



 Research Article

Who Wins First? A Six-Attack Machine-Learning Benchmark on OR-AND-XOR Arbiter PUFs, And A Methodological Warning About Comparing Them

Journal Website:
<http://sciencebring.com/index.php/ijasr>

Copyright: Original
content from this work
may be used under the
terms of the creative
commons attributes
4.0 licence.

Submission Date: February 12, 2026, **Accepted Date:** March 10, 2026,

Published Date: March 31, 2026

Crossref doi: <https://doi.org/10.37547/ijasr-06-03-08>

Ismoil Makhamatdjonov

Department of Computer Science, University of Sheffield, Uzbekistan

ABSTRACT

Published attacks on Physically Unclonable Functions (PUFs) usually test one or two attacker architectures against a single circuit family, leaving open a basic question: given the same PUF and CRP budget, which learning paradigm wins first, and why? This article benchmarks six attackers — logistic regression, least-squares regression, a multilayer perceptron, an LMN feature-lifting model, a denoising autoencoder, and a compact Transformer — against Arbiter OR-AND-XOR PUFs (AOX-PUFs) across three circuit families (aox-5, aox-6, aox-7; $n=64$) and twelve (x,y,z) mixes, with CRP budgets from 2,500 to ≈ 1.9 million on the University of Sheffield's Stanage HPC cluster. The hypothesis is that no single attacker dominates independent of circuit composition: linear attackers should suffice only while the decision boundary stays close to linear in the engineered challenge features, while non-linear attackers should become necessary once OR-AND-XOR composition pushes it away from linearity. The data confirm this. The MLP and autoencoder reach 98–99.5% accuracy fastest, by 25,000–75,000 CRPs on favourable mixes; logistic regression settles near 0.95 ± 0.02 ; LMN bifurcates sharply between near-chance (0.52–0.60) and near-ceiling (≈ 0.98) accuracy depending on mix; least-squares regression is the weakest linear baseline. A further finding is methodological: an under-trained Transformer can make an unrelated protocol-level defence look misleadingly effective, because its classical baseline never reaches its own ceiling. These results are benchmarked against the reliability-based, hybrid-PUF, and bias-exploiting attack literature, with a labelled illustrative calculation setting this study's own data against an independently reported deep-neural-network accuracy collapse at high XOR counts.

KEYWORDS

Physically unclonable function; Arbiter PUF; OR-AND-XOR PUF; machine-learning attack; reliability-based attack; linear additive delay model; benchmarking.

1. INTRODUCTION

Physically Unclonable Functions (PUFs) exploit uncontrollable manufacturing variation to derive device-specific challenge-response behaviour, and are widely proposed as low-cost primitives for authentication and key generation on resource-constrained hardware. Their practical security is set not by nominal non-linearity but by how well a motivated attacker, given enough data and the right model, can approximate the challenge-response function. This is the behaviour a modelling-attack benchmark is meant to characterise.

Most published evaluations report one attacker, occasionally two, against a given circuit — reasonable for showing that a PUF can or cannot be modelled, but silent on a question that matters for anyone selecting a defensive baseline: given the same PUF, CRPs, and budget, which learning paradigm wins first, and why? Different architectures encode different inductive biases — a linear decision boundary, a bottlenecked latent representation, a lifted monomial feature space, self-attention over challenge bits — and these interact with circuit structure in ways a single-attacker study cannot reveal.

The working hypothesis is that no single attacker family dominates independent of circuit composition for an Arbiter OR-AND-XOR PUF (AOX-PUF). Linear attackers (LR, LSR) should suffice only while the decision boundary stays close to linear in the engineered challenge features; non-linear attackers (MLP, AE, Transformer) and the feature-lifted LMN model

should outperform them once OR/AND/XOR composition pushes that boundary away from linearity, and this transition should not be uniform across the twelve mixes tested. [Section 2](#) develops this from the APUF's linear additive delay model; [Sections 3 and 4](#) confirm it, with LMN's sharp bifurcation the clearest evidence that the transition is itself mix-dependent rather than smooth.

