

RESEARCH ARTICLE OPEN ACCESS

Physical Unclonable Function Based on 3D-NAND Flash Array Structure With Multi-Chip Implementation

 Hwiho Hwang  | Sangwook Youn | Jinwoo Park | Dayeon Yu | Hyungjin Kim 

Division of Materials Science and Engineering and Department of Semiconductor Engineering, Hanyang University, Seoul, South Korea

Correspondence: Hyungjin Kim (hkim12@hanyang.ac.kr)

Received: 26 May 2025 | **Revised:** 15 September 2025 | **Accepted:** 9 June 2026

Keywords: charge trapping material | in-memory computing | NAND flash | physical unclonable function (PUF)

ABSTRACT

As a promising hardware-based security primitive, physical unclonable functions (PUFs) exploit stochastic, process-induced variations in physical properties to generate unique and unclonable device fingerprints. However, conventional CMOS-based PUFs require large and complex circuitry, leading to increased area and power consumption. To address this, we propose a PUF based on a 3D-NAND flash array utilizing intrinsic string current variations. Unlike prior write-based approaches, it employs read operations for CRP extraction, thereby avoiding stress-induced degradation. Challenges are applied via string select lines and bitlines, while each wordline functions as an independent PUF chip. A 1-bit response is generated by comparing summed string currents. Thanks to the high cell density of 3D-NAND flash, multiple stacked PUF chips with an exponentially large CRP capacity can be integrated into a single array. We fabricated 48×24 planar NAND flash arrays using charge-trap flash, confirming current variability due to process variation and randomly formed poly-Si grain boundaries. A $48 \times 24 \times 8$ 3D-NAND PUF was implemented using measured data, achieving near-ideal PUF metrics and passing autocorrelation analysis, all 16 NIST statistical tests and ML attack simulations. These results highlight the potential of 3D-NAND flash for scalable, reliable hardware security.

1 | Introduction

The Internet of Things (IoT) enables wireless connectivity between edge devices and cloud services, facilitating real-time data exchange for automation and improved efficiency [1–3]. As IoT continues to expand, sensitive data is increasingly processed at the edge, heightening exposure to security threats such as hacking and data leakage. This has driven demand for stronger and more reliable security solutions beyond conventional software-based approaches [4, 5]. In response, recent efforts have focused on hardware-based security, utilizing intrinsic physical properties as entropy sources for cryptographic applications [6–13]. A physical unclonable function (PUF) is a hardware-based security primitive that exploits inherent process variations to generate unique and unclonable physical fingerprints [14, 15]. A PUF can be modeled as a function $r = f(c)$, where a given challenge c pro-

duces a corresponding response r , forming a challenge–response pair (CRP). Based on the number of possible CRPs, PUFs are classified as either weak or strong. Weak PUFs provide a limited CRP set and are mainly used for on-chip key generation, whereas strong PUFs support exponentially large CRP spaces, enabling device authentication through external challenges. Therefore, ensuring high randomness and strong uniqueness in responses is critical for secure and reliable operation.

Conventional PUFs are typically implemented using complementary metal–oxide–semiconductor (CMOS) technology, including arbiter PUFs, ring oscillator PUFs, and SRAM PUFs, which exploit process variations reflected in multiplexer delays, oscillator frequencies, and SRAM power-up states, respectively [16–20]. However, these designs often require large and complex circuits, resulting in increased area and power consumption. Moreover,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2026 The Author(s). *Small* published by Wiley-VCH GmbH

even if CRPs are exposed through external attacks, they cannot be modified. To address these limitations, recent studies have focused on PUF designs based on non-volatile memory (NVM), which offer more compact architectures with improved area, power, and cost efficiency. Among NVM-based PUFs, resistive random-access memory (RRAM) has been the most extensively studied, leveraging crossbar arrays for high integration density and low-cost fabrication. RRAM-based PUFs utilize intrinsic variations in resistance states, current–voltage (I – V) nonlinearity, and switching behavior [21–28]. Despite their advantages, these designs can suffer from issues such as line resistance and sneak path currents. Ferroelectric memories, including ferroelectric tunneling junctions (FTJs) and ferroelectric field-effect transistors (FeFETs), have also been explored as entropy sources by exploiting variations in ferroelectric domain configurations [29–33]. Ferroelectric-based PUFs are promising due to strong entropy characteristics, CMOS compatibility, and low power consumption. Nonetheless, reliability concerns such as limited endurance and retention still persist. FeFET-based implementations also face challenges in 3D integration, which is key to achieving higher density and expanded CRP capacity.

Flash memory, one of the most mature and reliable NVM technologies, has also been investigated for PUF applications. These approaches primarily utilize threshold voltage (V_{th}) variations arising from write-induced variability in flash memory, such as program operations, program disturb, latency differences, and erase speed. Moreover, some studies have utilized the high integration density of NAND flash arrays to realize high-density PUFs [34–42]. However, write-based CRP extraction methods require extensive program cycles to generate sufficiently random and balanced responses. These operations stress flash cells, which inherently have limited endurance, and repeated program/erase (P/E) cycling degrades device characteristics, reducing the reliability of generated responses [43, 44]. Accordingly, PUFs based on intrinsic V_{th} variation in flash memory cells have also been explored. Flash memory exhibits inherent V_{th} variability due to process-induced fluctuations, such as variations in cell geometry, interface traps, line edge roughness, and random dopant fluctuations [45–48]. In 3D-NAND flash structures with polycrystalline silicon (poly-Si) channels, additional V_{th} variation arises from irregular grain boundaries and associated trap states, leading to intrinsic variation in string current [49–53]. However, prior work was limited to weak PUF implementations with small CRP spaces due to restricted CRP extraction [54]. Other studies achieved strong PUFs using 3D-NAND arrays but relied on simultaneous challenge input across multiple wordlines (WLs), which reduced current variation compared to single-WL access and degraded response randomness. Moreover, activating multiple WLs lowered the sensing current and increased read latency [55]. Nevertheless, 3D-NAND flash, with its vertically stacked structure and established commercial reliability, remains a compelling option for implementing cost-efficient and robust PUFs with large CRP capacity.

In this work, we present a PUF based on a 3D-NAND flash array, where CRPs are generated by exploiting the intrinsic variation of flash memory. First, we fabricated a 48×24 planar NAND array utilizing charge trap flash (CTF), evaluated its electrical characteristics, and confirmed intrinsic variation in bitline (BL) current, which can be used as an entropy source for PUF

implementation. Based on this, we propose a 3D-NAND flash-based PUF in which challenges are applied through string select lines (SSLs) and BLs, and responses are generated by comparing the summed BL currents. Furthermore, each WL is configured as an independent PUF instance, allowing for a large number of unique and random CRPs per chip. The vertically stacked structure of 3D-NAND also enables multi-chip integration within a single array. To validate this, we used measured data from eight fabricated 48×24 arrays to construct a virtual $48 \times 24 \times 8$ 3D-NAND flash array. The extracted CRPs exhibited near-ideal randomness according to standard PUF metrics, including uniformity, diffuseness, and uniqueness. The response bits also passed both the autocorrelation function (ACF) test and all 16 National Institute of Standards and Technology (NIST) randomness tests, confirming their independence and unpredictability. Finally, we evaluated robustness under environmental variations, including temperature and bias fluctuations. Under temperature variation, the BER at 85 °C was reduced to 5.42% with the application of TMV3, confirming resilience against thermal fluctuations. Under WL and BL bias variations of $\pm 20\%$, the BER remained low at 4.71% and 2.64% in the worst case, validating the stability of the proposed. Furthermore, machine learning (ML) attack simulations were performed, and the results showed that even after extensive training, unseen CRPs remained unpredictable, demonstrating strong resilience to modeling attacks and reliable randomness of the proposed PUF system.

