2026 15th International Conference on Communications, Circuits and Systems (ICCCAS) | 979-8-3195-0849-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICCCAS70754.2026.11672432

SRF-PUF: Secret Response Feedback for ML-Resistant Device Authentication

Ching-Che Chung, *Senior Member, IEEE*
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan
wildwolf@cs.ccu.edu.tw

Wei-Xiang Huang
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan
amycok1701@gmail.com

***Abstract*—Strong Physical Unclonable Functions (PUFs) are promising for lightweight device authentication. However, transmitting Challenge–Response Pairs (CRPs) over public networks exposes devices to eavesdropping, rendering them vulnerable to replay and Machine Learning (ML) modeling attacks. This paper presents a Secret Response Feedback PUF (SRF-PUF) and a companion protocol designed to conceal CRP relationships during authentication. The SRF-PUF integrates a self-test-assisted Arbiter PUF for reliable bit selection, a trigger-based LFSR with irregular polynomial updates for challenge obfuscation, and a response-feedback mechanism that disrupts static CRP mappings, even under Non-Volatile Memory (NVM) exposure. Instead of exchanging raw responses, the device transmits a 64-bit authentication signal derived from a secret pattern. Post-layout results demonstrate near-ideal statistics and high reliability (0.996 at 1.0 V over 0 – 75°C). The design achieves 98.75 – 100% authentication success across PVT corners. Furthermore, ML modeling accuracy remains limited to the random-guess level (50.3% for reverse-engineering and 53.0% in NVM-exposed scenarios) using 8k stable CRPs. Implemented in a TSMC 90-nm process, the prototype occupies 0.0193 mm² and consumes 0.693 mW at 50 MHz.**

Keywords—Physical unclonable function (PUF), device authentication, challenge obfuscation, response feedback, machine-learning attack resistance.

I. INTRODUCTION

With the rapid expansion of the Internet of Things (IoT), a massive number of connected endpoints—ranging from sensors and appliances to vehicles and industrial controllers—are required to exchange data over public and often untrusted networks. Before access to network services can be granted, device authentication is typically required to prevent impersonation, message manipulation, and unauthorized control. In resource-constrained IoT nodes, conventional approaches based on storing long-term secret keys in non-volatile memory (NVM) can be costly in area and power, while also remaining vulnerable to invasive readout or reverse-engineering attempts.

Physical unclonable functions (PUFs) have been widely studied as a hardware-rooted security primitive that leverages uncontrollable manufacturing variations to generate chip-unique behavior. In device authentication, a challenge–response pair (CRP) mechanism is typically adopted, where a device-specific response is produced when a challenge is applied and legitimacy is verified by a server using enrolled CRPs. Strong PUFs are particularly attractive for lightweight authentication because a large CRP space can be provided without storing explicit keys [2]. However, strong PUFs deployed in open environments face an inherent exposure problem: transmitted authentication information can be eavesdropped and collected at scale, enabling machine-learning (ML) modeling attacks. Once a sufficiently accurate surrogate model is trained, responses can be predicted and impersonation can be attempted without physical access to the chip. In addition, when response-bearing protocol data remain directly observable on public links, replay and correlation-based exploitation can be attempted.

Considerable effort has been devoted to hardening strong PUF authentication against ML by modifying circuits and by adding protocol-level transformations. Temporal dependency has been introduced by response-feedback mechanisms that feed responses back into an LFSR to poison training data with a compact hardware footprint (e.g., FLAM-PUF) [1]. Reliability has also been enhanced for arbiter-type PUFs because metastability and small delay margins can cause unstable responses across voltage/temperature variations; a Bit-Self-Test (BST) arbiter PUF has been proposed to generate a reliability flag through on-chip self-test so that unstable CRPs can be filtered [3]. Protocol-oriented protections have been investigated as well. Random set-based obfuscation (RSO) has been introduced by selecting stable-response keys (via TRNG) to XOR both challenges and responses, with key updates applied to bound the number of usable CRPs [4]. Active deception has been proposed by mixing genuine and fake strong PUF behaviors and returning decoy responses under suspicious query behavior [5]. Dynamic responding mechanisms have also been suggested, where externally supplied challenges are transformed into time-varying internal challenges using an LFSR combined with a weak PUF, and a companion authentication protocol has been designed to mitigate Hamming-weight-assisted CMA-ES attacks [6]. In addition, circuit-level obfuscation has been explored through stagewise obfuscated interconnections to disrupt linear additive delay modeling while improving reliability through filtering mechanisms [7], and nonmonotonic response quantization has been demonstrated to force nonmonotonic input–output behavior for improved resistance against advanced learning [8].

