Normal fit, which is particularly important for datasets exhibiting non-IID behaviour or limited sample sizes, thereby enhancing the reliability of derived metrics.

Key Finding	N must be sufficient for a good Normal fit, increased N is generally better.

C. The effects of array size

The size of digital keys or IDs generated by O-PUFs varies significantly across studies. Smaller arrays, when derived from the same initial imaged tags and feature sizes, may reduce noise by averaging over fewer pixels per array element, while larger arrays capture finer details and allow for more dataset variability. In arrays that are truly random, the inter-fHD should tend towards the 0.5 mark. For IID PUFs, the fHD distributions should follow the binomial distribution which means that the array size should affect the standard deviations of the inter-fHD distribution with the relationship $\sigma = \sqrt{n\mu}$. As shown in FIG. 17, the randomised IID data generated, increasing array size improves the distribution of fHD for the array. This observation aligns with the 'Law of Large Numbers'[73], which states that as sample size increases, averages converge closer to their true values.

However, when conducting this analysis with a non-IID dataset, the patterns observed in the samples used are not completely random. A small-scale pilot study was conducted on a set of $N = 20$ O-PUF tags, where the samples consist of a mixture of fluorescent dyes and were imaged using a high-DPI image scanner. The results of this study are illustrated in FIG. 18.

For each set, the images were binarized using modified version R-LBP (Reduced Local Binary Patterns) with radius = 3

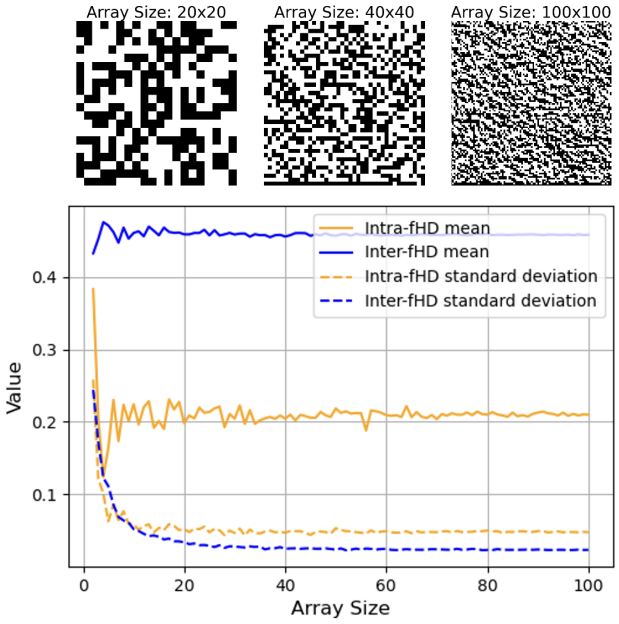


FIG. 17. Simulated data: showing the IID arrays follows the binomial σ relationship as the array size increases. While this does not affect the mean of the distribution, it does effect the standard deviation. This would give higher μ&σ-based metrics for higher array sizes.

and neighbourhood = 16. The only variation across tests was in the 'keysize', which determines the size of the output array. This was calculated for each fingerprint size from 2 to 100, and intra-fHD and inter-fHD plots were generated for each set. The means and standard deviations of these distributions were then stored and plotted in the final graph depicted in FIG. 18.

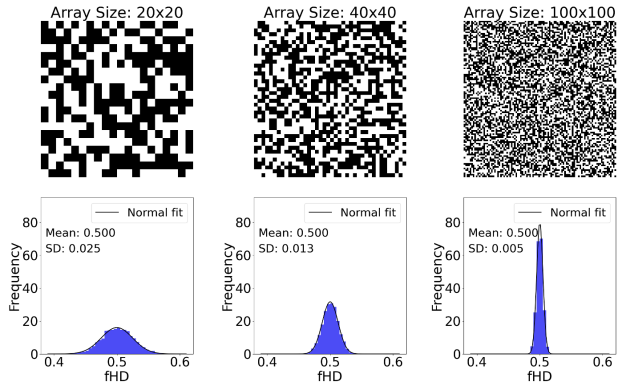


FIG. 18. Non-IID arrays: Variation in array size can be used to check for consensus in the means and standard deviations for the intra-fHD and inter-fHD metrics. It can be seen that in this case, array sizes below 20x20 do not capture the true nature of the O-PUF. This was tested on a N=20 inter-images and intra-images set for fluorescent molecule tags. These arrays were binarized using R-MLBP with different key sizes giving different array sizes.

FIG. 17 demonstrates that the standard deviations and

means stabilise around an array size of $n = 20$. However, this stabilisation is specific to the feature size of the test samples. The choice of key size or array size is critical for producing a fingerprint that effectively conveys information while minimising noise. As n (array size) increases, higher resolution increases the likelihood of capturing physical interdependencies among features or patterns.

In conclusion, if an array is IID (passes NIST-800-22), the array size will influence the σ value. As a result, μ- and σ-based metrics are not suitable for comparison unless the array size is explicitly stated, allowing other studies to use the same dimensions for consistency. However, for non-IID arrays (those that fail NIST-800-22), the standard deviation is not impacted by array size. This is due to the breakdown of statistical relationships for non-IID described in the section III A. As discussed in section IV A, metrics must incorporate σ to ensure valid comparisons in such cases.

Key Findings	If O-PUFs do not follow IID, array size doesn't affect standard deviation (σ). μ & σ-based metrics are a useful tool for analysis that are not affected by array size.

D. A case for the use of decidability or probability of cloning as key metrics.

Decidability, which is increasingly used in various bioinformatics fields[57], serves as a robust summary statistic for comparing the means and standard deviations of the two fHD fits for the O-PUFs. Although none of the reviewed papers have utilised this metric, this review advocates for its inclusion in future research efforts. Decidability is maximised when the intra-fHD and inter-fHD curves are well-separated with minimal standard deviations, indicating a higher ease of distinguishing an ID from other samples.

Another significant metric that holds potential as a commercial benchmark is the probability of cloning, discussed in Section III C 9. This metric quantifies the overlap between two fitted curves, indicating the material's effectiveness as an O-PUF in a generalised context. Decidability is a comprehensive metric providing a concise summary statistic for comparing materials, encapsulating both the reliability (intra-fHD mean) and uniqueness (inter-fHD mean) of O-PUFs and their level of security.

E. Relying on Normal distributions for theoretical FPR

When it comes to the difference in FPR calculation between studies, O-PUFs exemplify their interdisciplinary reach. While more biochemical studies often list experimental FPR, more cryptographic studies use theoretical FPR. While experimental rates rarely go below values such as 0/200 as this would require 200 tests, the theoretical rates can be as low as 10^{-300} as they are derived from calculations based sometimes on the tail-ends of gaussian/normal fHD distributions. Despite the fact that theoretical FPR can give a good security measurement that can be compared to other metrics for security and can extrapolate the potential of the system FPR outside of the experimental testing range, this can yield misleading results when considering the materials practically. For context, here is a table illustrating the estimated time required for experimental verification of different *Theoretical FPR* values, assuming one testing image can be taken per second and that the calculations are conducted to yield a True/False result.

TABLE VII. Comparison of theoretical FPR with the estimated time required to experimentally conclude these rates, expressed in billions of years. The table demonstrates that for FPR values below 10^{-2}, experimental verification becomes impractical within a reasonable time-frame. The time estimates are based on imaging one sample per second.

Theoretical FPR	Estimated time for same conclusion via Experimental FPR (billion years)
10^{-2}	3.17×10^{-15}
10^{-10}	3.17×10^{-7}
10^{-50}	3.17×10^{33}
10^{-100}	3.17×10^{83}
10^{-300}	3.17×10^{283}

TABLE VII presents the estimated time required to experimentally validate theoretical FPR. Testing FPR values below 10^{-2} becomes impractical, as verifying an FPR of 1/100 requires testing 100 tags, taking approximately 100 seconds. Given variations in imaging techniques, testing FPRs below 10^{-7} would take over a week, making them increasingly impractical. This challenge is amplified in optically-read PUFs due to time constraints imposed by the material system, though it is more feasible with E-PUFs.