The contribution is fourfold: a head-to-head ranking of six attacker architectures under matched conditions, extending the architecture-limited evaluation that introduced the OAX/AOX recomposition [\[1\]](#); the LMN configuration-sensitivity effect at a resolution not previously reported; a methodological pitfall the benchmark itself encountered, where an under-trained attacker makes a protocol-level defence look more effective than it is; and an explicit benchmarking of these results against the reliability-based [\[3\]](#), [\[5\]](#), hybrid-PUF [\[10\]](#), and bias-exploiting [\[1\]](#), [\[4\]](#) literature, with a labelled illustrative calculation ([Section 3.3](#)) against an independently reported structural-complexity finding [\[8\]](#). A systems designer choosing an AOX-PUF-based scheme needs to know which attacker family is the realistic threat, at what data cost, and whether published defences were tested against a competent baseline.

2. METHODS

2.1 The Arbiter PUF and Its Linear Additive Delay Model

The Arbiter PUF (APUF) compares the propagation delay of two nominally identical signal paths through a chain of switching elements; manufacturing-induced delay variation determines the response bit (Figure 1) [2]. This

variation is unclonable at the sub-micron scale but stable for a fixed device, so repeated measurement of the same challenge normally returns the same response, up to measurement noise.

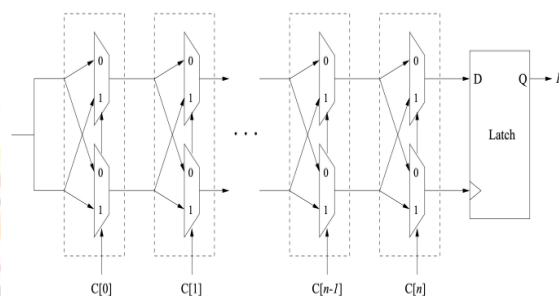


Figure 1. Arbiter PUF delay-chain structure. Adapted from Fei, Millwood, Gope, Miskelly, & Sikdar [2].

The APUF's challenge–response behaviour is well captured by the linear additive delay model, restated by Nguyen et al. [3] in the form used throughout this article. Writing the final delay difference between the two signal paths as Δ , and

w and Ψ for the weight and parity (feature) vectors derived from an n -bit challenge c :

$$\Delta = w[0]\Psi[0] + \dots + w[i]\Psi[i] + \dots + w[n]\Psi[n] = \langle w, \Psi \rangle, \quad r = \text{sign}(\Delta) \quad (1)$$

Equation (1) is the ground-truth model every attacker here is trying to approximate. Because the response is linear in the engineered parity features Ψ , not the raw challenge bits c , logistic regression is the natural first attacker: fitting a linear classifier to Ψ attempts to recover w directly, and when the feature map is known LR's inductive bias matches the target function exactly. Non-linear attackers earn their complexity only once composition stops being linear in Ψ . XOR-combining several APUF outputs already forces a linear attacker into an expanded

feature space [3]; the OR-AND-XOR (OAX/AOX) recombination evaluated here goes further, since combining OR-, AND-, and XOR-block outputs makes the final response non-linear in the per-chain Δ values, even though each chain stays linear in its own Ψ (Figure 2). LMN approximates that non-linearity with an engineered monomial basis; MLP and autoencoder learn non-linear representations directly; the Transformer targets longer-range interactions a chain-local feature map cannot express.

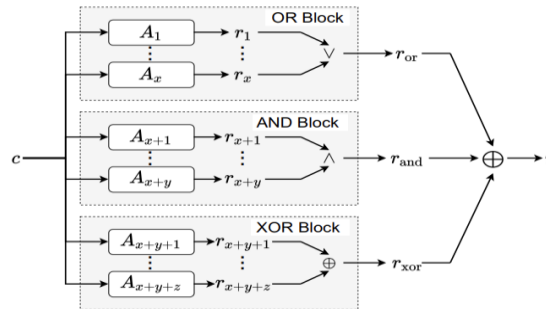


Figure 2. OAX-PUF (OR/AND/XOR) recomposition. Adapted from Yao et al. [1].