Despite these advances, practical limitations are often observed: auxiliary information or response-bearing outputs may remain observable on the public channel, replay risks may persist, additional NVM or configuration secrecy may be

This work was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 114-2221-E-194-060-.

required, and reliability degradation or implementation overhead can be introduced.

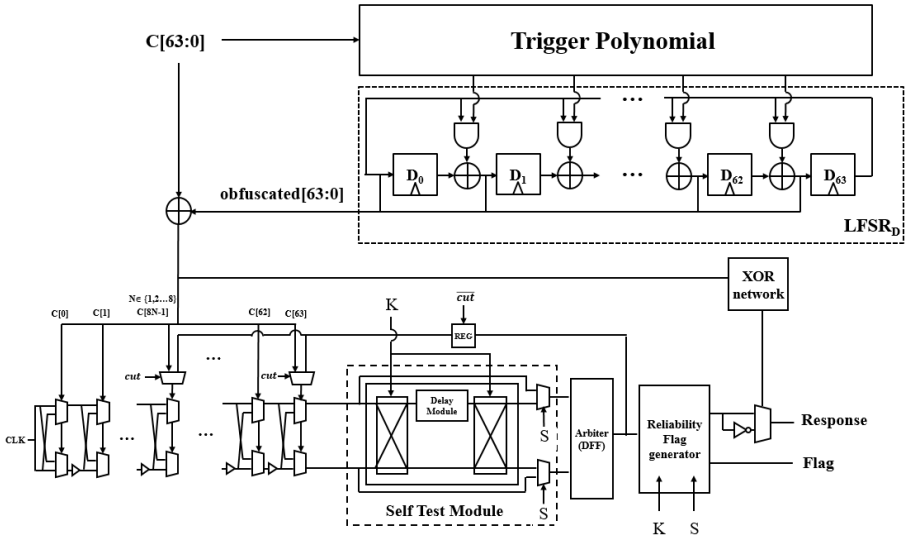

Fig. 1. The circuit of secret response feedback (SRF) PUF with trigger LFSR

From this landscape, a critical gap is identified: strong ML resistance alone is insufficient when usable CRP correspondence remains observable over a public channel. Even when learned accuracy is reduced, replay and correlation attacks can still be attempted when exchanged data remain structurally linked to underlying CRPs. Moreover, protocol confidentiality can collapse under NVM-exposure assumptions, where leaked seeds, polynomials, or enrollment-time configuration values allow obfuscation to be reconstructed and modeling to be accelerated. Therefore, a mechanism is required in which (i) CRP correspondence is intrinsically hidden during authentication, (ii) high reliability is maintained across process/voltage/temperature (PVT) variations to sustain authentication success rate, and (iii) robustness is preserved even when circuit information and selected NVM contents are assumed to be exposed.

To address these requirements, a Secret Response Feedback PUF (SRF-PUF) and a corresponding device–server authentication protocol are presented. In the proposed approach, challenge obfuscation is strengthened through a trigger-based LFSR mechanism with irregular polynomial updates, while static CRP correspondence is disrupted by response feedback in which selected challenge bits are replaced with internally generated feedback responses. Reliability is enhanced using a self-test-assisted flag generation mechanism to filter unstable CRPs, and response statistics are balanced through a lightweight XOR-network-assisted randomization scheme. Most importantly, during authentication, CRP correspondence is concealed by transmitting an authentication signal derived from a secret pattern rather than transmitting raw responses, thereby mitigating replay and large-scale CRP harvesting over public networks. Under a 64-bit authentication signal with a Hamming-distance tolerance threshold of 11 bits, 98.75–100% authentication success has been achieved across representative voltage – temperature corners, supported by approximately 99% response reliability. Uniformity and uniqueness have been maintained near ideal values (≈0.5). In ML evaluations, prediction accuracy has been constrained near the random-guess level in reverse-engineering scenarios and has remained near-random under NVM-exposure assumptions due to response-feedback-induced decorrelation.