Both theoretical and experimental FPRs require the same data collection process, involving sets of intra-images tested against fHD distributions. While theoretical FPRs yield more impressive results, experimental data provides practical validation. Presenting both together ensures transparency in data processing. Additionally, stating the total number of images tested is crucial, as experimental results like 0/200 images support a theoretical FPR of 1×10^{-10} more effectively than theoretical values alone. Conversely, a 1/200 experimental FPR indicates possible system noise, questioning the validity of extremely low theoretical FPRs.

Key Findings	When *theoretical FPR* is calculated, *experimental FPR* should also be calculated. Extremely low *theoretical FPR* should be considered as a hypothetical lowest bound.

F. Choosing figures of merit for longevity

When selecting FoM for determining the longevity of an O-PUF, researchers have employed various approaches. Metrics such as BER, FPR, and uniformity have been adapted to analyse how well an O-PUF maintains its fingerprint over time.

These metrics, already used in time-of-manufacture analysis, can be extended for time-dependent evaluations, as discussed in section V. This consistency in methodology ensures more reliable and comparable results across different studies.

Uniformity is widely used in the literature to assess the bias of an O-PUF's fingerprint, providing insight into the balance of 0s and 1s and, consequently, the information density. However, uniformity does not measure changes to the fingerprint ID over time. While it evaluates the presence of information, it does not determine whether that information is preserved, as metrics like BER or other ID-tracking methods do. For this reason, uniformity should not be used in isolation but rather as part of a broader suite of metrics that collectively assess both the quantity and stability of the information over time.

In addition to traditional metrics, the use of distribution evolution provides a powerful tool for visualising changes in the O-PUF's performance over time. By analysing the evolution of key distributions at regular intervals, researchers gain an intuitive understanding of how the fingerprint's uniqueness and reliability evolve. This method complements quantitative metrics by offering a direct, visual assessment of temporal performance, which is particularly valuable for identifying trends or anomalies in O-PUF behaviour.

Overall, combining these established and emerging methods provides a comprehensive framework for evaluating the long-term reliability of O-PUFs, ensuring their suitability for applications requiring sustained performance and security. This method is provided in the python package in later sections alongside more conventional metrics.

Key Findings	BER paired with FPR provides a useful metric for fingerprint longevity. For a fully comprehensive analysis, this research proposes 'distribution evolution' testing.

G. Important notes on NIST testing

NIST-800-22 testing, which is used to evaluate the randomness of cryptographic sequences, is one of the tests used in the literature. Different NIST-800-22 tests require varying minimum input bit-string lengths, derived from flattening 2D interfHD arrays into a 1D bit-string. While typically applied to help prove the unpredictability of each bit of a PUF by themselves, understanding these requirements is important for the subject of this review. Achieving the necessary 1 million bits can be challenging, but it is highly recommended for this type of testing.

As shown in FIG. 19, it's possible to get to 1000000 bits if a fingerprint of 100x100 is generated and 100 inter-images are taken. If the array size is smaller, more inter-images are needed - however with appropriate experiment planning, it is possible. Additionally, it is worth noting that Section 4.2.2 of the NIST-800-22 documentation mentions and recommends that, in order to obtain statistically significant results, the tests - each requiring a 10^6-bit input - should be repeated 55 times, with a p-value calculated from these repetitions. However, this is unfeasible for most O-PUFs, as it would require 5.5×10^7 bits. Therefore, this recommendation should be considered based on the specific research scenario.

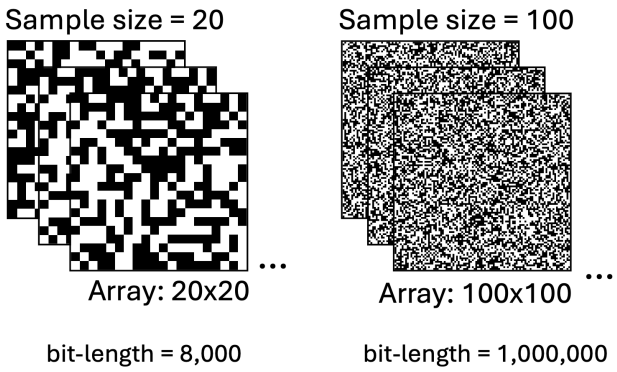


FIG. 19. A depiction of varying array setups illustrating the capability to achieve bit-string lengths (n) suitable for NIST-800-22 tests. The figure showcases two distinct fingerprint configurations: one with bits=n=72000, another with n=1000000. These data sets demonstrate the feasibility of generating bit-string lengths of 1000000 bits, a prerequisite for certain NIST-800-22 evaluations.

A note on 'Entropy': When considering 1D entropy metrics, such as Shannon entropy (as discussed in section III C 12), and the NIST-800-22 test suite, it's important to note the distinctions between these approaches. Both assess aspects of randomness, but they do so differently. The NIST-800-22 suite comprises 15 tests that evaluate a sequence's unpredictability, detecting certain predictable patterns across a range of criteria. In contrast, Shannon entropy provides an actual continuous entropy rate estimate, assuming the data is IID, which theoretically offers direct insight into the randomness level of the data itself. Importantly, the NIST tests are also limited to a 1D framework, as is the Shannon entropy calculation when applied to binary bit-strings. Neither approach captures the full spatial relationship between bits that might be better represented through 2D approaches to entropy metrics, which, as of yet, have not been explored in the literature.

A note on hashing: When applying hashing or seeded extractors to improve the randomness of O-PUF responses, it is crucial to consider their potential to mask underlying issues in the data. Hashing can spread local patterns, such as bias or correlation, across a wider scale, making the bit-string appear IID, which can result in the incorrect application of metrics. For instance, a biased or highly correlated input may yield a seemingly uniform hash output, but the underlying predictability remains.

While some PUFs are suitable for random key generation, O-PUFs are not. NIST-800-22 testing evaluates randomness in bit-strings, but non-IID O-PUFs fail the tests[58], as patterns emerge as the array size increases relative to feature size. FIG. 14 illustrates these limitations, discussed further in section III A.

Key Findings	NIST-800-22 testing recommends at least 1'000'00 bits for the full 15 tests. NIST-800-22 will fail on O-PUFs which are not IID.

H. Recommended metrics for a testing suite.

This section provides a comprehensive set of recommended metrics for evaluating O-PUFs, aimed at standardising the analysis and facilitating comparison across different studies. By adhering to these guidelines, researchers can ensure their work is both rigorous and comparable to other investigations in the field.

TABLE VIII. Summary of metrics and recommended components of the testing plan, including secondary variables necessary for comparability with other studies. This testing plan is implemented in the Python package.

Metric	**Figure of merit/metric**	**Secondary Variables**
Intra-hamming	μ_1 and σ_1 stated, Reliability (%)	N_1, Array size
Inter-hamming	μ_2 and σ_2 stated, Uniqueness (%), ENIB	N_2, Array size
Whole hamming Plot	Shown plot, Decidability	
Experimental & Theoretical Metrics	Experimental & Theoretical FPR, TPR, FNR, TNR	n, decision point
Randomness	NIST tests, Average uniformity (inters)	N_1,N_2 NIST variables
Over-time metrics	FPR, BER (or uniformity) and full Gaussian evolution	N_1, N_2, Array size

TABLE VIII shows the main recommended figures of merit for future O-PUF research, along with any secondary variables that should also be listed.