2.2 Baseline Response Uniformity of the OAX/AOX Composition

A second property governing attackability is uniformity: the proportion of 1-responses, ideally 0.5 for an unbiased primitive [1]. For an x-OR-PUF built from x sub-PUFs each with uniformity 0.5, the OR operation biases the output towards 1, since the output is 0 only if all x inputs are 0:

$$U_{OR} = 1 - \frac{1}{2^x} \quad (2)$$

By symmetry a y-AND-PUF is biased towards 0 by the same factor, while a z-XOR-PUF stays unbiased ($U_{XOR} = 1/2$) regardless of z [1]. Combining all three gate types, Yao et al. [1] show algebraically that the two opposing biases cancel exactly:

$$U_{OAX} = \frac{1}{2} \quad (3)$$

independent of (x,y,z). Before any BB84-style protocol-level defence is layered on top [7], the six attackers here face a response distribution with no exploitable marginal bias, so any accuracy advantage one shows over another comes from how well its architecture captures the OR/AND/XOR boundary, not from a skewed response distribution. The baseline is not unconditional, however: Wisiol and Pirnay [4] show that an XOR-APUF built from an even number of individually unbiased chains is itself

systematically biased, so Yao et al.'s [1] derivation can understate the exploitable bias present when an AOX mix's XOR block uses an even z (Section 4.4).

2.3 Reliability as an Attack Surface, and a Discussed-but-Not-Benchmarked Seventh Attacker

All six attackers here are accuracy-based: trained on single-measurement challenge–response pairs (c,r) and evaluated on held-out accuracy. A separate attack family instead exploits reliability — the consistency of a response to the same challenge across repeated measurements. Writing r_i for the i-th of M repeated measurements of a fixed challenge c, reliability is

$$R = \frac{1}{M} \sum_{i=1}^M r_i \quad (4)$$

the fraction of measurements returning 1 [3]. Becker [5] showed that R leaks information about the wire-delay structure: when $|\Delta|$ sits close to the APUF's noise threshold, the response becomes unreliable, so low-confidence challenges cluster near the decision boundary $\Delta = 0$ in Equation (1). Becker's divide-and-conquer CMA-ES attack exploits this directly, treating each weight as an independent Gaussian variable, and needs only a linear increase in CRPs as XOR-chain count grows, against the exponential growth that defeats naive accuracy-based attacks on the same

target. This seventh family is discussed in [Section 4.2](#) precisely because it is not benchmarked here: none of the six attackers use repeated-measurement data.

2.4 Six Attacker Architectures

Four are drawn from the standard pypuf baseline suite [6]: Logistic Regression (LR), a linear classifier on challenge features; Least-Squares Regression (LSR), a linear least-squares fit over the same feature space; LMN, which lifts challenge bits into a higher-order monomial basis before fitting a linear model; and the Multilayer Perceptron (MLP), a fully connected network capable of approximating arbitrary decision boundaries given sufficient capacity and data. Two further architectures extend the suite: the autoencoder (AE), trained on a joint reconstruction-and-classification loss over a compressed latent representation; and the Transformer, which applies multi-head self-attention over the challenge bit sequence to capture non-local interactions a chain-local feature map cannot express. LMN's lift and the three learned-representation attackers are all attempts to recover a decision function non-linear in Ψ once OR/AND composition sits on the per-chain linear model; which succeeds, and on which mixes, is the empirical question this benchmark answers.