The main contributions can be summarized as follows. First, an SRF-PUF architecture is provided that jointly targets reliability, statistical quality, and ML resistance through combined self-test filtering, response balancing, and response feedback decorrelation. Second, a hidden-CRP authentication protocol is introduced in which CRP correspondence is concealed on public links by transmitting only an authentication signal derived from a secret pattern, improving robustness against replay and CRP-collection-driven modeling. Third, feasibility is demonstrated through implementation in a TSMC 90-nm CMOS process, with compact area and low power reported under a practical operating frequency.

II. The Proposed SRF-PUF

Security primitives for IoT endpoints are constrained by limited area/power budgets and the absence of secure non-volatile storage. Strong PUFs are attractive because a large challenge-response space enables lightweight authentication without long-term secret storage. However, strong PUFs deployed over public networks remain exposed to two practical threats. First, challenge-response pairs (CRPs) can be eavesdropped during transmission and accumulated for machine-learning (ML) modeling, enabling response prediction. Second, protocol-level leakage can still enable replay or impersonation when responses (or response-derived information) remain directly observable. A protection mechanism is therefore required that increases resistance to ML attacks at the circuit level while preventing usable CRP relationships from being revealed during device-server exchanges.

A Secret Response Feedback PUF (SRF-PUF) and a companion authentication protocol are presented. A hidden-

CRP principle is adopted: response bits generated inside the PUF are prevented from being directly exposed on the public network, while server-side authentication is still supported. The circuit architecture is illustrated in Fig. 1, where three objectives are jointly addressed: (i) CRP correlation is concealed through challenge obfuscation and response feedback, (ii) reliability is increased through a self-test mechanism that labels unstable outputs, and (iii) response uniformity is maintained despite reliability-based discarding. The protocol flow is partitioned into an enrollment phase and an authentication phase, as shown in Fig. 2 and Fig. 3, respectively.

As depicted in Fig. 1, a 64-bit input challenge vector *C[63:0]* is obfuscated before being applied to the delay-based PUF core. Obfuscation is performed by XORing the external challenge with an internally generated obfuscation signal *obfuscated[63:0]* produced by a Trigger LFSR. The effective challenge applied to the PUF core is thus decoupled from the challenge observed at the interface, reducing the utility of eavesdropped CRPs even when the circuit structure is assumed to be known. Unlike a conventional fixed-polynomial LFSR, the Trigger LFSR performs irregular polynomial updates so that stable obfuscation regularities are not retained over long CRP acquisition periods.

The PUF core is constructed around an arbiter-style delay comparison. The response bit is determined by the relative arrival times at the arbiter, which is vulnerable to metastability when the phase difference is small. To prevent unstable response bits from degrading authentication, a self-test mode is integrated. A per-instance reliability flag is generated and is later used to identify stable outputs for enrollment and to support robust verification during authentication.

A response-processing stage is integrated to address a common side-effect of reliability screening. When unstable CRPs are discarded, a skewed output distribution can emerge because values near the arbiter dead zone tend to be mapped to a preferred logic value. To correct this imbalance, an XOR-based post-processing network is applied so that controlled inversion of responses is induced with an approximately balanced probability. This post-processing is performed without altering the physical delay paths, thereby avoiding reliability degradation that would otherwise occur if extra nonlinear structures were inserted inside the PUF delay stages.

A Bit-Self-Test (BST)-style operation is employed to identify response bits that are likely to flip under voltage or temperature variation. The circuit is operated in multiple modes controlled by internal signals (denoted by control lines such as *S* and *K* in Fig. 1). In response-output mode, the arbiter output is latched and treated as the candidate response. In test-output mode, controlled delay offsets are inserted into either the upper or the lower path, and the arbiter output is re-evaluated under these offset conditions. A reliability decision is then formed by comparing the two test outputs.