Secondary Variables:

- Array Size**:** The nxn size of the binarized fingerprint produced by the analysis.
- N_1**:** the number of repeat images inputted for the intra-fHDs calculations.
- N_2**:** the number of different images inputted for the inter-fHDs calculations.
- n**:** the number of testing instances (intra-images/inter-images) for experimental and theoretical FPR+ testing.
- Decision point**:** (also known as threshold value) the specific value at which an experimental test distinguishes between a True and False outcome.
- NIST variables**:** the NIST-800-22 documentation can be consulted for the specifics of variable input needed to run the 15 tests.

Using this suite of tests, which includes means and standard deviations from Normal fits, ensures comprehensive analysis and facilitates direct comparison across studies. It accounts for variables like array size and input sample count, aiding interpretation of anomalies in the data. NIST-800-22 testing is included to help inform about the IID or non-IID nature of the O-PUF and therefore can be used to inform the tests the researcher chooses to focus on. Additionally, by moving beyond mean-based metrics, this approach addresses the limitations of non-IID O-PUFs, promoting standardisation and comparability, even for those not employing the Python package discussed in the next section.

V. ACCOMPANYING OPEN-SOURCE PYTHON PACKAGE

The Python package described in this report enables calculation of all the main figures of merit, facilitating comparability across studies and streamlining future research efforts. Freely accessible on `GitHub`, it will be installable using '`pip install pyopticalpuf`' and requires minimal Python knowledge for use on a wide variety of imaged samples. Comprehensive documentation and `.ipynb` Jupyter notebooks are also included on the Github page, providing step-by-step guides to installation, function usage, and producing final displays. An example testing set is also provided to verify the code.

The package supports pre-processing steps like cropping and binarization, offering four main algorithms: Reduced Local Binary Pairs (R-LBP), Adaptive High Boost (AHB), Sauvola, and Otsu's method. Each algorithm includes optimisable parameters to suit specific patterns in the imaged O-PUF. Custom binarization methods can also be added, offering flexibility and adaptability for diverse applications. Flowcharts describing the processing present in both forms of testing can be found in FIG. 20 and FIG. 21. The outputs themselves, in the form of FIG. 22 and FIG. 23 can be found on the subsequent pages.

A. Time-of-manufacture testing

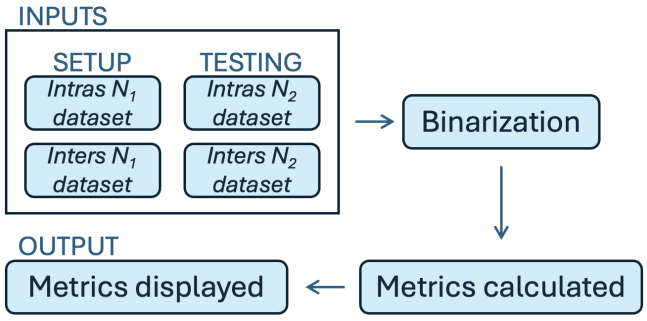


FIG. 20. A flowchart illustrating the inputs, process, and output of the code. The process begins with a set of N intra-images and inter-images, divided into setup and testing sections. These images are then binarized using either a pre-defined algorithm from the package or a custom-built one. All relevant metrics are subsequently calculated and displayed in the output report .png file.

This section provides an overview of the comprehensive testing capabilities enabled by the Python package `pyopticalpuf`. By running the code provided in the repository, all major figures of merit and supporting metrics for O-PUF analysis can be calculated and visualised. The output includes key metrics, fHD histograms, confusion matrices, and NIST-800-22 results, ensuring reproducibility and facilitating comparisons across studies. FIG. 22 demonstrates an example of the output in detail while FIG. 20 shows a generalised flow chart of the process.

As shown in FIG. 20, the code takes sets of images from the user and produces a full output report, as shown in FIG. 22. The code binarizes the images and then calculates fHDs, plotting the fHDs and calculating the relevant metrics. The IDs that are produced are also analysed using non-fHD-based metrics such as NIST-800-22 and uniformity. By consolidating all the definitions and providing easy-to-access code, the package offers a valuable tool for researchers to standardise their analysis, ensuring consistency and comparability across studies.

B. Time-dependent testing

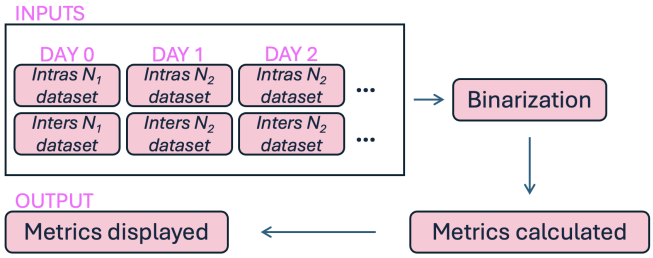
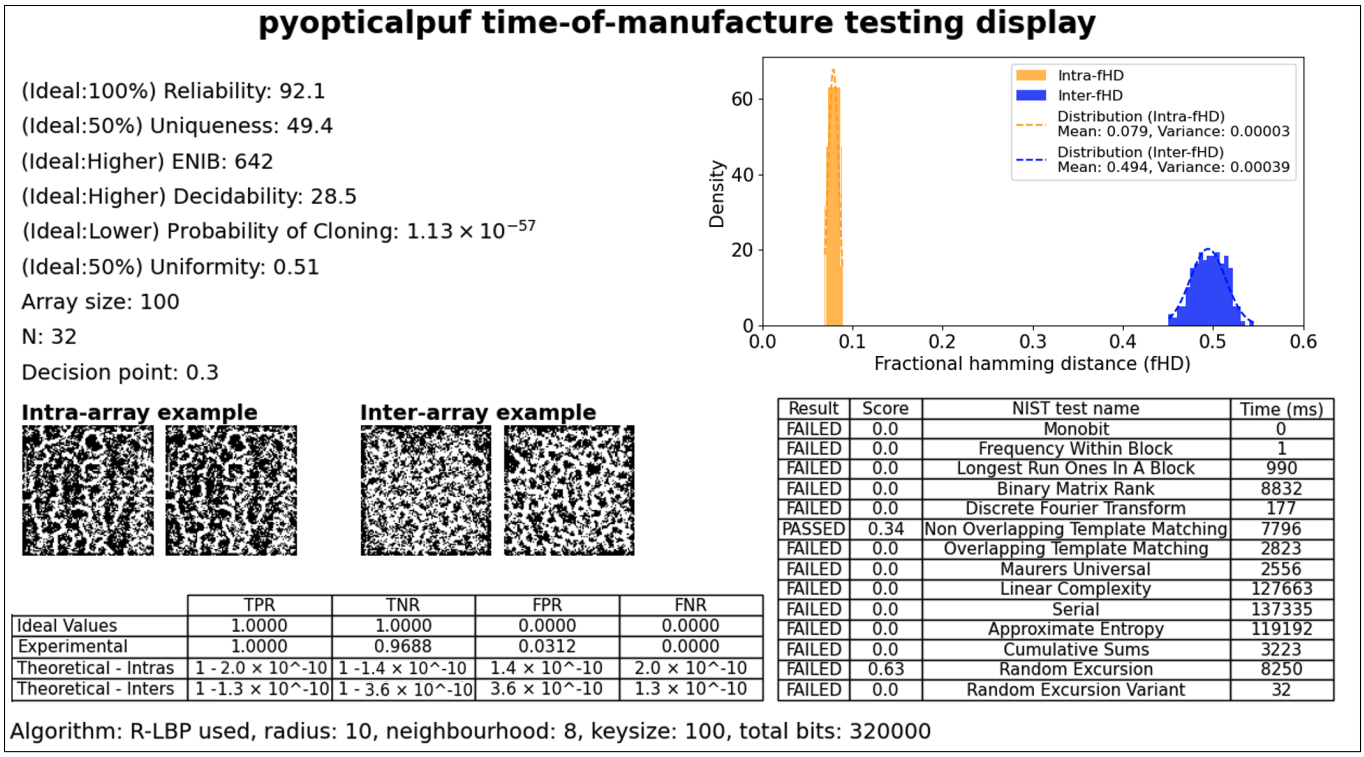


FIG. 21. Flowchart demonstrating the inputs, output, and process of testing O-PUF samples over time. Each day, a set of intra-images and a set of inter-images are taken. These images are binarized using either an algorithm from the package or a custom choice. The fHD comparisons between the sets are then used for Normal fits and metric calculations.