2.5 Experimental Protocol

All six attackers were evaluated against three AOX-PUF families (aox-5, aox-6, aox-7; $n = 64$), each instantiated across four (x, y, z) mixes (twelve total), using pypuf [6] for CRP generation and the LR, LSR, LMN, and MLP baselines; AE and Transformer were implemented separately under an identical protocol. The mixes ranged from balanced configurations to those dominated

by a single gate type. Two budget regimes were used: 2,500 to 75,000 CRPs for the short-budget snapshot ([Section 3.1](#)), and up to ≈ 1.9 million CRPs to check whether short-budget rankings persist at scale, on the University of Sheffield's Stanage HPC cluster. Hyperparameters and training length were fixed per architecture rather than tuned per mix, with no early stopping — this isolates architecture and budget effects but also underlies the under-training artifact of [Section 4.6](#). No repeated-measurement data were collected, so [Section 2.3](#)'s seventh attack family remains discussed, not benchmarked.

3. RESULTS

3.1 Short-Budget Snapshot: Who Wins First?

By approximately 70,000 CRPs — the point at which most learning curves had visibly stabilised — the six attackers separated into clearly different regimes ([Table 1](#)). The MLP and the autoencoder were the fastest learners overall, reaching ≈ 98 – 99.5% accuracy on favourable mixes; the autoencoder's advantage comes from a learned, denoised latent representation rather than explicit feature engineering, and on several mixes it matched or slightly exceeded the MLP. Logistic regression settled reliably around 0.95 ± 0.02 regardless of mix. LMN was the most configuration-dependent attacker in the benchmark: on most mixes it barely exceeded chance-adjacent accuracy (0.52–0.60) even at the largest short-budget checkpoint, while on a few mixes it climbed to ≈ 0.98 — a bifurcation not observed for LR, MLP, or AE on the same mixes. Least-squares regression was the weakest attacker overall, reaching 0.85–0.90 when the input basis was revealed but near chance otherwise.

**Table 1**

Snapshot comparison of six attacker architectures against classical (unprotected) AOX-PUF CRPs at $\approx 70,000$ CRPs.

Attacker	Accuracy at $\sim 70k$ CRPs (best mixes)	Governing factor
MLP	~ 0.98 (e.g., aox-5 (2,2,1))	High representational capacity; latches onto XOR-style structure almost immediately
Autoencoder	~ 0.98 - 0.995 (best mixes)	Bottlenecked latent representation nearly as effective as MLP; can overfit and degrade with more data under a fixed training schedule
Logistic Regression	0.95 ± 0.02	Simple linear boundary; reliable and fast but capped below MLP/AE on most mixes
LMN (feature-lifting)	0.52 - 0.60 on most mixes; ~ 0.98 on a few	Highly configuration-sensitive: performance bifurcates sharply by (x,y,z) mix
Least-Squares Regression	0.85 - 0.90 (basis revealed); near chance without it	Weakest overall; strong model-feature mismatch on encoded-style inputs
Transformer	0.50 - 0.60	Systematically underfits at this model size and training schedule, independent of PUF mix

3.2 Configuration Sensitivity Across the Attack Set

No single attacker's ranking is mix-invariant across the twelve mixes tested: LR, MLP, and AE all show mix-dependent variation of several percentage points in how fast they reach their ceilings, consistent with the original observation that OAX/AOX mix choice shifts the classical learnability boundary [1]. The effect is strongest for LMN, whose accuracy switches sharply between near-chance and near-ceiling regimes as the mix changes, with little intermediate behaviour — the clearest case here of an attacker

whose effectiveness cannot be summarised by one number independent of circuit configuration. LSR's basis-known-versus-state-only gap, by contrast, is a property of the attacker-feature interaction rather than of the mix, and persists across all three AOX families tested.

3.3 An Illustrative CRP-Budget Calculation

Table 3 combines this study's own benchmark data with the reported finding of Wang et al. [8]. It is not a fused result: the top six rows are this study's own data (CRP budgets, AOX-PUF instances, and attacker implementations all

produced for this benchmark); the bottom row is Wang et al.'s [8] finding, using a fixed four-layer DNN, a plain OAX-PUF rather than this study's AOX-5/6/7 families, and a different experimental

axis — XOR-module count rather than CRP budget. The two halves are not blended into a single curve.