Reliability flag generation is summarized as follows. First, the nominal arbiter output is captured in a register during response-output mode. Next, in test-output mode, the upper path is passed through a programmable delay module while the lower path remains unmodified; the resulting arbiter output is stored as a first test sample. Then, the lower path is passed through the delay module while the upper path remains unmodified; the resulting arbiter output is stored as a second test sample. The two test samples are XORed to generate the reliability flag. If the samples match, the intrinsic delay difference ΔD is inferred to be larger than the applied delay window and the bit is treated as stable. If the samples differ, ΔD is inferred to be close to the metastable region and the bit is treated as unstable for stable-set selection. Reliability is quantified through on-chip digital operations, and the delay window can be tuned to trade throughput for stability.

Reliability-based discarding can introduce response bias. When a large fraction of CRPs is filtered out, the remaining set can contain an imbalanced ratio of '0' and '1', which degrades uniformity and may leak structural information. To mitigate this, conditional inversion is induced based on selected challenge bits. An inversion control signal is generated from a subset of the 64 challenge bits and is used to select whether the nominal response or its inverted form is released. By selecting a challenge-bit combination with an approximately 0.5 probability of asserting inversion, the output distribution is driven closer to the ideal 50/50 balance. Importantly, this operation is confined to response post-processing; the delay-chain structure and arbiter decision remain unchanged.

Although the self-test mechanism can deliver high reliability, excessive CRP discarding is undesirable because usable entropy is reduced and enrollment overhead is increased. An additional mechanism is therefore integrated to enlarge the phase difference between the two competing paths in a challenge-dependent manner, thereby reducing the probability that the arbiter operates near its metastable region. As shown in Fig. 1, buffers are inserted into selected portions of one delay path (e.g., the lower path) between multiplexers across the delay stages. When a path switch is exercised by a given challenge, extra delay is introduced and the phase-difference margin is increased. Because insertion is challenge-dependent, the phase difference is effectively randomized across CRPs rather than applied as a fixed offset, enabling fewer CRPs to be discarded while maintaining a high reliability level.

Even with challenge obfuscation, exposure of internal configuration (e.g., LFSR seeds or polynomials) can weaken protection if the effective internal challenge becomes reconstructible. To strengthen resilience under such leakage, a response feedback mechanism is employed so that part of the effective challenge is derived from internal, challenge-dependent PUF behavior that is not transmitted during authentication. The response feedback mechanism operates in two conceptual stages. In the first stage, a feedback response R_{FB} is generated under a designated mode and is retained internally. In the second stage, selected bits of the next applied challenge are replaced with bits of R_{FB}. For example, substitution of periodic bit positions (e.g., the 8th, 16th, 24th, ..., 64th bits) creates a composite challenge that is partially determined by the PUF's own prior response. This breaks the static mapping between an externally observed challenge and the internal effective challenge, since the effective challenge depends on a response value that is neither fixed nor externally observable in the authentication phase. Instability propagation is limited by ensuring that feedback is regenerated under controlled operation, so that marginal behavior is localized to the associated CRP

instance rather than permanently corrupting subsequent operation.

LFSR-based challenge obfuscation is commonly adopted to raise modeling difficulty, but a fixed polynomial can imprint periodicity that becomes observable with large datasets. To avoid this, the Trigger LFSR in Fig. 1 is operated with an irregular polynomial update mechanism. Polynomial bits are conditionally inverted based on the occurrence of a specific challenge-bit pattern at a prescribed interval (e.g., a pattern match checked every 1,000 challenges). The obfuscation sequence is therefore prevented from remaining stationary over long CRP acquisition periods, and cyclic regularity typically associated with a conventional LFSR is disrupted. This irregular update behavior blocks reliable alignment of collected CRPs to a single stationary transformation and prevents the obfuscation generator from being trivially reduced to a fixed mapping across the collected dataset.

Circuit-level protections alone are insufficient when responses are directly transmitted over a public channel, because eavesdropped responses can still be used for replay, impersonation attempts, or offline model building. The SRF-PUF protocol is therefore designed so that responses are concealed during transmission and an authentication signal is transmitted instead. The protocol is separated into enrollment and authentication, following Fig. 2 and Fig. 3.