This Python package introduces a novel approach to O-PUF analysis over time. Beyond standard BER and FPR, it enables over-time fHD distribution analysis. 'Distribution evolution' examines intra-fHD and inter-fHD distributions and their key metrics, including reliability, uniqueness, ENIB, decidability, and cloning probability. This complements standard FPR and BER tests, where a database image from day 0 is compared to later readings. However, distribution evolution offers deeper insights into O-PUF behaviour over time. FIG. 21 defines the inputs and outputs of time-dependent testing, while FIG. 23 shows the display output, generated via 'over-time-testing.ipynb' in the GitHub repository.



	TPR	TNR	FPR	FNR
Ideal Values	1.0000	1.0000	0.0000	0.0000
Experimental	1.0000	0.9688	0.0312	0.0000
Theoretical - Intras	1 - 2.0 × 10^-10	1 -1.4 × 10^-10	1.4 × 10^-10	2.0 × 10^-10
Theoretical - Inters	1 -1.3 × 10^-10	1 - 3.6 × 10^-10	3.6 × 10^-10	1.3 × 10^-10

Result	Score	NIST test name	Time (ms)
FAILED	0.0	Monobit	0
FAILED	0.0	Frequency Within Block	1
FAILED	0.0	Longest Run Ones In A Block	990
FAILED	0.0	Binary Matrix Rank	8832
FAILED	0.0	Discrete Fourier Transform	177
PASSED	0.34	Non Overlapping Template Matching	7796
FAILED	0.0	Overlapping Template Matching	2823
FAILED	0.0	Maurers Universal	2556
FAILED	0.0	Linear Complexity	127663
FAILED	0.0	Serial	137335
FAILED	0.0	Approximate Entropy	119192
FAILED	0.0	Cumulative Sums	3223
FAILED	0.63	Random Excursion	8250
FAILED	0.0	Random Excursion Variant	32

FIG. 22. **Time-of-manufacture testing output of pyopticalpuf**. By following the code in the file 'Instantaneous-testing.ipynb' all major figures of merit for the O-PUF can be calculated. **(top-left)** The main $\mu - based$ and $\mu\&\sigma - based$ metrics are listed including: Reliability, Uniqueness, ENIB, Decidability, Probability of Cloning and Uniformity. Additional important secondary variables are also listed including N (the number of total images) and array size. Additionally, the threshold point, which is the point chosen for True/False determinations in experimental FPR calculations. **(top-right)** The histogram of the fHDs is plotted and the Normal fit applied. The intra-fHDs are seen on the left and the inter-fHDs on the right. **(bottom-right)** NIST-800-22 testing results are shown with all the tests having their result, score, name and time-taken to complete. **(bottom-left)** True Positive Rates (TPR), FPR, FNR and TPR of both the experimental and theoretical variety are shown here. Above this the inter-example and intra-examples allow a look at how the binarization affected the image and a check for the researcher to verify no essential detail has been lost. Below the FPR+ table, also known as a confusion matrix, the algorithm used is specified along with the input parameters. **(whole)** This is an example output in .png form. This allows for comprehensive testing, backwards-compatible testing and useful data for comparison studies of future O-PUFs. Any results can also be displayed separately as well as in the final output.

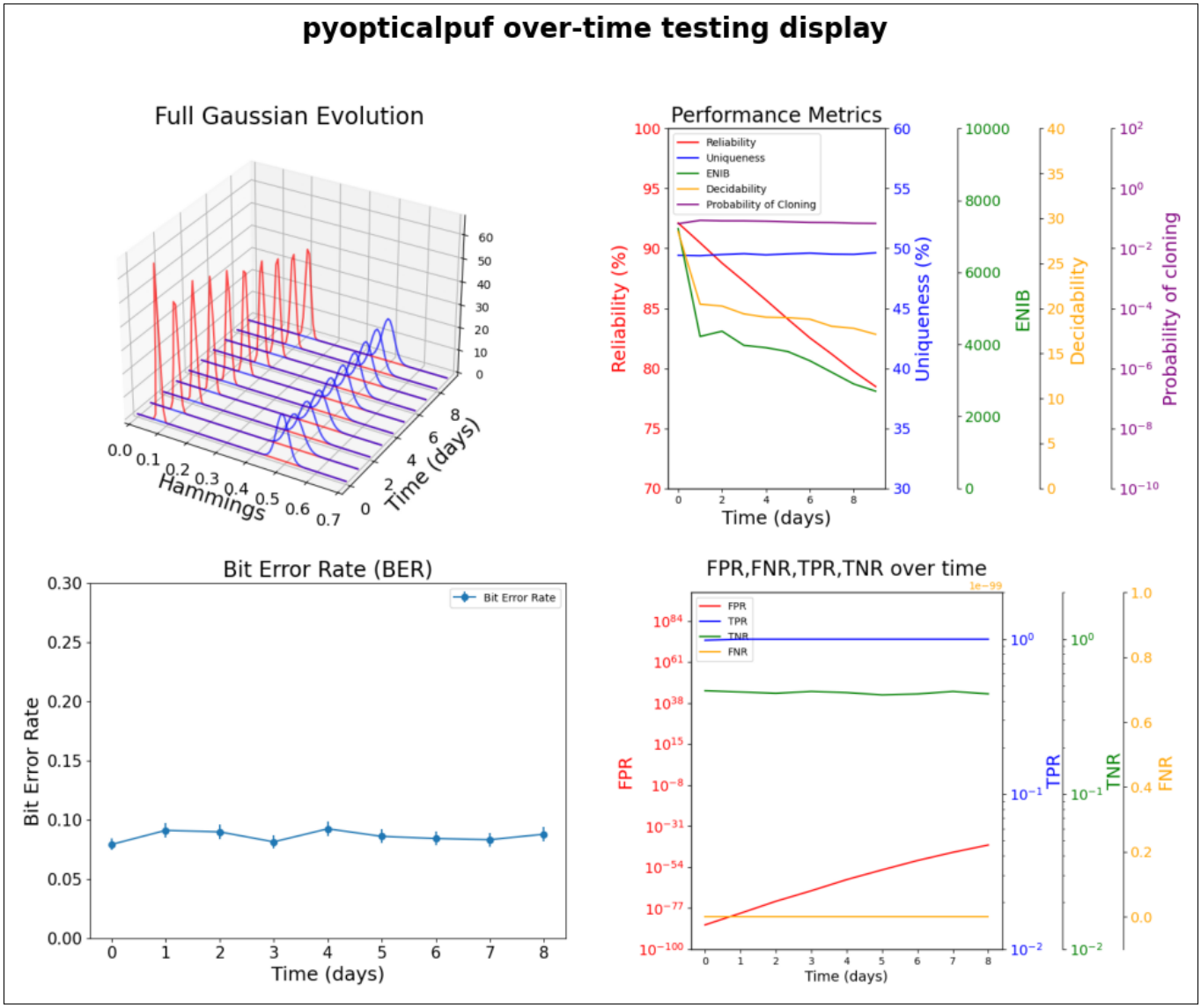


FIG. 23. The **output of pyopticalpuf** for time-dependent testing. **(top-left)** The Normal distribution plots over time illustrate the variability of individual sets of fHD measurements taken across different days. Changes in intra-fHD and inter-fHD are observed as the response of the O-PUF evolves over time. Ideally, the fitted curves show little variation, indicating stable ID. **(top-right)** Main metrics derived from the curves are tracked over time, including Reliability, Uniqueness, ENIB, and Decidability. Higher values on the y-axis indicate better performance for all four metrics, enabling a conclusive evaluation of stability. **(bottom-left)** Bit Error Rate is calculated by comparing each set of IDs to the original set from day 0. **(bottom-right)** The False Positive Rate (FPR), here theoretical, is calculated by comparing fHD from later days to the Normal-fitted histogram plot from day 0. **(whole)** This display encompasses the main tests used in the literature, augmented by the novel Normal evolution plot.