2 | Results and Discussion

2.1 | Device Fabrication and Characteristics

Figure 1a illustrates the fabrication process flow of the 48×24 planar NAND flash array. The process began with the formation of a 300 nm-thick buried oxide (BOX) layer on a p-type bulk silicon wafer via wet oxidation in a furnace. Next, a 100 nm-thick amorphous silicon (a-Si) layer was deposited at 550 °C using low-pressure chemical vapor deposition (LPCVD) with 100 sccm of SiH_4 . Solid-phase crystallization (SPC) was then performed by annealing the a-Si at 600 °C for 24 h in a furnace to form poly-Si, which serves as the channel layer. The active region was subsequently patterned through etching and isolation. The gate stack was fabricated as follows: a 5 nm-thick tunneling oxide was deposited at 767 °C and 200 mTorr via LPCVD using 40 sccm of dichlorosilane (DCS, SiH_2Cl_2) and 80 sccm of N_2O . Then, a 6 nm-thick Si_3N_4 charge trapping layer was deposited at 800 °C and 200 mTorr using 40 sccm of DCS and 400 sccm of NH_3 . A 9.5 nm-thick Al_2O_3 blocking oxide layer was subsequently deposited at 340 °C using atomic layer deposition (ALD) with trimethylaluminum (TMA, $\text{Al}(\text{CH}_3)_3$) as the precursor. A 50 nm-thick TiN metal gate was then formed using metal-organic chemical vapor deposition (MOCVD) at 450 °C and 1.5 Torr. To improve adhesion between the TiN metal gate and the photoresist, a 25 nm-thick SiO_2 hard mask was deposited at 380 °C using plasma-enhanced CVD (PECVD) with 700 sccm of tetraethoxysilane (TEOS, $\text{Si}(\text{OC}_2\text{H}_5)_4$) and 700 sccm of O_2 . The hard mask was patterned by wet etching using a buffered oxide etchant (BOE, $\text{NH}_4\text{F}:\text{HF} = 50:1$), and the TiN gate was patterned through wet etching in a DI water: H_2O_2 (4:1) solution at 60 °C. Self-aligned ion implantation was performed using As^+ ions at a dose of $2 \times 10^{15} \text{ cm}^{-2}$ and an energy of 40 keV to

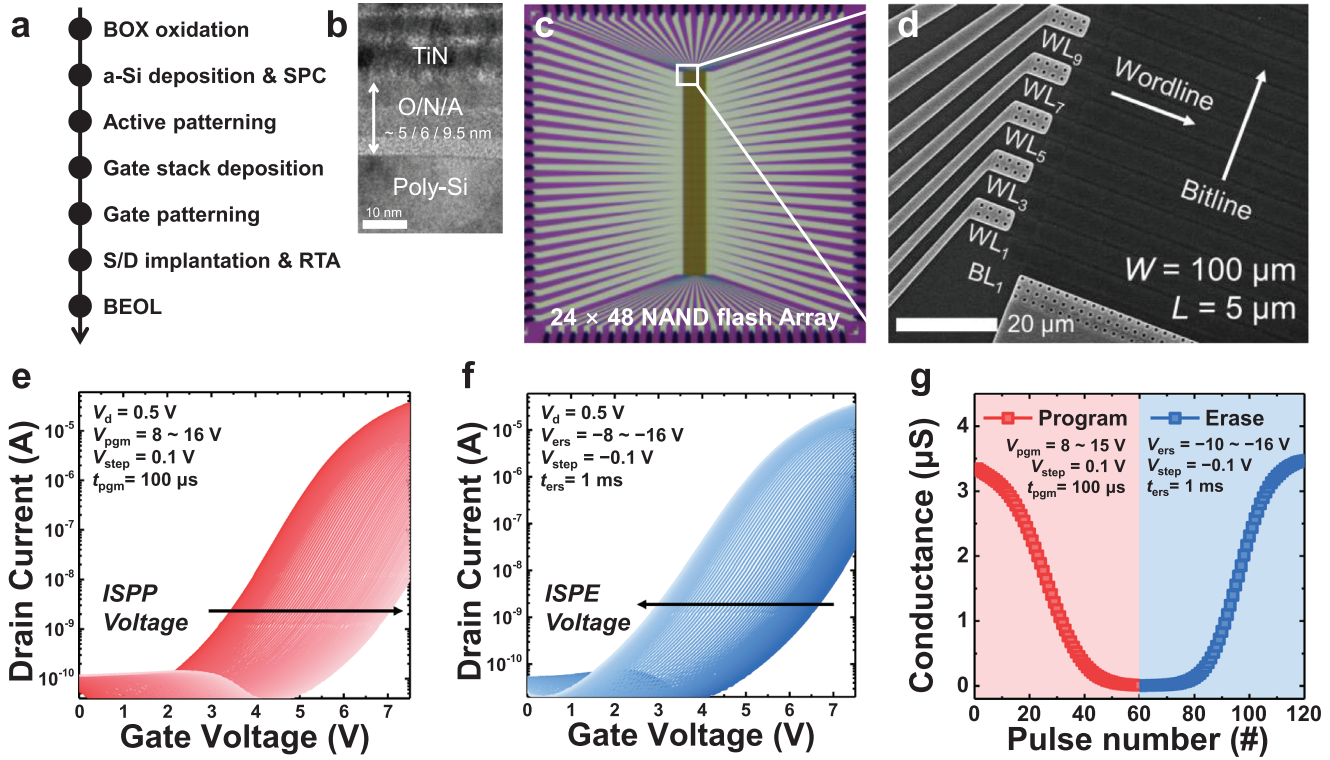


FIGURE 1 | Charge trapping NAND flash array. (a) Process flow of NAND flash array. (b) Cross-sectional TEM image of the TANOS gate stack. (c) Optical microscope image of the fabricated 48×24 NAND flash array. (d) SEM image of a fragment of the fabricated 48×24 NAND flash array. (e,f) Measured I_d - V_g transfer curves under ISPP and ISPE operations. (g) Conductance modulation in response to ISPP and ISPE pulses.

form the n^+ source and drain regions. The implanted dopants were activated via rapid thermal annealing (RTA) at 950°C for 10 s. Finally, the back-end-of-line (BEOL) process was carried out, including deposition of the interlayer dielectric (ILD) using PECVD, formation of contact holes, and metallization.

Figure 1b presents a cross-sectional transmission electron microscopy (TEM) image of the fabricated gate stack based on TANOS ($\text{TiN}/\text{Al}_2\text{O}_3/\text{Si}_3\text{N}_4/\text{SiO}_2/\text{Si}$) flash cell stack. The corresponding energy-dispersive spectroscopy (EDS) elemental mapping is shown in Figure S1, confirming the presence of well-defined and uniformly stacked layers. Figure 1c displays an optical microscopy image of the fabricated 48 (WL) $\times$ 24 (BL) NAND flash array, while Figure 1d shows a scanning electron microscopy (SEM) image of a part of the array. Each device within the array has a channel width of $100\ \mu\text{m}$ and a length of $5\ \mu\text{m}$ ($W/L = 100/5\ \mu\text{m}$). The electrical characteristics of the device are shown in Figure 1e, including the drain current–gate voltage (I_d - V_g) transfer curves obtained using the incremental step pulse program (ISPP) method. A drain voltage (V_d) of $0.5\ \text{V}$ was applied, while the program voltage (V_{pgm}) was increased from 8 to $16\ \text{V}$ in $0.1\ \text{V}$ steps with a pulse width (t_{pgm}) of $100\ \mu\text{s}$. Figure 1f shows the corresponding transfer curves obtained during incremental step pulse erase (ISPE) method. The device was initially programmed at $16\ \text{V}$ for $100\ \mu\text{s}$ and then erased by applying negative pulses ranging from -8 to $-16\ \text{V}$ in $-0.1\ \text{V}$ steps, each with a pulse width (t_{ers}) of $1\ \text{ms}$. The transfer were curves measured with constant read bias V_d of $0.5\ \text{V}$. In both ISPP and ISPE measurements, the gradual modulation of the V_{th} can be clearly observed, confirming the stable program (PGM) and erase (ERS) operations. Figure 1g

summarizes the device conductance extracted from both ISPP and ISPE measurements, where conductance (G) was calculated from measured I_d at a fixed gate voltage. The results clearly show a monotonic modulation of conductance in response to the applied PGM and ERS pulses.