The enrollment phase shown in Fig. 2 is executed in a trusted environment. One-time-programmable NVM is programmed with secret configuration values associated with the obfuscation generator, including polynomial parameters and a random seed. A challenge generator (denoted as $LFSR_C$) is seeded by a server-provided seed *Ca* to generate a deterministic sequence of challenges used for model construction. During enrollment, internal obfuscation and response feedback are enabled so that the actual effective challenges applied to the PUF can be derived within the trusted environment. For each generated challenge, both the feedback response R_{FB} and the nominal response *R* are collected together with the associated reliability flags. Stable responses are identified through the flags and used to form a stable dataset for server-side model training. In addition, a secret pattern used for subsequent authentication-signal generation is derived from the programmed NVM parameters and is stored on the server side.

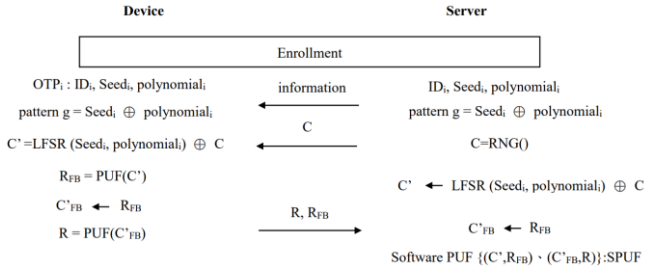

Fig. 2. The enrollment phase of the SRF-PUF protocol.

The authentication phase shown in Fig. 3 is performed after deployment over a public network. A fresh challenge seed *Ca* is transmitted from the server to the device. The challenge generator $LFSR_C$ produces a deterministic challenge sequence *Cb*, which is obfuscated by XOR with the Trigger LFSR output. Response feedback is executed internally to derive modified challenges C'_{FB} from feedback responses, and a response sequence *R[k]* is generated.

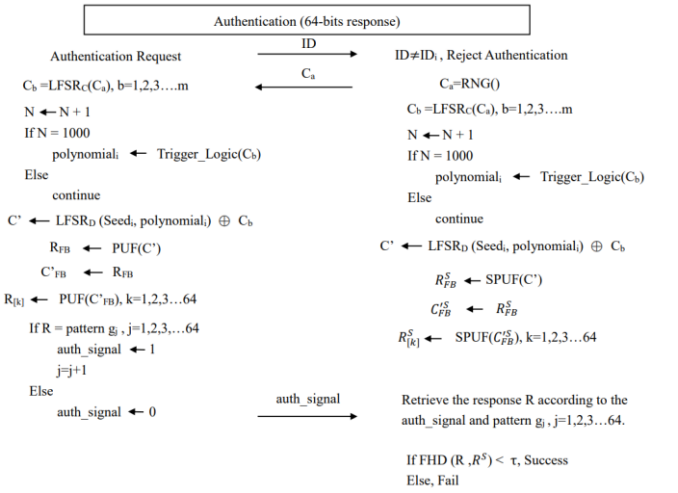

Fig. 3. The authentication phase of the SRF-PUF protocol.

Instead of transmitting *R[k]*, an authentication signal *auth_signal* is produced by comparing the internally generated response sequence against the secret pattern for a prescribed number of iterations (e.g., a 64-iteration construction for a 64-bit authentication signal). Reliability flags may also be output to indicate bit positions that were screened as stable, allowing robust matching at the server under environmental variation. On the server side, the stored model and enrollment-derived pattern are used to reconstruct the expected response sequence and the corresponding expected authentication signal. Authentication is decided by comparing the received and expected authentication signals under a Hamming-distance threshold τ, enabling tolerance to residual modeling error and environmental perturbations. Because the exchanged value is an authentication signal derived through a secret pattern, replay utility is reduced when fresh seeds are used across sessions, and CRP correspondence required for straightforward impersonation is suppressed.

Layered protection is thus provided across attacker capabilities. When circuit structure is assumed to be publicly known, challenge obfuscation by the Trigger LFSR and hidden-CRP signaling suppress direct learning of internal CRP mappings. When NVM parameters are assumed to be exposed, response feedback prevents reconstruction of effective challenges from external observations because internal feedback required for challenge substitution is not transmitted during authentication. When public-network eavesdropping is assumed, response exposure is avoided by transmitting an authentication signal rather than raw responses, thereby limiting replay-oriented misuse and large-scale CRP harvesting.

III. Experimental Results

The experimental validation of the proposed SRF-PUF was conducted using a TSMC 90-nm CMOS standard-cell implementation, with post-layout simulations and protocol-level evaluations performed under voltage/temperature variations. The layout of the test chip is illustrated in Fig. 4, where the core

TABLE I. PERFORMANCE OF THE PROPOSED PUF.