VI. CONCLUSION

O-PUFs represent a significant area in the field of anti-counterfeiting, offering a low-cost, easily mass-producible method of authentication that could revolutionise security for products and documents. As a subset of Physically Unclonable Functions, O-PUFs provide a unique fingerprint ID through the characterisation of a physical authentication material. However, unlike other PUFs, O-PUFs do not always adhere to the IID assumption, as demonstrated in section III A, which necessitates the use of modified analytical methods. As O-PUF technology transitions from research to commercial application, it is imperative to establish a standardised method for its analysis to support consistent and reliable implementation.

O-PUF analysis has employed a diverse array of metrics over the past decade. In this study, a set of 30 papers on optical physically unclonable functions from the past decade was reviewed. This broad range of metrics includes: fHD means (reliability and uniqueness), uniformity, NIST-800-22, FPR, entropy, and BER. Each metric has been evaluated in this review, and the mathematical definitions are explicitly defined in section III C.

Key findings include:

- **O-PUF IID behaviour** cannot be assumed as it typically is for electronic PUFs. This is due to the 2D nature of the input array and the presence of multi-pixel visible features in the fingerprint, meaning the bit-string produced by O-PUFs may not follow IID. This informs later decisions as this results in the Normal fits of the intra-fHD and inter-fHD data not following the binomial distribution shape where the standard deviation can be predicted mathematically. This results in more need for distribution evolution analysis and standard deviation based metrics for analysis.

- **NIST-800-22 testing** can be used to detect if the O-PUF that is being evaluated is IID. An O-PUF that fails a majority of the NIST-800-22 tests is most likely non-IID and therefore should rely on σ-related metrics instead of only μ-based metrics. However, if the O-PUF is IID, then σ-related metrics must be used with caution as σ is affected by the array size chosen.

- **Mean-based metrics** were used as the **only** metrics used in a large proportion of studies. This allows for easy comparison but makes it difficult to assess the effectiveness of the O-PUF material if IID. Mean-based metrics (such as the means, reliability and uniqueness) should ideally be supported with other fHD-based metrics such as standard deviations, ENIB, Decidability, Probability of cloning and FPR, which involve the standard deviation or have clear proof that the data is IID.

- An **Open-source, easy-to-use Python testing package** has been coded and is described in section IV H, which includes all relevant figures of merit that may be needed to compare the work to others. Frameworks for testing over time are also provided, including BER, FPR and Normal distribution evolution.

- **Secondary variables** should be stated clearly, such as the number of images binarized and inputted into the fHD plot, denoted N. Additionally, one should include the pixel size of the array - as these can affect the key figures of merit if the images follow IID. N can affect the fidelity of the fit and the array size affects the standard deviation of the inter-fHD curves.

- **Distribution evolution over time** is a novel method for O-PUF analysis over time. By taking a full set of readings and constructing Normal distributions for the intra-fHDs and inter-fHDs at each time step, the full properties of the O-PUF can be tracked over time. This will allow for more thorough comparisons and evaluation in addition to FPR and BER - which are used more generally in the literature.

In summary, this study introduces primary testing criteria that, combined with the mathematical definitions in Section III C, enable comprehensive O-PUF analysis across both instantaneous and longitudinal assessments. By encompassing key metrics from the past decade of literature, this framework ensures compatibility with prior studies while improving clarity through secondary variables like N and array size. A Python package, with additional GitHub jupyter notebook walk-throughs, consolidates these definitions to streamline analysis and enhance usability.

Looking into the future of the field, as O-PUFs continue to cross from the realms of research into the commercial space, the testing procedure requires a clear outline to avoid inefficient use of time and resources on materials which cannot be transformed into viable security products. Additionally, a clear comparison architecture could allow for the optimisation of binarization algorithms - which is another key area of advancement in the field. A clear testing suite may assist new researchers whose existing materials may have O-PUF potential to avoid pitfalls in the analysis. In future works, a database or comparison of various O-PUF technologies could be developed and standards for acceptability established. These standards would expand on ideal values for specific figures of merit and would set lower limits for what these parameters should be to make a good O-PUF.

Whilst this study endeavoured to find a majority of O-PUF publications released in the 2013-2024 range, it is likely some papers may not have been found. It should still be a representative sample of the field. The definition of an O-PUF can be broad but in this study any system that used optical imaging techniques to get the data were considered, with some variation. Also included were one or two review papers on specific subsets of the field, purely to evaluate the metrics in their contents. It will be interesting to see another review assessing any progress in the use of metrics in this field in the coming decades.

A key area for future work would be the development of a standardised framework for evaluating O-PUF quality. Es-

tablishing clear thresholds and benchmarks for what constitutes a "good" O-PUF would enable more consistent comparisons across studies. A thorough investigation into this, supported by numerical results from the literature, would help define reliable performance criteria. Additionally, a future review assessing the evolution of metric usage in the field over the coming decades would provide valuable insights into advancements in O-PUF evaluation and standardisation.

In conclusion, the recommendations for testing procedures and the code attached to this research should foster greater standardisation in the field as it continues to mature. As O-PUFs move toward commercial applications, this report aims to serve as a centralised resource for the mathematical and computational definitions of key figures of merit in O-PUF analysis, supported by the accompanying Python package. Additionally, this study underscores the critical importance of testing for IID within datasets during metrics evaluation, ensuring the reliability and validity of comparisons across studies. By adopting these practices, future research can achieve greater comparability and effectively benchmark against past O-PUF studies.

Data availability statement: *The data that support the findings of this study are available from the corresponding author upon reasonable request.*

VII. BIBLIOGRAPHY

[1] K. Jacobs and K. Jacobs, "Independent identically distributed (iid) random variables," Discrete Stochastics , 65–101 (1992).

[2] R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," Science **297**, 2026–2030 (2002).

[3] M. S. Kim, G. J. Lee, J. W. Leem, S. Choi, Y. L. Kim, and Y. M. Song, "Revisiting silk: a lens-free optical physical unclonable function," Nature Communications **13**, 247 (2022).

[4] T. McGrath, I. E. Bagci, Z. M. Wang, U. Roedig, and R. J. Young, "A puf taxonomy," Applied physics reviews **6** (2019).

[5] A. Maiti, V. Gunreddy, and P. Schaumont, "A systematic method to evaluate and compare the performance of physical unclonable functions," Embedded systems design with FPGAs , 245–267 (2013).

[6] R. Helinski, "Evaluating physical unclonable functions." Tech. Rep. (Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2020).

[7] C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, "Physical unclonable functions and applications: A tutorial," Proceedings of the IEEE **102**, 1126–1141 (2014).

[8] C. Mesaritakis, M. Akriotou, A. Kapsalis, E. Grivas, C. Chaintoutis, T. Nikas, and D. Syvridis, "Physical unclonable function based on a multimode optical waveguide," Scientific reports **8**, 9653 (2018).

[9] J. Shin, R. Ko, T. Park, Y. J. Kim, B. C. Jang, and H. Yoo, "Multidimensional physically unclonable functions: Optoelectronic variation-induced multi-key generation from small molecule pn heterostructures," Advanced Functional Materials , 2314949 (2024).

[10] A. Anastasiou, E. I. Zacharaki, A. Tsakas, K. Moustakas, and D. Alexandropoulos, "Laser fabrication and evaluation of holographic intrinsic physical unclonable functions," Scientific Reports **12**, 2891 (2022).

[11] S. M. Park, G. Park, and D. K. Yoon, "Paintable physical unclonable functions using dna," Advanced Materials **35**, 2302135 (2023).