2.2 | Physical Unclonable Function Based on 3D-NAND Flash Array Structure

Figure 2a illustrates a 3D schematic of the proposed PUF system based on a 3D-NAND flash array structure composed of i SSLs, j BLs, and k WLs. Each WL layer is configured as an independent PUF chip, and the desired chip is activated by applying a gate WL voltage (V_{WL}) to the corresponding WL layer. Meanwhile, unselected layers are biased with a pass voltage (V_{PASS}) to prevent IR drop and interference with selected chip operations. The challenge is defined by the selection and biasing of specific SSLs and BLs, where each bit of the challenge corresponds to either activating (for bit “1”) or floating (for bit “0”) a given line. SSLs are selected from 2^i possible combinations by applying a string select voltage (V_{SSL}) to each active SSL. The j BLs are divided into two groups using $jC_{j/2}$ combinations, and each group is sequentially biased with a BL voltage (V_{BL}) to perform a current comparison operation. During a single CRP extraction cycle, the selected SSLs and the first group of BLs are activated and biased with V_{SSL} and V_{BL} , respectively. This leads the BL current from all the selected cells to flow simultaneously toward the BLs, and the resulting summed BL current is sensed. In the subsequent cycle, the same procedure is repeated for the second BL group. A 1-bit response is

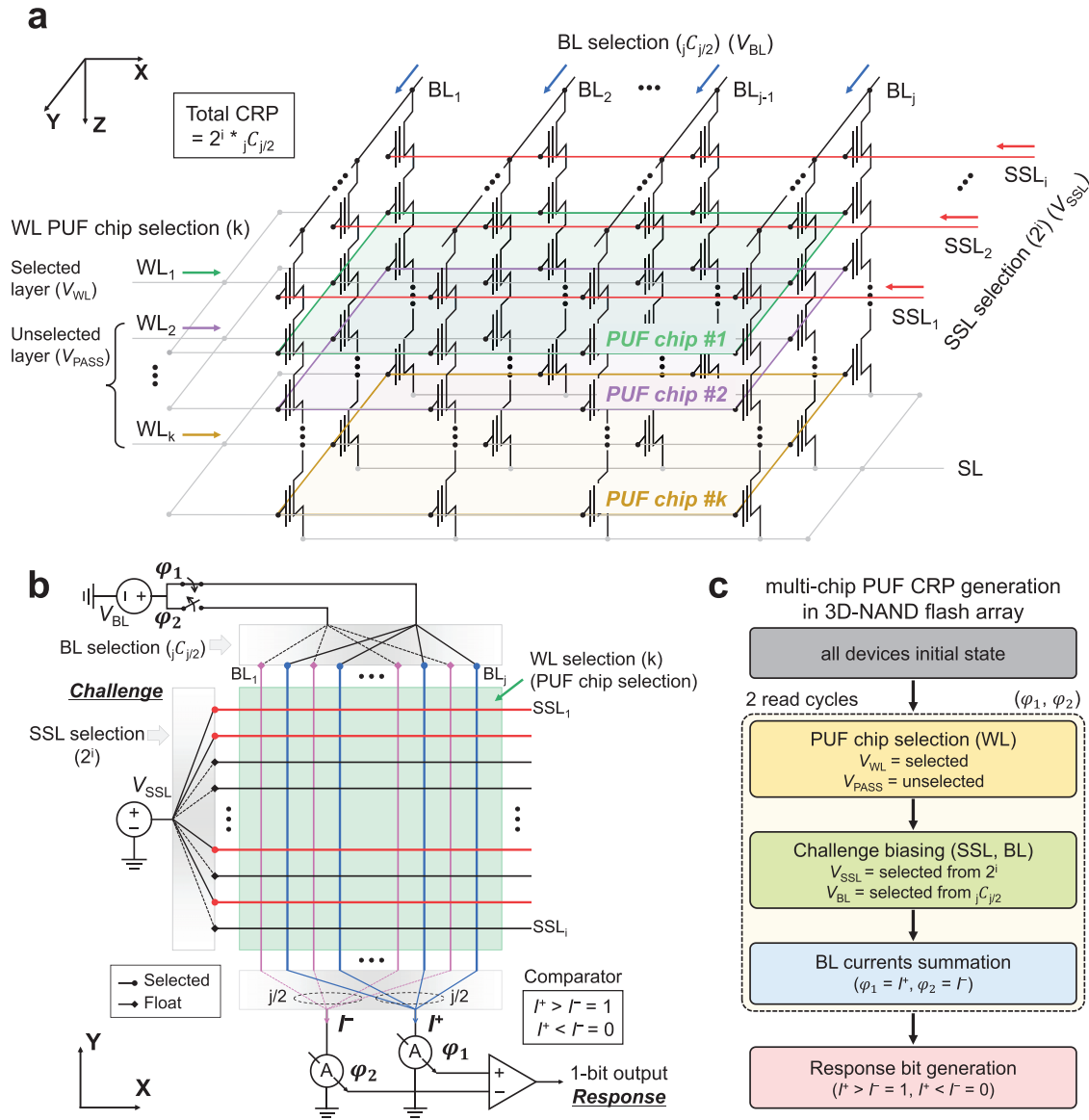


FIGURE 2 | Multi PUF chip implementation in 3D-NAND flash. (a) Schematic view of PUF system based on a 3D-NAND flash array. (b) x-y plane schematic of PUF system. WLS are used to select individual PUF chips, challenges are applied through SSLs and BLs, and a 1-bit response is generated by comparing the summed currents from two read cycles. (c) Flowchart of the CRP generation process for the multi-chip PUF based on a 3D-NAND flash array.

then generated by comparing the two measured BL currents. As a result, each WL layer in the array acts as an independent PUF chip, capable of generating a total of $2^i \times \binom{j}{j/2}$ CRPs, and an array consisting with k WLS provides k independent PUF chips.

Here, the proposed scheme is demonstrated with a 3D array configuration constructed from eight 48×24 planar NAND flash arrays, effectively forming a 48 (WL) $\times$ 24 (BL) $\times$ 8 (SSL) 3D-NAND flash structure. This configuration provides 48 independent PUF chips, each supporting approximately $\sim 6 \times 10^8$ CRPs based on $2^8 \times {}_{24}C_{12}$. Since current commercial 3D vertical-NAND (V-NAND) flash memory typically contains over 1 Tb of cells, including more than 16 000 BLs (corresponding to page size), over 1000 SSLs and more than 64 WL layers (corresponding to pages per block), the number of CRPs can be

exponentially increased accordingly [56–59]. Therefore, this PUF design has the potential to support an exponentially larger CRP space, depending on the number of SSLs and BLs. Moreover, over 64 independent PUF chips can be integrated into a single array, enabling scalable multi-chip implementation and supporting strong PUF characteristics high resistance to adversarial attacks due to the large CRP space.

Figure 2b illustrates the operation schematic of a single selected WL layer, which serves as one PUF chip, in the x-y plane of the 3D-NAND flash array. As described earlier, the challenge bits determine which SSLs and BLs are selected and biased accordingly. During the first read cycle (ϕ_1), the summed BL current from the first selected group is denoted as I^+ . In the second read cycle (ϕ_2), the summed BL current from the remaining group is denoted as I^- . A 1-bit response is generated by comparing the

two currents, resulting in response bit “1” when $I^+ > I^-$ and “0” when $I^+ < I^-$. This operation is similar to cell selection in a conventional $i \times j$ crossbar array composed of i rows and j columns. Consequently, the CRP generation using a 3D-NAND flash PUF chip is demonstrated using an 8×24 array, where a 32-bit challenge input produces a single-bit response output. Figure 2c presents a flowchart summarizing the CRP generation process of the proposed PUF system. Since this method utilizes intrinsic variations of the flash memory devices, no additional write operation is required prior to CRP generation. All cells remain in their initial state, and only two read cycles (ϕ_1, ϕ_2) are performed to generate each CRP. Although this work conceptually describes a simple current comparison for response generation, the can be practically implemented using two identical passive capacitors that integrate the accumulated summed string current from each read cycle, converting it into voltage. The resulting output voltages can then be compared using a voltage comparator. Alternatively, the summed BL currents can be converted and amplified into voltage through a transimpedance amplifier (TIA), allowing precise current comparison for reliable response generation [54, 55]. In commercial 3D-NAND flash arrays, the read latency is typically on the order of a few to several tens of microseconds per wordline. Accordingly, the presented CRP generation method with two consecutive read cycles exhibits a total latency on the order of several tens of microseconds [60–62]. This read-only operation represents a significant advantage in terms of both delay and power consumption compared to other flash memory-based PUFs that rely on repeated P/E operations for bit initialization or balancing. Since this method generates CRPs directly from intrinsic current variations, P/E operations are not required during normal authentication, and may only be involved in limited stages such as initial enrollment. Furthermore, the resulting throughput readily satisfies the timing requirements of practical authentication protocols in IoT and related applications, which typically incur delays on the order of several to tens of milliseconds [63–65]. Therefore, this approach demonstrates strong potential for real-world deployment, providing a large number of random CRPs in real time for energy-efficient authentication.