Metric	Condition / Setup	Result (Mean with Min–Max Range)
Uniformity	Post-layout sim, 10k CRPs Corners: Worst (SS), Typ (TT), Best (FF)	Worst: 0.5052 (0.5042–0.5061) Typ: 0.5045 (0.5037–0.5051) Best: 0.5048 (0.5041–0.5057)
Uniqueness	16 chips, 64-bit responses Corners: Worst (SS), Typ (TT), Best (FF)	Worst: 49.10% (48.91–49.28) Typ: 48.97% (48.80–49.19) Best: 48.77% (48.65–48.91)
Reliability	Temp sweep (0–75°C) @1.0V Voltage sweep (0.9–1.2V) @25°C	Temp: 0.9962 (0.9948–0.9974) Volt: 0.9905 (0.9897–0.9916)
ML Resistance (Rev. Engineering)	Training: 8k stable CRPs Algorithms: {MLP, LR, SVM, GB, CMA-ES}	Accuracy: 50.30% (47.56–52.25) (Ideal is 50%)
ML Resistance (NVM Exposed)	Scenario: Leaked seed/polynomial Feedback Mechanism: Active	Accuracy: 53.01% (50.81–55.38) (Attacker restricted near 50%)
Server Model Acc.	Software model for Authentication Training: 8k stable CRPs	Accuracy: 92.93% (89.81–95.38) (Server accuracy ≫ Attacker accuracy)
Auth. Success Rate	Threshold τ=11/64 Corners: Worst (0.9V, 75°C), Typ, Best	Worst: 98.75% Typ: 100% Best: 100%

size is 330 µm × 330 µm and the die size including I/O pads is 866 µm × 866 µm. The active area is 0.0193 mm², which is composed of the SRF-PUF area (0.0024 mm²), the Modified_LFSR area (0.0144 mm²), and the controller area (0.0025 mm²). Under authentication-mode operation, the maximum operating frequency is 50 MHz, and the power consumption is 0.6932 mW at 50 MHz.

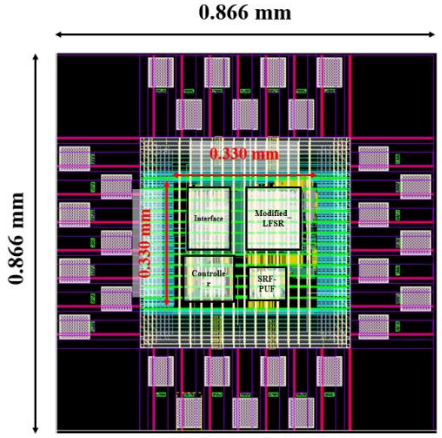

Fig. 4. The layout of proposed SRF-PUF chip.

To support both enrollment and deployment use cases, multiple operating modes were considered. During enrollment, Challenge–Response Pairs (CRPs) were collected to enable statistical characterization (uniformity, uniqueness, and reliability) and to train a server-side software model for authentication. During authentication, only the authentication-related output was assumed to be observable on the public network, while direct responses (and the corresponding CRP linkage) were concealed by the protocol design. This separation was used to evaluate both intrinsic PUF quality and system-level resilience against adversaries operating under different knowledge assumptions.

Post-layout Monte Carlo SPICE simulations were carried out across representative voltage–temperature corners, including a worst-case corner, a typical corner, and a best-case corner. The evaluation flow emphasized the use of the reliability flag mechanism to select stable CRPs for model training and authentication, while preserving response balance through the adopted output randomization strategy. A consolidated summary of key performance metrics is provided in Table I, including uniformity, uniqueness, reliability, machine-learning (ML) resistance under multiple threat scenarios, server-model accuracy, and end-to-end authentication success rate.