[12] C. Mesaritakis, M. Akriotou, and D. Syvridis, "Laser induced speckle as a foundation for physical security and optical computing," in *2018 Photonics in Switching and Computing (PSC)* (IEEE, 2018) pp. 1–3.

[13] Y. Cao, S. S. Zalivaka, L. Zhang, C.-H. Chang, and S. Chen, "Cmos image sensor based physical unclonable function for smart phone security applications," in *2014 International Symposium on Integrated Circuits (ISIC)* (IEEE, 2014) pp. 392–395.

[14] M. Liao, J. Yuan, F. Huang, P. Wang, W. Wang, S. Luo, and Y. Yao, "On-chip silicon optical scattering physical unclonable function towards hardware security," Journal of Lightwave Technology **41**, 1487–1494 (2022).

[15] T. Zhang, Z. Shu, L. Zhang, Y. Chen, Z. Feng, Y. Hu, F. Huang, P. Wang, D. Li, Y. Yao, *et al.*, "Random nanofracture-enabled physical unclonable function," Advanced Materials Technologies **6**, 2001073 (2021).

[16] N. B. Kiremitler, A. Esidir, G. A. Drake, A. F. Yazici, F. Sahin, I. Torun, M. Kalay, Y. Kelestemur, H. V. Demir, M. Shim, *et al.*, "Tattoo-like multi-color physically unclonable functions," Advanced Optical Materials , 2302464 (2023).

[17] A. Esidir, N. Kayaci, N. B. Kiremitler, M. Kalay, F. Sahin, G. Sezer, M. Kaya, and M. S. Onses, "Food-grade physically unclonable functions," ACS applied materials & interfaces **15**, 41373–41384 (2023).

[18] H. Sun, S. Maji, A. P. Chandrakasan, and B. Marelli, "Integrating biopolymer design with physical unclonable functions for anticounterfeiting and product traceability in agriculture," Science Advances **9**, eadf1978 (2023).

[19] A. Esidir, N. B. Kiremitler, M. Kalay, A. Basturk, and M. S. Onses, "Unclonable features via electrospraying of bulk polymers," ACS Applied Polymer Materials **4**, 5952–5964 (2022).

[20] P. Kehayias, E. Bussmann, T.-M. Lu, and A. M. Mounce, "A physically unclonable function using nv diamond magnetometry and micromagnet arrays," Journal of Applied Physics **127** (2020).

[21] N. Torun, I. Torun, M. Sakir, M. Kalay, and M. S. Onses, "Physically unclonable surfaces via dewetting of polymer thin films," ACS applied materials & interfaces **13**, 11247–11259 (2021).

[22] X. Lu, L. Hong, and K. Sengupta, "Cmos optical pufs using noise-immune process-sensitive photonic crystals incorporating passive variations for robustness," IEEE Journal of Solid-State Circuits **53**, 2709–2721 (2018).

[23] D. Dermanis, A. Bogris, P. Rizomiliotis, and C. Mesaritakis, "Photonic physical unclonable function based on integrated neuromorphic devices," Journal of Lightwave Technology **40**, 7333–7341 (2022).

[24] Y. Xu, Y. Lao, W. Liu, Z. Zhang, X. You, and C. Zhang, "Mathematical modeling analysis of strong physical unclonable functions," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **39**, 4426–4438 (2020).

[25] N. Kayaci, R. Ozdemir, M. Kalay, N. B. Kiremitler, H. Usta, and M. S. Onses, "Organic light-emitting physically unclonable functions," Advanced Functional Materials **32**, 2108675 (2022).

[26] A. T. Erozan, M. Hefenbrock, M. Beigl, J. Aghassi-Hagmann, and M. B. Tahoori, "Image puf: A physical unclonable function for printed electronics based on optical variation of printed inks," Cryptology ePrint Archive (2019).

[27] Y. Gao, S. F. Al-Sarawi, and D. Abbott, "Physical unclonable functions," Nature Electronics **3**, 81–91 (2020).

[28] J. Miao, M. Li, S. Roy, Y. Ma, and B. Yu, "Sd-puf: Spliced digital physical unclonable function," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **37**, 927–940 (2017).

[29] Y. Gao, D. C. Ranasinghe, S. F. Al-Sarawi, O. Kavehei, and D. Abbott, "Emerging physical unclonable functions with nanotechnology," IEEE access **4**, 61–80 (2016).

[30] Y. Cao, L. Zhang, S. S. Zalivaka, C.-H. Chang, and S. Chen, "Cmos image sensor based physical unclonable function for coherent sensor-level authentication," IEEE Transactions on Circuits and Systems I: Regular Papers **62**, 2629–2640 (2015).

[31] L. M. Dias, J. F. Ramalho, T. Silvério, L. Fu, R. A. Ferreira, and P. S. André, "Smart optical sensors for internet of things: Integration of temperature monitoring and customized security physical unclonable functions," IEEE Access **10**, 24433–24443 (2022).

[32] A. Vijayakumar, V. C. Patil, and S. Kundu, "On testing physically unclonable functions for uniqueness," in *2016 17th International symposium on quality electronic design (ISQED)* (IEEE, 2016) pp. 368–373.

[33] F. Zerrouki, S. Ouchani, and H. Bouarfa, "Quantifying security and performance of physical unclonable functions," in *2020 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)* (IEEE, 2020) pp. 1–4.

[34] F. K. Wilde, *Metrics for physical unclonable functions*, Ph.D. thesis, Technische Universität München (2021).

[35] N. Wisiol, C. Gräbnitz, C. Mühl, B. Zengin, T. Soroceanu, N. Pirnay, K. T. Mursi, and A. Baliuka, "pypuf: Cryptanalysis of physically unclonable functions," (2021).