Figure 3a shows the I_{BL} - V_{WL} transfer curves measured from all 1,152 cells in the fabricated 48×24 NAND flash array. A V_{BL} of 0.5 V and a V_{PASS} of 5.5 V were applied, and the inset shows the same results in linear scale. Although all the cells were in their initial state, a noticeable spread in I_{BL} can be observed, indicating intrinsic device-to-device variation. As discussed earlier, such variation arises from process-induced fluctuations in physical parameters, such as cell width and length, random dopant fluctuation, and interface traps, all of which can affect the threshold voltage and string current. Furthermore, since the fabricated array employs a poly-Si channel formed by crystallizing amorphous silicon, additional variability can arise from material imperfections, such as the random size, orientation, and spatial distribution of grain boundaries, as well as trap states within them. These combined sources of intrinsic variation result in a unique and unclonable current distribution across the array, effectively forming a physical fingerprint. In addition, in practical 3D-NAND flash arrays, further variations in read operation and current arise from interlayer coupling-induced interference and IR drop along the vertically stacked channel resistance [59, 66]. From the perspective of the PUF system, these effects can

broaden the distribution and can be exploited as an additional entropy source, thereby contributing to the generation of more randomized CRPs. While the reduction of on-current due to IR drop inevitably can decrease the sensing margin, which may lead to reliability degradation, the incorporation of supplementary error correction codes (ECC) can be adopted to mitigate this drawback [66]. Figure 3b shows the distribution of sampled I_{BL} at a fixed V_{WL} ($V_{WL} = 4.5$ V) from four of the eight fabricated arrays. The remaining four distributions are provided in Figure S2. Each array exhibits a Gaussian distribution with an average current of approximately 18 nA. At the same time, the distributions are distinguishable from one another, demonstrating unique current profiles for each array. This consistency in mean current reflects that all arrays were fabricated on the same wafer under identical process conditions. However, the observed variations, both across arrays and among cells within each array, are due to inherent fluctuations introduced during fabrication. This confirms that the observed variability can serve as a random entropy source for PUF implementation.

To validate the proposed PUF using the measured variations, the eight fabricated 48×24 planar arrays were combined to form a $48 \times 24 \times 8$ 3D-NAND flash array. Figure 3c displays a BL current colormap from one selected PUF chip, activated by applying $V_{WL} = 4.5$ V to a specific WL. Due to the Gaussian-distributed current characteristics observed across all eight arrays, each sharing a similar mean values, the selected chip exhibits a uniform yet randomly distributed BL current distributions across its 8×24 array. Figure 3d shows the distribution of differential summed current ($I^+ - I^-$), measured from 100 000 randomly generated challenges through the two-cycle readout. Here, I^+ and I^- correspond to the total BL current from the first and second read cycles, respectively. A response bit is assigned as “1” if this differential current is greater than zero, and “0” otherwise. The distribution of these differential current values follows a Gaussian distribution, confirming that the presented selection and readout preserves the intrinsic randomness of the device-level variations. Accordingly, the generated 1-bit responses inherit this intrinsic randomness, validating their suitability for robust PUF operations.

2.3 | PUF Performance Metrics and Reliability

The performance of the 3D-NAND flash-based PUF system was evaluated based on standard PUF metrics, using CRPs generated at $V_{WL} = 4.5$ V and 25°C . Uniformity (UF) represents the balance between “0” and “1” values in the response bits from a single PUF chip and ideal value is 50%. Figure 4a presents the distribution of Hamming weights for response bits generated from 1M random challenges. The responses were grouped into 1000 packets of 1000 bits each, and the Hamming weight was calculated for each packet. The resulting distribution of UF shows an average of approximately 49.97%, indicating excellent uniformity close to the ideal value of 50%. Diffuseness (DF) quantifies the degree of variation between response bits generated from different challenges applied to the same PUF chip. It is evaluated by measuring the Hamming distance (HD) between response bit packets, and the ideal value is 50%. Figure 4b shows the distribution of fractional HDs calculated from 1000 randomly generated 64-bit response keys. A total of 499, 500 pairwise comparisons ($_{1000}C_2$)

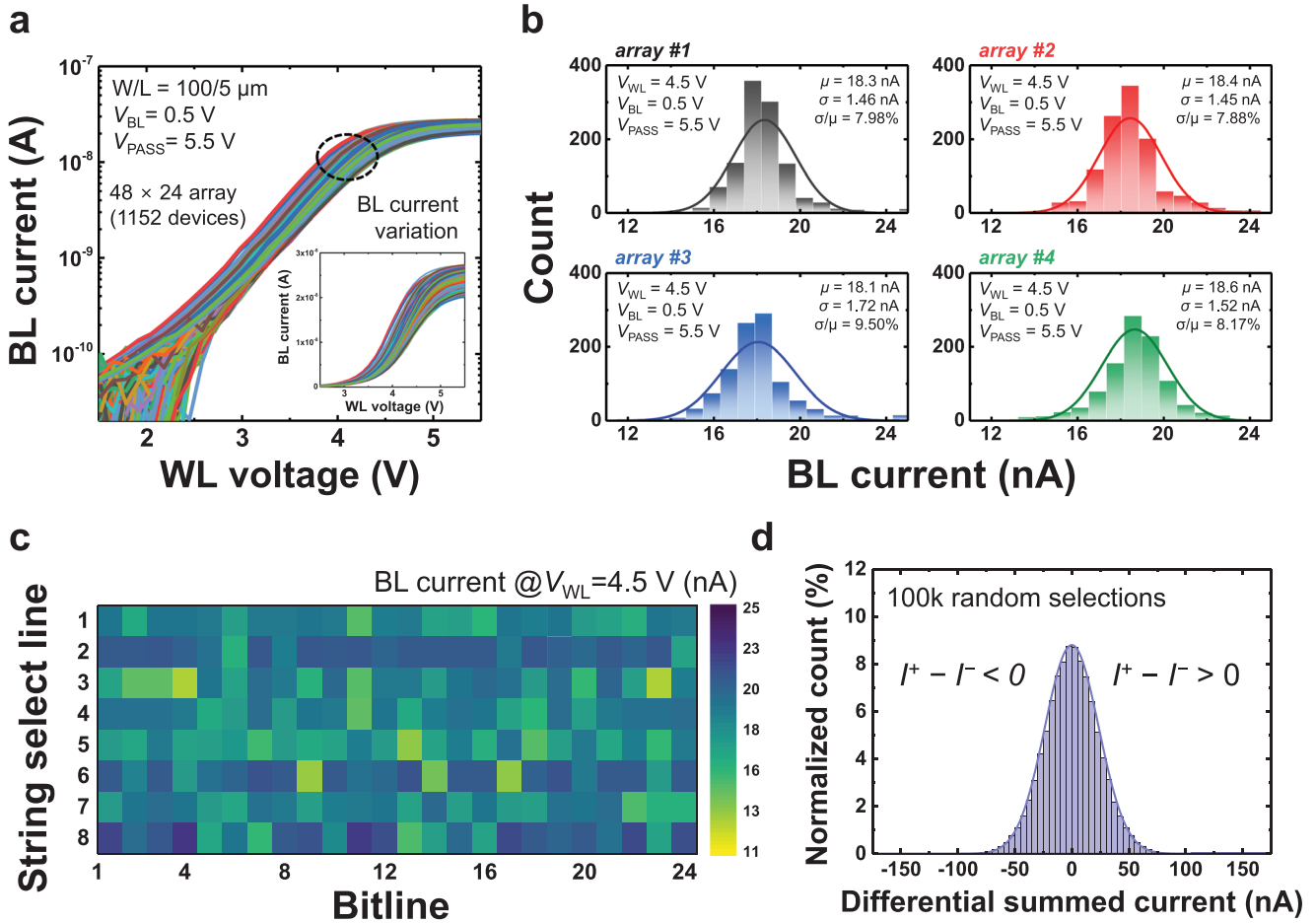


FIGURE 3 | Response generation with a string current variation. (a) Measured I_{BL} - V_{WL} characteristics of 1,152 devices in the fabricated 48×24 planar NAND flash array in the initial state. The inset shows the data in linear scale. (b) BL current distributions from four arrays. (c) Color map of the BL current distribution in a 3D-NAND flash array constructed using measured BL currents from eight 48×24 arrays. (d) Distribution of differential summed currents from 100 000 random selections using the proposed CRP generation method.

were performed, resulting in an average DF of approximately 50.01%, closely aligned with the ideal value. Uniqueness (UQ) evaluates how differently two distinct PUF chips respond to the same input challenge, which corresponds to inter-chip HD. An ideal uniqueness value is 50%. To assess this, we applied identical challenges to all 48 PUF chips in the demonstrated $48 \times 24 \times 8$ array structure. Figure 4c presents the distribution of fractional HDs computed from 1,000 64-bit response keys per chip under fixed challenge conditions. In total, 1,128,000 pairwise comparisons ($1000 \times {}_{48}C_2$) were performed, each representing the HD between two 64-bit response keys generated by different chips under the same challenge. The average UQ was measured to be approximately 50.07%, indicating that the PUF system generates sufficiently uncorrelated responses across 48 PUF chips.