Uniformity and uniqueness. Uniformity was evaluated using 10,000 CRPs under the three corners. As reported in Table I, the mean uniformity remained close to the ideal value of 0.5 across corners, achieving 0.5052 (worst), 0.5045 (typical), and 0.5048 (best), with tight min–max ranges. These results indicate that the response distribution was maintained near-balanced even under corner variations, which is essential for preventing bias-driven leakage and for improving the effectiveness of subsequent matching. Uniqueness was assessed using 64-bit responses from 16 instances, and the inter-chip Hamming distance remained close to the ideal 50%: 49.10% (worst), 48.97% (typical), and 48.77% (best) as summarized in Table I. The near-ideal uniqueness suggests that strong device-to-device discrimination was retained despite the additional on-chip mechanisms introduced for obfuscation and protocol support.

Reliability under PVT variations. Reliability was characterized through temperature and voltage sweeps, focusing on the stability of responses when environmental conditions change. As shown in Table I, the reliability reached 0.9962 (mean) during a temperature sweep of 0–75°C at 1.0 V, and 0.9905 (mean) during a voltage sweep of 0.9–1.2 V at 25°C. These values correspond to approximately 99% stability and indicate that the proposed architecture can sustain low bit-flip probability across a meaningful operating range, which directly affects both authentication success rate and the server-side tolerance required for robust matching.

Machine-learning resistance. reverse engineering and NVM-exposed scenarios. Two adversarial settings were evaluated to reflect realistic attack progressions. In the first scenario (reverse engineering/eavesdropping-oriented), the attacker was assumed to observe transmitted authentication-

TABLE II. PERFORMANCE COMPARISONS OF THE PUFS.

	[4] IEEE TCAS'21	[5] IEEE TCAD'21	[6] IEEE TVLSI'22	[7] IEEE IOT'23	[8] IEEE TVLSI'23	Proposed work
Technology	28nm	28nm	45nm	28nm	65nm	90nm
Type of PUF	Arbiter PUF	Lightweight PUF	Arbiter PUF	Arbiter PUF	Ring oscillator PUF	Arbiter PUF
Experimental method	FPGA	FPGA	FPGA	FPGA	Chip	Simulation
Machine Learning Prediction Accuracy	~55% Training data: 100,000	~50% Training data: 10,000	~53.67% Training data: 50,000	~50% Training data: 10,000,000	~50% Training data: 2,000,000	~50% Training data: 1,000,000
Uniformity	N/A	0.470	N/A	~0.5	0.505	~0.5
Reliability	0.948	0.939	N/A	~0.99	~0.866	~0.994
Uniqueness	0.498	~0.4	N/A	~0.45	0.5	~0.493
Voltage range	N/A	0.8~1.0(V)	N/A	N/A	N/A	0.9~1.1(V)
Temperature range	N/A	0~70 (ºC)	N/A	-20~100 (ºC)	0~50 (ºC)	0~75 (ºC)
Power	N/A	101mW @ 100MHz	N/A	N/A	0.237mW	0.6932mW @ 50MHz
Reverse engineering attack	√	√	√	√	√	√
Replay attack	√	×	√	×	×	√
Hidden CRPs	×	×	√	×	×	√
NVM attack	×	×	No NVM	No NVM	No NVM	√

related information and attempt to reconstruct a predictive model. Using 8,000 stable CRPs as training data, multiple learning algorithms (including MLP, LR, SVM, GB, and CMA-ES) were evaluated. As reported in Table I, the mean prediction accuracy was 50.30% with a range of 47.56%–52.25%, which is effectively indistinguishable from random guessing and indicates that the exposed information was insufficient for learning the underlying CRP mapping.

In the second scenario, exposure of enrollment-time secrets (e.g., seed/polynomial stored in NVM/OTP) was assumed, which weakens obfuscation mechanisms relying solely on hidden initialization. Under this stronger attacker model, the response-feedback mechanism was used as the primary protection layer to disrupt CRP correspondence. As summarized in Table I, modeling accuracy was constrained to 53.01% on average (range 50.81%–55.38%) under the evaluated setting, remaining close to the random-guessing baseline. Additional large-data simulations further indicated that, if extremely large datasets are available, certain algorithms can partially increase prediction accuracy; however, the authentication protocol was designed to preserve a practical separation between the server and attacker capabilities through concealed CRP linkage and thresholded matching.