[36]H. Kareem, K. Almousa, and D. Dunaev, "Matlab gui-based tool to determine performance metrics of physical unclonable functions," in *2022 Cybernetics & Informatics (K&I)* (IEEE, 2022) pp. 1–5.
[37]A. Clauset, "A brief primer on probability distributions," in *Santa Fe Institute* (2011).
[38]A. Maiti and P. Schaumont, "Improved ring oscillator puf: An fpga-friendly secure primitive," Journal of cryptology **24**, 375–397 (2011).
[39]C. Böhm, M. Hofer, and W. Pribyl, "A microcontroller sram-puf," in *2011 5th International Conference on Network and System Security* (IEEE, 2011) pp. 269–273.
[40]S. Hemavathy and V. K. Bhaaskaran, "Arbiter puf—a review of design, composition, and security aspects," IEEE Access **11**, 33979–34004 (2023).
[41]K. D. Longmate, N. M. Abdelazim, E. M. Ball, J. Majaniemi, and R. J. Young, "Improving the longevity of optically-read quantum dot physical unclonable functions," Scientific Reports **11**, 10999 (2021).
[42]U. Rührmair, C. Hilgers, S. Urban, A. Weiershäuser, E. Dinter, B. Forster, and C. Jirauschek, "Optical pufs reloaded," Cryptology ePrint Archive (2013).
[43]S. Sivaraju, V. Mani, A. Umaamaheshvari, P. D. Banu, T. Anuradha, and S. Srithar, "An attack resistant physical unclonable function smart optical sensors for internet of things for secure remote sensing," Measurement: Sensors **29**, 100882 (2023).
[44]Z. Wang, H. Wang, P. Wang, and Y. Shao, "Robust optical physical unclonable function based on total internal reflection for portable authentication," ACS Applied Materials & Interfaces (2024).
[45]P. Mall, R. Amin, A. K. Das, M. T. Leung, and K.-K. R. Choo, "Puf-based authentication and key agreement protocols for iot, wsns, and smart grids: A comprehensive survey," IEEE Internet of Things Journal **9**, 8205–8228 (2022).
[46]Q. Li, F. Chen, J. Kang, P. Wang, J. Su, F. Huang, M. Li, and J. Zhang, "Intrinsic random optical features of the electronic packages as physical unclonable functions for internet of things security," Advanced Photonics Research **3**, 2100207 (2022).
[47]G. Park, H. Park, J. M. Wolska, J. G. Park, and D. K. Yoon, "Racemized photonic crystals for physical unclonable function," Materials Horizons **9**, 2542–2550 (2022).
[48]A. Fernández-Benito, M. Hoyos, M. A. López-Manchado, and T. J. Sørensen, "A physical unclonable function based on recyclable polymer nanoparticles to enable the circular economy," ACS Applied Nano Materials **5**, 13752–13760 (2022).
[49]T. Silvério, L. Dias, J. F. Ramalho, S. F. Correia, L. Fu, R. A. Ferreira, and P. S. André, "Functional mobile-based two-factor authentication by photonic physical unclonable functions," AIP Advances **12** (2022).
[50]R. Maes, *Physically unclonable functions: Concept and constructions* (Springer, 2013).
[51]A. Maiti, I. Kim, and P. Schaumont, "A robust physical unclonable function with enhanced challenge-response set," IEEE Transactions on Information Forensics and Security **7**, 333–345 (2011).
[52]J. Daugman, "The importance of being random: statistical principles of iris recognition," Pattern recognition **36**, 279–291 (2003).
[53]S. Nocentini, U. Rührmair, M. Barni, D. S. Wiersma, and F. Riboli, "All-optical multilevel physical unclonable functions," Nature Materials **23**, 369–376 (2024).
[54]A. Esidir, M. Ren, S. Pekdemir, M. Kalay, N. Kayaci, N. Gunaltay, H. Usta, X. Huang, and M. S. Onses, "Structurally colored physically unclonable functions with ultra-rich and stable encoding capacity," Advanced Functional Materials , 2417673 (2024).
[55]K. Wang, J. Shi, W. Lai, Q. He, J. Xu, Z. Ni, X. Liu, X. Pi, and D. Yang, "All-silicon multidimensionally-encoded optical physical unclonable functions for integrated circuit anti-counterfeiting," Nature Communications **15**, 3203 (2024).
[56]Y. Gu, C. He, Y. Zhang, L. Lin, B. D. Thackray, and J. Ye, "Gap-enhanced raman tags for physically unclonable anticounterfeiting labels," Nature communications **11**, 516 (2020).
[57]J. Daugman, "How iris recognition works," in *The essential guide to image processing* (Elsevier, 2009) pp. 715–739.
[58]L. E. Bassham III, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, E. B. Barker, S. D. Leigh, M. Levenson, M. Vangel, D. L. Banks, *et al.*, "Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications," (2010).
[59]K. Marton and A. Suciu, "On the interpretation of results from the nist statistical test suite," Science and Technology **18**, 18–32 (2015).
[60]R. Konig, R. Renner, and C. Schaffner, "The operational meaning of min- and max-entropy," IEEE Transactions on Information theory **55**, 4337–4347 (2009).
[61]C. E. Shannon, "A mathematical theory of communication," The Bell system technical journal **27**, 379–423 (1948).
[62]S. Zhu, Y. Ma, T. Chen, J. Lin, and J. Jing, "Analysis and improvement of entropy estimators in nist sp 800-90b for non-iid entropy sources," IACR Transactions on Symmetric Cryptology , 151–168 (2017).
[63]U. Rührmair and D. E. Holcomb, "Pufs at a glance," in *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2014) pp. 1–6.
[64]T. Silvério, L. Dias, R. A. Ferreira, and P. S. André, "Optical authentication of physically unclonable functions using flexible and versatile organic-inorganic hybrids," in *2021 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)* (IEEE, 2021) pp. 1–3.
[65]Z. C. Meijs, H. S. Yun, P. Fandre, G. Park, D. K. Yoon, and L. Isa, "Pixelated physical unclonable functions through capillarity-assisted particle assembly," ACS Applied Materials & Interfaces **15**, 53053–53061 (2023).
[66]J.-L. Zhang, Q. Wu, Y.-P. Ding, Y.-Q. Lv, Q. Zhou, Z.-H. Xia, X.-M. Sun, and X.-W. Wang, "Techniques for design and implementation of an fpga-specific physical unclonable function," Journal of Computer Science and Technology **31**, 124–136 (2016).
[67]U. Ruhrmair and J. Solter, "Puf modeling attacks: An introduction and overview," in *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (2014).
[68]F. Ganji, S. Tajik, F. Fäßler, and J.-P. Seifert, "Strong machine learning attack against pufs with no mathematical model," in *Cryptographic Hardware and Embedded Systems–CHES 2016: 18th International Conference, Santa Barbara, CA, USA, August 17-19, 2016, Proceedings 18* (Springer, 2016) pp. 391–411.
[69]N. Saadvikaa, K. J. Saketi, A. Gopishetti, B. Degala, and K. K. Anumandla, "Puf modeling attacks using deep learning and machine learning algorithms," Engineering Proceedings **56**, 187 (2023).
[70]D. Divyanshu, A. K. Goyal, and Y. Massoud, "Physical unclonable function using photonic spin hall effect," Scientific Reports **14**, 14393 (2024).
[71]J. Tobisch, A. Aghaie, and G. T. Becker, "Combining optimization objectives: New modeling attacks on strong pufs," IACR Transactions on Cryptographic Hardware and Embedded Systems , 357–389 (2021).
[72]A. Diaspro, G. Chirico, C. Usai, P. Ramoino, and J. Dobrucki, "Photobleaching," Handbook of biological confocal microscopy , 690–702 (2006).
[73]P. Révész, *The laws of large numbers*, Vol. 4 (Academic Press, 2014).

VIII. APPENDIX A

Glossary of terms (excluding terms explicitly defined in the text) in alphabetical order:

- Array - A 2D matrix which is used to represent a response. A binary array is a type of array containing only 1s and 0s.
- Authentication - The process of verifying the authenticity of a product or connection. The most common use-case for PUF work. (= anti-counterfeiting)
- Binarisation algorithms - Algorithms used to convert intra-images and inter-images into two-dimensional binary arrays, with the choice of algorithm depending on the researcher.
- Bit-string (=string) (=sequence) - 1D data. A binary version would be composed of 1s and 0s.