Figure 4d shows the speckle pattern of 1000 64-bit response keys obtained from a single selected PUF chip. The pattern demonstrates that the response bits are uniformly distributed without noticeable bias. To further evaluate the bit-level correlation, an autocorrelation function (ACF) test was performed using 100,000 response bits. As shown in Figure 4e, the autocorrelation values fall within the 95% confidence interval, confirming the absence of linear correlation in the extracted responses. In addition, the

randomness of the response bits was evaluated using the National Institute of Standards and Technology (NIST) SP 800-22 test suite. Figure 4f shows the results of the 16 NIST tests conducted on 1M response bits, all of which were successfully passed, validating the high quality of randomness from the response bits generated by the PUF scheme.

A PUF system must reliably reproduce consistent response bits under varying operating conditions when the same challenge is applied. This property, referred to as reliability, is quantified by the bit error rate (BER), which represents the HD between the originally enrolled response and the one regenerated under different conditions. Ideally, the BER should be 0%. To further evaluate the robustness of the proposed PUF system, we examined the BER under temperature variations. Figure 5a presents box plots of the BER extracted from 1000 CRPs per PUF chip across 48 chips, both without error correction and with temporal majority voting (TMV) applied. Reference CRPs were enrolled at room temperature (25°C) and compared with responses obtained at elevated temperatures of 55°C and 85°C under identical challenges. The average BER was 0.74% at room temperature, increasing to 9.6% at 85°C . To mitigate these errors, we applied TMV3, in which the final response bit is determined by majority voting over three repeated read operations. With

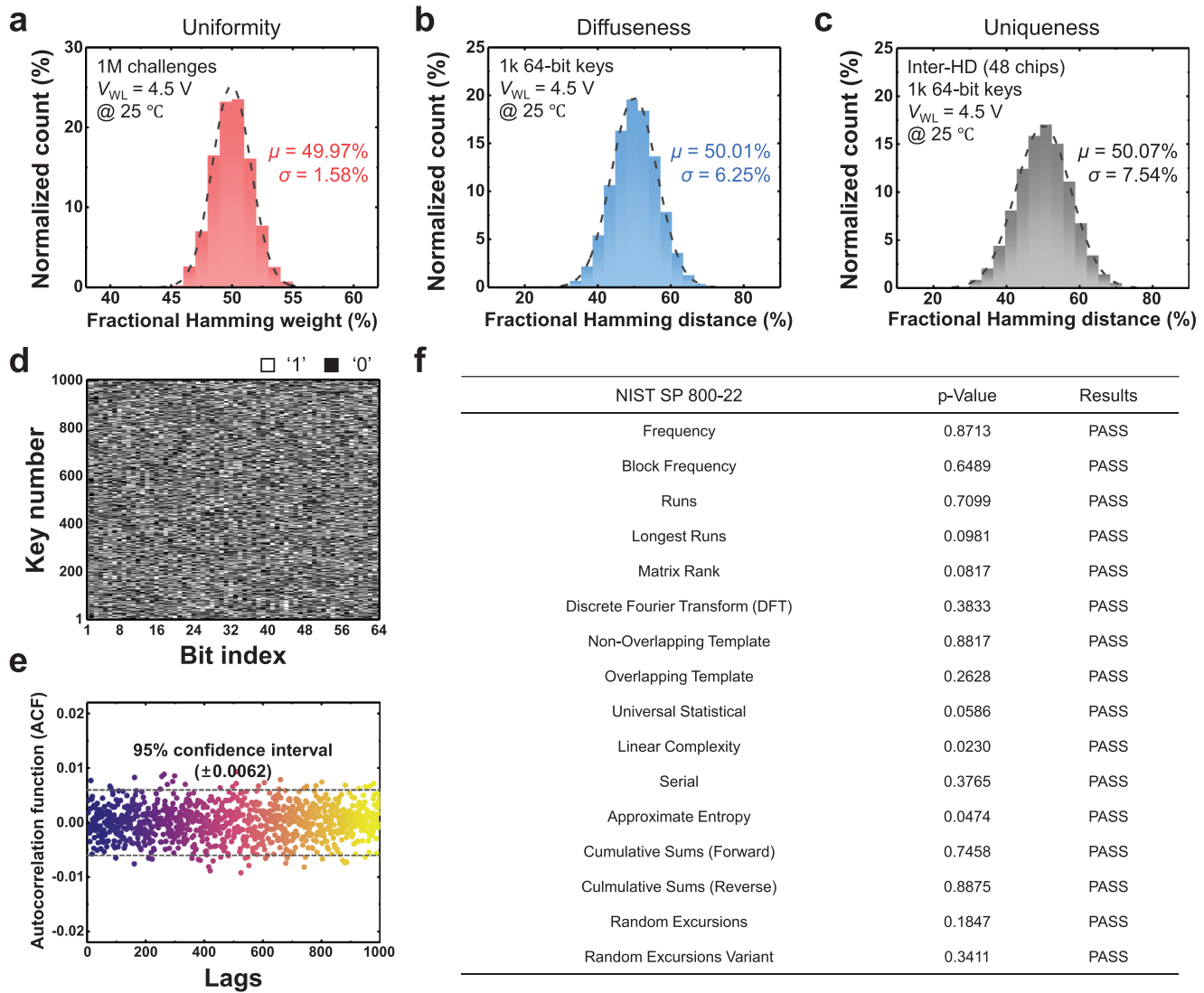


FIGURE 4 | PUF metrics. (a) Distribution of fractional Hamming weight of 1-bit responses from 1M challenges (uniformity). (b) Distribution of fractional Hamming distance of 1k 64-bit response keys (diffuseness). (c) Distribution of fractional Hamming distance of 1k 64-bit response keys across 48 PUF chips (uniqueness). (d) Speckle pattern of 1k 64-bit response keys. (e) Autocorrelation test results for 100k 1-bit responses. (f) NIST SP 800-22 test results for 1M 1-bit responses.

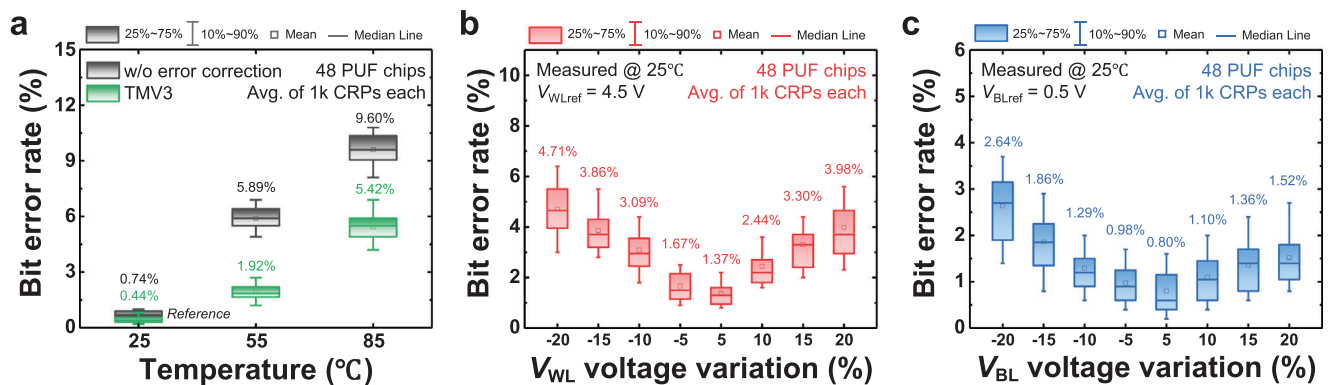


FIGURE 5 | Verification of reliability. (a) Bit error rate extracted from 1000 CRPs across 48 PUF chips under temperature variations. Bit error rate under voltage bias variations of (b) WL voltage and (c) BL voltage.