Server modeling accuracy and authentication success rate. A server-side software model was trained using 8,000 stable CRPs collected at nominal conditions to emulate the expected deployment flow. The resulting server-model accuracy reached 92.93% on average (range 89.81%–95.38%) as listed in Table I, enabling reliable reconstruction of expected authentication behavior under the protocol. System-level authentication was then evaluated using a 64-bit authentication signal and a Hamming-distance threshold set to $\tau = 11/64$. Under this configuration, the authentication success rate achieved 98.75% at the worst corner and 100% at the typical and best corners, as reported in Table I. These results demonstrate that high authentication success can be sustained while maintaining strong resistance against modeling-based impersonation, given that attacker prediction accuracy is kept far below the server's effective verification capability.

Comparison with prior art. A performance comparison against representative recent PUF protection schemes is summarized in Table II, covering technology node, PUF type, experimental method, ML attack accuracy and training size, and key quality metrics (uniformity, reliability, uniqueness), together with security features such as reverse-engineering resistance, replay-attack considerations, hidden-CRP capability, and NVM-attack resilience. As indicated in Table II, many existing approaches achieve ML resistance near 50% but may lack simultaneous protection against replay attacks or may not explicitly support hidden CRPs on public networks, while other designs introduce reliability degradation or require additional assumptions (e.g., secure memory storage). In contrast, the proposed SRF-PUF integrates concealed CRP correspondence

with response-feedback-based disruption of modeling attempts, while maintaining approximately 0.994 reliability and near-ideal uniformity/uniqueness, together with low power consumption (0.6932 mW @50 MHz) in TSMC 90-nm CMOS process implementation.

IV. Conclusion

An SRF-PUF architecture and a hidden-CRP device–server protocol were developed to harden strong-PUF authentication over public networks. Challenge obfuscation was reinforced through a trigger-based LFSR with irregular polynomial updates, and CRP correspondence was further broken through response feedback, in which selected challenge bits were replaced by internal feedback responses. Stable-bit screening was enabled by a self-test-assisted reliability flag, while uniformity loss from discarding was compensated by an XOR post-processing network. In TSMC 90-nm CMOS post-layout results, 0.0193 mm^2 and 0.693 mW at 50 MHz were obtained, with ~0.505 uniformity, ~49% uniqueness, ~0.996 reliability, 98.75–100% authentication success, and ~50–53% ML predict accuracy.

References

[1] Linjun Wu, Yupeng Hu, Kehuan Zhang, Wenjia Li, Xiaolin Xu, and Wanli Chang, "FLAM-PUF: A response–feedback-based lightweight anti-machine-learning-attack PUF," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 1, pp. 4433-4444, Nov. 2022.

[2] G. Edward Suh and Srinivas Devadas, "Physical unclonable functions for device authentication and secret key generation," *in Proceedings of 44th ACM/IEEE Design Automation Conference (DAC)*, Jun. 2007, pp. 9-14.

[3] Zhangqing He, Wanbo Chen, Lingchao Zhang, Gaojun Chi, Qi Gao, and Lein Harn, "A highly reliable arbiter PUF with improved uniqueness in FPGA implementation using bit-self-test," *IEEE Access*, vol. 8, pp. 181751-181762, Oct. 2020.

[4] Jiliang Zhang and Chaoqun Shen, "Set-based obfuscation for strong PUFs against machine learning attacks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 288-300, Jan. 2021.

[5] Chongyan Gu, Chip-Hong Chang, Weiqiang Liu, Shichao Yu, Yale Wang, and Máire O'Neill, "A modeling attack resistant deception technique for securing lightweight-PUF-based authentication," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 6, pp. 1183-1196, Jun. 2021.

[6] Yale Wang, Chenghua Wang, Chongyan Gu, Yijun Cui, Máire O'Neil, and Weiqiang Liu, "A generic dynamic responding mechanism and secure authentication protocol for strong PUFs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 9, pp. 1256-1268, Sep. 2022.

[7] Chongyao Xu, Litao Zhang, Man-Kay Law, Xiaojin Zhao, Pui-In Mak, and Rui P. Martins, "Modeling-attack-resistant strong PUF exploiting stagewise obfuscated interconnections with improved reliability," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 16300-16315, Sep. 2023.

[8] Kleber Stangherlin, Zhuanhao Wu, Hiren Patel, and Manoj Sachdev, "Enhancing strong PUF security with nonmonotonic response quantization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 1, pp. 55-64, Jan. 2023.