Table 3

. Illustrative comparison combining this study's own benchmark data with the reported finding of Wang et al. [8]. Rows 1-6: this study's own experimental data. Row 7: an independently reported finding, included for comparison only.

Attacker / data source	Approx. CRP budget to reach $\geq 95\%$ accuracy (unprotected AOX baseline)	Data source
MLP	~10,000-25,000 CRPs (favourable mixes)	This study (own experimental data)
Autoencoder	~10,000-25,000 CRPs (favourable mixes)	This study (own experimental data)
Logistic Regression	~50,000-75,000 CRPs (crosses 95% only on favourable mixes near the top of the tested range; settles at 0.95 ± 0.02)	This study (own experimental data)
LMN (favourable mixes only)	~75,000 CRPs; most mixes never cross 95% within the tested range (bifurcation)	This study (own experimental data)
Least-Squares Regression	Does not reach 95% within 2,500-75,000 CRPs (caps at 0.85-0.90, basis revealed)	This study (own experimental data)
Transformer	Does not reach 95% within 2,500-75,000 CRPs (caps at 0.50-0.60)	This study (own experimental data)
Fixed 4-layer DNN vs. XOR-module count	Accuracy collapses to ~50% (chance) once pure-XOR module count reaches 6 (64-bit OAX-PUF) or 5 (128-bit OAX-PUF), CRP budget and architecture held fixed	Wang et al. [8] (reported finding; different PUF instances, bit-widths, and attacker implementation)



In this study's own data, the two best-performing architectures (MLP, autoencoder) cross 95% accuracy on favourable mixes using roughly 10,000–25,000 CRPs — three to seven times fewer than the budget at which the weaker linear baselines approach, but do not always reach, the same threshold. Wang et al. [8], holding CRP budget and a fixed four-layer DNN constant and instead varying defender structural complexity, report accuracy collapsing to chance ($\approx 50\%$) once pure-XOR module count reaches 6 for their 64-bit configuration (5 for 128-bit). The two results point to the same tradeoff from opposite directions: attacker sophistication (MLP/AE over LR/LSR) buys an order-of-magnitude reduction in the CRPs needed to defeat a moderately complex mix, while defender structural complexity, attacker held fixed, can defeat even a capable neural attacker. Neither axis alone settles which side wins in practice — Wang et al.'s defender has not been tested against this study's best attacker, and this study has not tested its MLP or autoencoder at the XOR depths where

Wang et al.'s DNN collapses. Table 3 is a research-gap statement, not a security guarantee in either direction; Section 4.5 returns to it.

4. DISCUSSION

4.1 Relative to Yao et al. [1]

Yao et al. [1], who introduced the OAX/AOX recomposition, reported LR reaching 99.09% at 1.4 million CRPs for one mix, and expected MLP/DNN-class attackers to progress further where linear models plateau. This benchmark corroborates both: LR lands in a comparable band at similar CRP scales, and the MLP exceeds LR on nearly every mix tested. The contribution beyond Yao et al.'s architecture-limited evaluation is four further attacker types — LSR, LMN, AE, Transformer — run under an identical protocol, surfacing the LMN bifurcation and the Transformer artifact a two-attacker study could not reveal. Table 2 summarises this and the comparisons below.

Table 2

. Comparison of this study's six-attacker benchmark against the prior work it builds on and the attack/defence literature it is benchmarked against.

Prior work	Prior finding	This study's finding	Relationship
Yao et al. [1]	LR reaches 99.09% at 1.4M CRPs for one mix; MLP/DNN expected to progress further.	Confirms both; adds LSR, LMN, AE, Transformer, surfacing the LMN bifurcation.	Extends to a six-attacker comparison.
Wang et al. [8]	Fixed 4-layer DNN collapses to $\sim 50\%$ once XOR count reaches 5-6.	MLP/AE reach ~ 0.98 - 0.995 well inside the tested XOR depths (Table 3).	Qualifies: result is architecture- and depth-specific (§4.5).