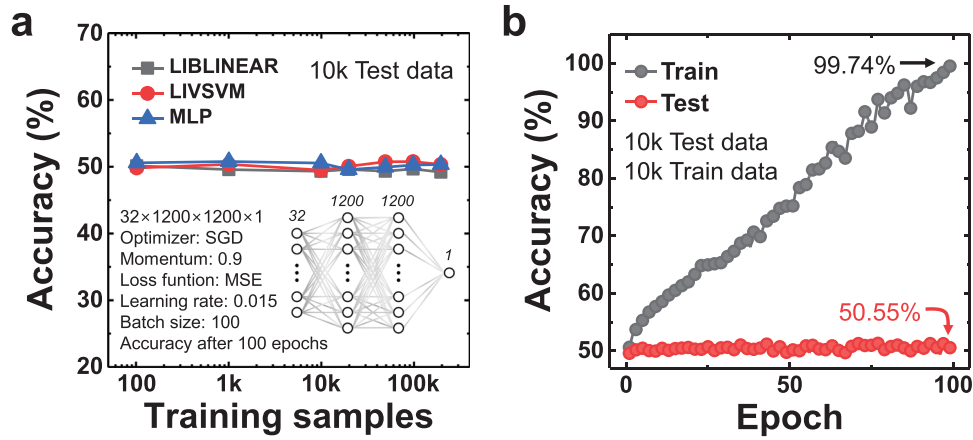


FIGURE 6 | Machine learning attack. (a) Prediction accuracy of three ML attack models as a function of the number of training samples. (b) Prediction accuracy of training and test data over 100 epochs using a 32-1200-1200-1 MLP model.

TMV3, the average BER at 85°C was reduced to 5.42%, which can be further reduced by incorporating well-established techniques such as ECC or helper-data algorithms [28, 54, 67, 68]. In addition to temperature dependence, reliability was also evaluated under bias voltage variations of WLs and BLs. The reference CRPs were measured at V_{WL} of 4.5 V ($V_{WLref} = 4.5$ V) and V_{BL} of 0.5 V ($V_{BLref} = 0.5$ V) at room temperature. For each condition, the applied voltages were varied from -20% to $+20\%$ in steps of $\pm 5\%$. Figure 5b shows box plots of the measured BER from 1,000 CRPs per PUF chip under V_{WL} variations, across all 48 PUF chips, with the corresponding mean values. Figure 5c presents the results under V_{BL} variations using the same procedure. At the maximum fluctuation of -20% , average BERs of 4.71% and 2.64% were measured for WLs and BLs variations, respectively, confirming the robustness of the PUF system against operating voltage fluctuations.

We further evaluated the resistance of the proposed PUF system to ML attacks using the generated CRPs. In addition to the intrinsic variations of each PUF chip, modeling resilience was enhanced by adopting a multi-chip response-combination. In this method, multiple chips generate independent response bits under the same input challenge, and the outputs are combined through a logical XOR operation to produce the final response. Among the 48 PUF chips demonstrated in this work, three were selected for this, and their outputs were aggregated to form each response bit. ML attack simulations were performed using three representative models: LIBLINEAR, LIBSVM, and a multi-layer perceptron (MLP). The detailed parameters for each attack are provided in the Experimental Section. Figure 6a shows the prediction accuracy of the models as a function of the number of training samples, with 10,000 unseen CRPs used as the test set. For the MLP, a 32-1200-1200-1 network was trained for 100 epochs, and the prediction accuracy for the test data was evaluated. Even with 200k training samples, all three models converged to $\sim 50\%$ accuracy, indicating that unseen CRPs could not be effectively learned or predicted. Figure 6b further illustrates the MLP training results: using two non-overlapping CRP sets of 10k each for training and testing, the model reached 99.74% accuracy on the training data after 100 epochs, while the accuracy on the test data remained at 50.55%. These results confirm that the CRPs generated by the proposed 3D-NAND flash-based PUF

system, including those derived from the multi-chip combination, exhibit strong resistance to ML attacks and maintain high randomness.

Finally, Table S1 compares our approach with recent NAND flash-based PUFs in terms of entropy source, demonstration complexity, multi-chip capability, and key performance metrics. Unlike prior program- or erase-based s, the proposed method directly exploits intrinsic string-current variations and uniquely demonstrates scalable multi-chip PUF operation, enabling a very large CRP space and stable performance under elevated temperatures. These results highlight not only the robustness of the proposed but also its strong potential for deployment in secure, reliable, and scalable hardware security applications based on 3D-NAND flash technology.

3 | Conclusion

In this study, we presented a PUF system based on a 3D-NAND flash array that exploits intrinsic string current variations. By utilizing the high-density configuration of 3D-NAND flash, the proposed enables the generation of a large number of CRPs and realizes a multi-chip PUF implementation, in which each WL layer functions as an independent chip. A 48×24 planar NAND flash array based on CTF memory was fabricated, and its electrical characteristics were validated. The measured results confirmed notable variations in string current both within and across arrays, which were due to process-induced variability and randomly distributed grain boundaries in the poly-Si channel. By combining eight arrays, a $48 \times 24 \times 8$ 3D-NAND configuration was implemented, and the CRPs generated from the proposed exhibited near-ideal values across key randomness metrics, including UF, DF, and UQ. Additionally, the extracted response bits passed both the ACF and NIST tests, confirming their high randomness and independence. To further assess robustness under environmental variations, the BER was evaluated under elevated temperatures and fluctuations in WL and BL bias voltages. With the application of TMV3 under temperature variation, the BER at 85°C was reduced to 5.42%. Under $\pm 20\%$ bias voltage variation, the BER remained below 4.71% in the worst case without TMV3, demonstrating stable and reliable PUF

operation. Finally, ML attack simulations were performed on the CRPs generated by the proposed PUF system using three ML models. The results indicated, even after training, unseen CRP datasets remained unpredictable, confirming strong resilience and randomness. These findings demonstrate the robustness and scalability of the proposed 3D-NAND flash-based PUF, indicating its strong potential for deployment in practical hardware security systems.

4 | Experimental Section

4.1 | Electrical Measurements

All the electrical characteristics of the fabricated NAND flash device were measured using the Keysight B1500A semiconductor parameter analyzer, with source measure unit (SMU), and semiconductor pulse generator unit (SPGU). To characterize the 48×24 planar NAND flash array, the Keysight E5250A low-leakage switch matrix and 96-channel probe card were utilized. For randomness verification of the PUF, the BL current of each device in the entire array was sampled under fixed WL and BL voltages.

4.2 | ML Attack Simulations

To evaluate the robustness of the generated CRPs against ML attacks, simulations were conducted in Python using the scikit-learn library. Three representative models were considered: LIBLINEAR, LIBSVM, and MLP. The MLP architecture consisted of two hidden layers with a 32-1200-1200-1 configuration, where the 32-bit challenge input encodes the selection of 8 SSLs and 24 BLs, and the output predicts the final response bit (0 or 1). Training was performed using stochastic gradient descent (SGD) with a momentum of 0.9, mean square error (MSE) loss, a learning rate of 0.015, and a batch size of 100. To enhance modeling resistance, responses from three independent PUF chips subjected to the same challenge were combined into a single 1-bit output using a logical XOR operation, and this aggregated response was employed for ML attack prediction.