- Challenge-response pairs (CRPs) - A pair consisting of an input (challenge) and its corresponding response, which is used by PUFs for authentication.
- Challenge - A controlled stimulus applied to the PUF.
- Electronic Physically Unclonable Functions (E-PUFs) - A subset of PUFs that exploit variations in semiconductor manufacturing, such as differences in transistor threshold voltages or circuit delays, to produce unique and unclonable responses. Defined and compared to O-PUFs in section III A.
- Figures of merit (=metrics) - Mathematical methods used to numerically evaluate the quality of a PUF.
- Fractional Hamming distance (fHD) (=HD in the introduction) - A mathematical measure of similarity between two arrays, as defined in equation 1. A value of 0 indicates identical arrays, 0.5 indicates significant difference, and 1 indicates completely opposite arrays.
- fHD distributions (=intra-fHD and inter-fHD distributions) - The fraction hamming distance values are plotted as a histogram and the normal fits applied are the distributions, these have means and standard deviations.
- FPR+ - A short-cut term for: False Positive Rate, False Negative Rate, True Negative Rate and True Positive Rate.
- HD-based measurements (=HD-based metrics) - Metrics based on comparisons involving Hamming distances. These include 'mean-based' and 'mean-and-standard-deviation-based' metrics.
- Identity (ID) (=fingerprints) (=patterns) - A PUF response, often represented as an array for 2D data and a sequence for 1D data.
- Independently and Identically Distributed (IID) - A mathematical assumption that is used to describe the nature of the responses to PUFs. IID response will behave differently to non-IID and this will result in different relationships between the means and standard deviations of the fHD distributions as shown in FIG. 3. Defined formally in section III A.
- Inter-images - The second set of images, images of different IDs, used to determine uniqueness.
- Intra-images - As part of the main O-PUF analysis mechanism, repeat images are taken of the same IDs. This is later used to determine reliability of the ID.
- μ-based metrics; $\mu\&\sigma$-based metrics; other metrics - As stated in section II, these are classification based on which elements of the fHD distributions are used in the metric calculations.
- μ_1 - Mean of the intra-hamming distribution.
- μ_2 - Mean of the inter-hamming distribution.
- Optical Physically Unclonable Functions (O-PUFs) - A subset of PUFs that use image-based challenge-response pairs, defined further in section III A.
- Physically Unclonable Functions (PUFs) - A physical security primitive that relies on inherently non-reproducible properties to produce challenge-response pairs.
- Pixels (=bits (when considering an optical response)) - Pixels, often indexed as (I, J), represent locations in an image or array. After binarization, each pixel takes a value of 0 or 1. This term can refer to the bits in the image pre-binarization and the bits in the array post-binarization.
- Response - The output uniquely determined by the PUF's physical randomness.
- Sample (=tag) (=material) - for O-PUFs which rely on a physical material that is imaged for the challenge-response, the sample is the individual physical entity used for unique identification.
- Sample Size (N) - The number of intra-image and inter-images. This assumes that both are the same. N_1 and N_2 can be used if these numbers are different. Analysis of the samples sizes used in the literature can be found in TABLE II.
- Set {C} - The set of the intra-arrays.
- Set {D} - The set of the inter-arrays.
- time-dependent (=over-time) - Metrics that assess performance over time, enabling evaluation of ID degradation.
- Time-of-manufacture (=static) - Metrics that assess the PUF at a specific moment, assuming negligible time variation.
- Weak and Strong PUFs - As defined in section III D, these are PUF subcategories distinguished by the number of CRPs they generate.
- XOR - A binary operator that takes pairs of 1s and 0s and when performed follows the rules shown in FIG. 7. It is used in the calculation of fHD.
- σ_1 - Standard deviation of the intra-hamming distribution.
- σ_2 - Standard deviation of the inter-hamming distribution.

IX. APPENDIX B

The following are brief descriptions of the tests that constitute the NIST SP800-22 Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications:[58]

- **Test 1: Frequency (Monobit) Test:** This test examines the proportion of zero to ones across the entire input sequence, verifying the level of equality between the two states falls within expected confidence bounds.
- **Test 2: Frequency Test within a block:** This test examines the proportion of zero to ones across smaller, local blocks of bits within the input sequence, again testing that the level of equality between the two states falls within expected limits.
- **Test 3: Runs Test:** This test examines the length and frequency of occurrence for continuous sub-sequences that contain the same bit parity (known as runs), over the total input sequence. The resulting distribution is then analysed to verify that it conforms to what is expected for ideal randomness (within appropriate certainty bounds).
- **Test 4: Test for the Longest Run of Ones in a Block:** This test splits the input sequence into smaller blocks, counts the length of the largest contiguous subsequence of one bit within the block, and checks that the resulting distribution conforms to that which is expected for ideal randomness (within appropriate certainty bounds).
- **Test 5: Binary Matrix Rank Test:** This test evaluates the rank (the number of independent columns or rows) of matrices formed from the input sequence, looking to detect linear dependencies that would not be part of an ideally random sequence (within appropriate certainty bounds).
- **Test 6: Discrete Fourier Transform (Spectral) Test:** This test applies a Discrete Fourier Transform to the input sequence, converting from the sequential domain to a distribution in the frequency domain. The heights of whatever peaks may be in this resulting spectrum is then evaluated to test for non-random periodic features in the initial input sequence.
- **Test 7: Non-overlapping Template Matching Test:** This test searches across the length of the input sequence to find, or match with, a defined collection of small bit sequences, known as templates. The frequency of occurrence for these bit patterns is then analysed to ensure they conform to what is expected for ideal randomness (within appropriate certainty bounds). In this version of the test there is no overlapping, which is to say that any bit in the input sequence that is part of the matching template once would not be included in further matching of the same template as the search window moves further along. The moving search window 'skips' bits that have already been matched.
- **Test 8: Overlapping Template Matching Test:** This test counts the number of matches of the input sequence with a defined collection of small bit sequence templates, as with the Non-overlapping Template Matching Test. However, in this test bits that have already counted for the matching of a certain template can be included in matches with the same template as the search window moves along, and so overlapping of the template is possible.
- **Test 9: Maurer's "Universal Statistical" Test:** This test examines the number of non-matching bits that lie between matched recurring patterns. This test evaluates the level of compressibility of the input sequence, in other words to what level the sequence can be encoded using fewer bits than it actually contains (through the reuse of recurring patterns) without the loss of any information about the sequence. Here an ideally random sequence should be maximally incompressible.
- **Test 10: Linear Complexity Test:** This test evaluates the length of the linear feedback shift register (LFSR) required to emulate subsequence blocks of the total input sequence. A linear feedback shift register is a pseudorandom number generator based on the shift register, which can be considered as a length of cells (or register) each containing a bit state, where the contents of each cell is passed along one cell (shifted) on a certain cue. In the form LFSR used in this test, the input bit state to the shift register depends on a linear (XOR-ed) combination of the bit states taken at certain points (known as taps) across the register's length, with output at the far end being the next PRNG output state. The more random a sequence is, the longer and therefore more complex the LFSR needs to be to emulate it.
- **Test 11: Serial Test:** This test searches for, and counts the frequency of, the patterns or templates that make up every permutation of bits for of a certain length, allowing for overlapping as with the Overlapping Template Matching Test. In other words, for a pattern of length 3 this test would search for and count each of the 8 possible bit states patterns across the sequence. Every combination of bit states in a pattern subsequence should be equally likely, and so this test checks that the number of matching patterns found for each possible bit state template of the same length is equal (within appropriate confidence bounds).
- **Test 12: Approximate Entropy Test:** This test compares the (overlapping) frequency of occurrence for each possible combination of bit states at a certain pattern length (as with the Serial Test), alongside the frequencies of each state for a pattern one bit longer. This is equivalent to the sum of occurrence for each different pattern of bits found in an overlapping and moving window of a certain length, and length one bit more, passed across the sequence under test. For these two sets of state counts the Shannon entropy (as seen in the body of this work) is calculated and the difference between the two window size entropies is found. This difference in entropy estimates for the two window sizes is then compared to what is expected in a theoretical ideally random distribution.
- **Test 13: Cumulative Sums (Cusum) Test**: This test

first remaps the bits of the input sequence from 0, 1 to -1, 1, changing the zero bit to -1. From this, the test examines the highest magnitude value reached by summing along sub-sequences of increasing length in forward and reverse directions (the maximal excursion of the random walk here defined). For ideally random sequences this maximum cumulative sum should be close to zero.

- **Test 14: Random Excursions Test:** Using the same -1, 1 remapping as in the Cumulative Sums (Cusum) Test, this test examines the number of times summing over the sequence reaches a certain sum value (across the range $\pm$1-4) and returns back to a sum value of zero. This can be considered as the number of cycles to and from a certain (1 dimensional) location in a random excursion defined by the sequence, and is evaluated to ensure each count follows the distribution expected for an ideally random source (within appropriate confidence bounds).

- **Test 15: Random Excursions Variant Test:** Using the same -1, 1 remapping as in the Cumulative Sums (Cusum) and Random Excursions Test, this test examines the number of times the cumulative sum of the sequence equals a certain value (across the range $\pm$1-9) as the summation is performed. This can be considered as the number of visits to a certain (1 dimensional) location that occurs during a random excursion defined by the remapped input sequence. This distribution is then evaluated to ensure each count follows the distribution expected for an ideally random source (within appropriate confidence bounds).