Acknowledgements

This work was supported in part by the NRF funded by the Korean government (RS-2023-00260527, 15%, RS-2025-07402970, 50%), in part by the IITP (Institute for Information & Communications Technology Planning & Evaluation) grant funded by the Korean government (RS-2021-II212052, 10%, RS-2024-00506767, 20%), in part by KIAT grant funded by the Korean government (RS-2024-00410066, HRD Program for Industrial Innovation, 5%), and in part by the Brain Korea 21 Four Program.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," *Computer Networks* 54 (2010): 2787–2805.
2. J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications," *IEEE Internet of Things Journal* 4 (2017): 1125–1142.
3. Z. Zhao, K. M. Barijough, and A. Gerstlauer, "DeepThings: Distributed Adaptive Deep Learning Inference on Resource-Constrained IoT Edge Clusters," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37 (2018): 2348–2359.
4. F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine Learning in IoT Security: Current Solutions and Future Challenges," *IEEE Communications Surveys & Tutorials* 22 (2020): 1686–1721.
5. R. T. Tiburski, C. R. Moratelli, S. F. Johann, et al., "Lightweight Security Architecture Based on Embedded Virtualization and Trust Mechanisms for IoT Edge Devices," *IEEE Communications Magazine* 57 (2019): 67–73.
6. J. Park, H. Kim, and H. Kim, "Bias-Independent True Random Number Generator Circuit using Memristor Noise Signals as Entropy Source," *Advanced Intelligent Systems* 7 (2025): 2400648.
7. R. Carboni and D. Ielmini, "Stochastic Memory Devices for Security and Computing," *Advanced Electronic Materials* 5 (2019): 1900198.
8. H. Jiang, D. Belkin, S. E. Savel'ev, et al., "A Novel True Random Number Generator Based on a Stochastic Diffusive Memristor," *Nature Communications* 8 (2017): 882.
9. M. S. Song, T.-H. Kim, H. Hwang, S. Ahn, H. Nili, and H. Kim, "Optimization of Random Telegraph Noise Characteristics in Memristor for True Random Number Generator," *Advanced Intelligent Systems* 5 (2023): 2200358.
10. G. Kim, J. H. In, Y. S. Kim, et al., "Self-Clocking Fast and Variation Tolerant True Random Number Generator Based on a Stochastic Mott Memristor," *Nature Communications* 12 (2021): 2906.
11. H. Hwang, M. S. Song, S. Youn, and H. Kim, "True Random Number Generator Using Memcapacitor With Charge Trapping Layer," *IEEE Electron Device Letters* 45 (2024): 1464–1467.
12. K. S. Woo, J. Han, S. Yi, et al., "Tunable Stochastic Memristors for Energy-Efficient Encryption and Computing," *Nature Communications* 15 (2024): 3245.
13. D. Yu, S. Ahn, S. Youn, J. Park, and H. Kim, "True Random Number Generator Using Stochastic Noise Signal of Memristor With Variation Tolerance," *Chaos, Solitons & Fractals* 189 (2024): 115708.
14. C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, "Physical Unclonable Functions and Applications: A Tutorial," *Proceedings of the IEEE* 102 (2014): 1126–1141.
15. Y. Gao, S. F. Al-Sarawi, and D. Abbott, "Physical Unclonable Functions," *Nature Electronics* 3 (2020): 81–91.
16. B. Gassend, D. Clarke, M. Van Dijk, and S. Devadas, "Silicon Physical Random Functions," in *Proceedings of the 9th ACM Conference on Computer and Communications Security* (2002), 148–160.
17. G. E. Suh and S. Devadas, "Physical Unclonable Functions For Device Authentication and Secret Key Generation," in *Proceedings of the 44th Design Automation Conference* (2007): 9–14.
18. A. Maiti and P. Schaumont, "Improved Ring Oscillator PUF: An FPGA-Friendly Secure Primitive," *Journal of Cryptology* 24 (2011): 375–397.
19. D. P. Sahoo, D. Mukhopadhyay, R. S. Chakraborty, and P. H. Nguyen, "A Multiplexer-Based Arbiter PUF Composition with Enhanced Reliability and Security," *IEEE Transactions on Computers* 67 (2017): 403–417.
20. K. Liu, Y. Min, X. Yang, H. Sun, and H. Shinohara, "A 373-F² 0.21%-Native-BER EE SRAM Physically Unclonable Function With 2-D

- Power-Gated Bit Cells and V_{SS} Bias-Based Dark-Bit Detection,” *IEEE Journal of Solid-State Circuits* 55 (2020): 1719–1732.
21. H. Nili, G. C. Adam, B. Hoskins, et al., “Hardware-Intrinsic Security Primitives Enabled by Analogue State and Nonlinear Conductance Variations in Integrated Memristors,” *Nature Electronics* 1 (2018): 197–202.
22. A. Chen, “Utilizing the Variability of Resistive Random Access Memory to Implement Reconfigurable Physical Unclonable Functions,” *IEEE Electron Device Letters* 36 (2014): 138–140.
23. R. Liu, H. Wu, Y. Pang, H. Qian, and S. Yu, “Experimental Characterization of Physical Unclonable Function Based on 1 kb Resistive Random Access Memory Arrays,” *IEEE Electron Device Letters* 36 (2015): 1380–1383.
24. H. Jiang, C. Li, R. Zhang, et al., “A Provable Key Destruction Based on Memristive Crossbar Arrays,” *Nature Electronics* 1 (2018): 548–554.
25. J. Park, T.-H. Kim, S. Kim, G. H. Lee, H. Nili, and H. Kim, “Conduction Mechanism Effect on Physical Unclonable Function Using Al_2O_3/TiO_x Memristors,” *Chaos, Solitons & Fractals* 152 (2021): 111388.
26. M. Mahmoodi, H. Nili, Z. Fahimi, S. Larimian, H. Kim, and D. Strukov, “Ultra-Low Power Physical Unclonable Function With Nonlinear Fixed-Resistance Crossbar Circuits,” *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), 30.1. 1–30.1. 4.
27. J. Park, T.-H. Kim, S. Kim, et al., “Highly Reliable Physical Unclonable Functions Using Memristor Crossbar With Tunneling Conduction,” *2022 International Electron Devices Meeting (IEDM)* (IEEE, 2022), 18.3. 1–18.3. 4.
28. B. Gao, B. Lin, Y. Pang, et al., “Concealable Physically Unclonable Function Chip With a Memristor Array,” *Science Advances* 8 (2022): abn7753.
29. J. Yu, K. K. Min, Y. Kim, et al., “A Novel Physical Unclonable Function (PUF) Using 16×16 Pure- HfO_x Ferroelectric Tunnel Junction Array for Security Applications,” *Nanotechnology* 32 (2021): 485202.
30. S. Kim, K. Lee, M.-H. Oh, J.-H. Lee, B.-G. Park, and D. Kwon, “Physical Unclonable Functions Using Ferroelectric Tunnel Junctions,” *IEEE Electron Device Letters* 42 (2021): 816–819.
31. H. Shao, Y. Zhou, W. Huang, et al., “A Novel FeFET Array-Based PUF: Co-Optimization of Entropy Source and CRP Generation for Enhanced Robustness in IoT Security,” in *2023 International Electron Devices Meeting (IEDM)* (IEEE, 2023), 1–4.
32. Y. Miao, X. Zhou, Z. Wang, et al., “Electrically-Reconfigurable Extremely-High Density Physical Unclonable Cryptographic Keys Based on Aurivillius Ferroelectrics,” *Advanced Functional Materials* 34 (2024): 2314883.
33. T. Li, X. Guo, F. Müller, et al., “Demonstration of High-Reconfigurability and Low-Power Strong Physical Unclonable Function Empowered by FeFET Cycle-to-Cycle Variation and Charge-Domain Computing,” *Nature Communications* 16 (2025): 189.
34. P. Prabhu, A. Akel, L. M. Grupp, et al., “Extracting Device Fingerprints From Flash Memory by Exploiting Physical Variations,” *Proceedings of the 4th International Conference on Trust and Trustworthy Computing* (2011): 188–201.
35. Y. Wang, W. Yu, S. Wu, G. Malysa, G. E. Suh, and E. C. Kan, “Flash Memory For Ubiquitous Hardware Security Functions: True Random Number Generation and Device Fingerprints,” in *2012 IEEE Symposium on Security and Privacy* (IEEE, 2012): 33–47.
36. S. Q. Xu, W. Yu, G. E. Suh, and E. C. Kan, “Understanding Sources of Variations in Flash Memory for Physical Unclonable Functions,” in *2014 IEEE 6th International Memory Workshop (IMW)* (IEEE, 2014): 1–4.
37. M.-S. Kim, D.-I. Moon, S.-K. Yoo, S.-H. Lee, and Y.-K. Choi, “Investigation of Physically Unclonable Functions Using Flash Memory for Integrated Circuit Authentication,” *IEEE Transactions on Nanotechnology* 14 (2015): 384–389.
38. S. Sakib, A. Milenković, M. T. Rahman, and B. Ray, “An Aging-Resistant NAND Flash Memory Physical Unclonable Function,” *IEEE Transactions on Electron Devices* 67 (2020): 937–943.
39. L. T. Clark, J. Adams, and K. E. Holbert, “Reliable Techniques for Integrated Circuit Identification and True Random Number Generation Using 1.5-Transistor Flash Memory,” *Integration* 65 (2019): 263–272.
40. M. S. Song, S. Ahn, H. Hwang, and H. Kim, “Memcapacitor Crossbar Array With Charge Trapping Layer for Physical Unclonable Function in NAND Flash Architecture,” in *2023 Int. Electron Devices Meeting (IEDM)* (IEEE, 2023), 1–4.
41. S.-H. Park, R.-H. Koo, Y. Yang, J. Im, J. Ko, and J.-H. Lee, “Concealable Physical Unclonable Functions Using Vertical NAND Flash Memory,” *Nature Communications* 16 (2025): 5155.
42. Y.-L. Chou, C. C. Cheng, G. W. Wu, et al., “A Novel BTBT Erase-Based PUF Generation Method With High Throughput in a 3-D NAND Flash Memory,” *IEEE Transactions on Electron Devices* 72 (2025): 3528.
43. D. Resnati, G. Nicosia, G. M. Paolucci, A. Visconti, and C. M. Compagnoni, “Cycling-Induced Charge Trapping/Detrapping in Flash Memories—Part I: Experimental Evidence,” *IEEE Transactions on Electron Devices* 63 (2016): 4753–4760.
44. A. Fayrushin, K. Seol, J. Na, S. Hur, J. Choi, and K. Kim, “The New Program/Erase Cycling Degradation Mechanism of NAND Flash Memory Devices,” in *2009 IEEE Int. Electron Devices Meeting (IEDM)* (IEEE, 2009), 1–4.
45. A. Asenov, S. Kaya, and J. H. Davies, “Intrinsic Threshold Voltage Fluctuations in Decanano MOSFETs Due to Local Oxide Thickness Variations,” *IEEE Transactions on Electron Devices* 49 (2002): 112–119.
46. A. Asenov, S. Kaya, and A. R. Brown, “Intrinsic Parameter Fluctuations in Decananometer Mosfets Introduced by Gate Line Edge Roughness,” *IEEE Transactions on Electron Devices* 50 (2003): 1254–1260.
47. A. Spessot, C. M. Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, “Compact Modeling of Variability Effects in Nanoscale nand Flash Memories,” *IEEE Transactions on Electron Devices* 58 (2011): 2302–2309.
48. G. Roy, A. Ghetti, A. Benvenuti, A. Erlebach, and A. Asenov, “Comparative Simulation Study of the Different Sources of Statistical Variability in Contemporary Floating-Gate Nonvolatile Memory,” *IEEE Transactions on Electron Devices* 58 (2011): 4155–4163.
49. C.-W. Yang and P. Su, “Simulation and Investigation of Random Grain-Boundary-Induced Variabilities for Stackable NAND Flash Using 3-D Voronoi Grain Patterns,” *IEEE Transactions on Electron Devices* 61 (2014): 1211–1214.
50. R. Degraeve, S. Clima, V. Putcha, et al., “Statistical Poly-Si Grain Boundary Model With Discrete Charging Defects and its 2D and 3D Implementation for Vertical 3D NAND Channels,” *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), 5.6. 1–5.6. 4.
51. W.-J. Tsai, W. Lin, C. Cheng, et al., “Polycrystalline-silicon Channel Trap Induced Transient Read Instability in a 3D NAND Flash Cell String,” *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016): 11.3. 1–11.3. 4.
52. A. Subirats, A. Arreghini, E. Capogreco, et al., “Experimental and Theoretical Verification of Channel Conductivity Degradation due to Grain Boundaries and Defects in 3D NAND,” *2017 IEEE Int. Electron Devices Meeting (IEDM)* (IEEE, 2017), 21.2. 1–21.2. 4.
53. D. Resnati, A. Mannara, G. Nicosia, et al., “Characterization and Modeling of Temperature Effects in 3-D NAND Flash Arrays—Part I: Polysilicon-Induced Variability,” *IEEE Transactions on Electron Devices* 65 (2018): 3199–3206.
54. J. Lee, M. Kim, M. Jeong, G. Shin, and Y. Lee, “A 20F 2 /Bit Current-Integration-Based Differential nand-Structured PUF for Stable and V / T Variation-Tolerant Low-Cost IoT Security,” *IEEE Journal of Solid-State Circuits* 57 (2022): 2957–2968.

55. S. Sahay, M. Klachko, and D. Strukov, "Hardware Security Primitive Exploiting Intrinsic Variability in Analog Behavior of 3-D NAND Flash Memory Array," *IEEE Transactions on Electron Devices* 66 (2019): 2158–2164.
56. S. Lee, C. Kim, M. Kim, et al., "A 1Tb 4b/cell 64-Stacked-WL 3D NAND Flash Memory With 12MB/s Program Throughput," *2018 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2018): 340–342.
57. N. Shibata, K. Kanda, T. Shimizu, et al., "A 1.33-Tb 4-Bit/Cell 3-D Flash Memory on a 96-Word-Line-Layer Technology," *IEEE Journal of Solid-State Circuits* 55 (2019): 178–188.
58. J. H. Yuh, Y.-L. J. Li, H. Li, et al., "A 1-Tb 4-b/cell 4-Plane 162-Layer 3-D Flash Memory With 2.4-Gb/s IO Interface," *IEEE Journal of Solid-State Circuits* 58 (2022): 316–328.
59. A. Goda, "3-D NAND Technology Achievements and Future Scaling Perspectives," *IEEE Transactions on Electron Devices* 67 (2020): 1373–1381.
60. M. Kim, S. W. Yun, and J. Park, "A 1Tb 3b/cell 8th-Generation 3D-NAND Flash Memory With 164MB/s Write Throughput and a 2.4 Gb/s Interface," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2022), 136–137.
61. D. Kang, W. Jeong, C. Kim, et al., "256 Gb 3 b/Cell V-nand Flash Memory With 48 Stacked WL Layers," *IEEE Journal of Solid-State Circuits* 52 (2016): 210–217.
62. W. Cheong, C. Yoon, S. Woo, et al., "A Flash Memory Controller for 15μs Ultra-Low-Latency SSD Using High-Speed 3D NAND Flash With 3μs Read Time," in *2018 IEEE Int. Solid-State Circuits Conf. (ISSCC)* (IEEE, 2018), 338–340.
63. P. Gope and B. Sikdar, "Lightweight and Privacy-Preserving Two-Factor Authentication for IoT Devices," *IEEE Internet of Things Journal* 6 (2018): 580–589.
64. C.-M. Chen, B. Xiang, Y. Liu, and K.-H. Wang, "A Secure Authentication Protocol for Internet of Vehicles," *IEEE Access* 7 (2019): 12047–12057.
65. J. Ni, X. Lin, and X. S. Shen, "Efficient and Secure Service-Oriented Authentication Supporting Network Slicing for 5G-Enabled IoT," *IEEE Journal on Selected Areas in Communications* 36 (2018): 644–657.
66. S. K. Park and J. Moon, "Characterization of Inter-Cell Interference in 3D NAND Flash Memory," *IEEE Transactions on Circuits and Systems I: Regular Papers* 68 (2021): 1183–1192.
67. J. Delvaux, D. Gu, D. Schellekens, and I. Verbauwhede, "Helper Data Algorithms for PUF-Based Key Generation: Overview and Analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34 (2015): 889–902.
68. S. K. Mathew, S. K. Satpathy, M. A. Anders, et al., "16.2 A 0.19 pJ/b PVT-Variation-Tolerant Hybrid Physically Unclonable Function Circuit for 100% Stable Secure Key Generation in 22nm CMOS," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014), 278–279.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File: sml174218-sup-0001-SuppMat.docx.