Wang et al. [9]	Covariance/Boolean attacks keep eroding strong-PUF hardening vs. linear baselines.	AE matches/exceeds MLP; Transformer underperforms both at equal budget.	Sophistication alone, untrained, is not sufficient.
Wisioł et al. [6] (pypuf)	Standard reference suite (LR, LSR, LMN, MLP).	Adds AE, Transformer under an identical protocol; flags a training check.	Procedural extension to the suite.
Becker [5]; Nguyen et al. [3]	Reliability-based CMA-ES scales linearly, not exponentially, in XOR count.	Not benchmarked (no reliability data); explains the LMN bifurcation (§4.2).	Discussed, not replicated.
Fei et al. [10]	Transition Theorem: hybrid PUFs (incl. OAX) are not ML-immune.	MLP/AE independently reach near-ceiling accuracy on several AOX mixes.	Corroborates via an independent method (§4.3).

4.2 Relative to the Reliability-Based Attack Literature [3], [5]

Because all six attackers here are accuracy-based, the comparison with reliability-based attacks is qualitative. Equation (4) names what such an attacker exploits that none of the six do: per-challenge reliability R rather than a single measured response r . Becker's [5] result that reliability-based CMA-ES scales linearly, not exponentially, in XOR-chain count bears directly on the LMN bifurcation (Sections 3.1–3.2): LMN's chance-adjacent performance on most mixes is consistent with a linear-in-the-lifted-basis attacker hitting the same exponential wall that motivated Becker's alternative, while the mixes where LMN nearly matches the MLP plausibly have a monomial basis that aligns with the decision boundary. Nguyen et al.'s [3] analysis of multi-stage compositions suggests the OR and

AND stages, outside Becker's XOR-only analysis, would need their own reliability model before a seventh attacker could be benchmarked fairly here.

4.3 Relative to the Hybrid-PUF / Transition-Theorem Literature [10]

Fei et al. [10] attack hybrid PUFs — heterogeneous Feed-Forward and OAX-PUFs claimed secure against machine-learning attack — with a Transition Theorem identifying the optimal attacker model class for a given hybrid structure. Their conclusion that OAX-PUFs are not immune corroborates this study's finding that the MLP and autoencoder reach near-ceiling accuracy on several AOX-5/6/7 mixes. The studies differ in method, not conclusion: Fei et al. [10] derive the optimal attacker class analytically; this benchmark runs a fixed menu of six and lets the data show which wins, a natural basis for using

the Transition Theorem to predict winners ahead of time and checking that prediction against [Table 1](#).

4.4 Resolving the Uniformity-Versus-Bias Tension [1], [4]

[Section 2.2](#) flagged a tension between Yao et al.'s [1] derivation that the full OAX-PUF composition has uniformity exactly 0.5, and Wisiol and Pirnay's [4] finding that an XOR-APUF built from an even number of individually unbiased chains is nonetheless systematically biased. Yao et al.'s [1] algebra assumes each sub-PUF already has uniformity exactly 0.5, an idealisation that does not generally hold for even-chain XOR-APUFs [4]. Several of this study's twelve mixes use an even z , so those mixes may carry a small residual bias that [Equation \(3\)](#) does not capture, which could make the composition marginally easier for LR or LSR to exploit. This benchmark cannot separate that effect from boundary shape; it is flagged as a testable follow-up rather than resolved here.

4.5 The Attacker-Architecture-Versus-Defender-Structural-Complexity Tradeoff

[Table 3](#) carries a practical implication for anyone evaluating an OAX/AOX-style defence: showing that a fixed architecture fails against a sufficiently deep composition, as Wang et al. [8] do for their four-layer DNN at XOR-module count $\geq 5-6$, is not the same as showing the composition secure against machine-learning attack in general. Switching architecture (LR/LSR to MLP/AE) is worth an order of magnitude in CRP efficiency on the compositions tested here; a defender who tested only one fixed architecture at increasing depth has tested one axis of the attacker's choices, not both. A complete evaluation should vary architecture and defender depth jointly — neither study does so alone, and [Table 3](#) is offered to motivate that, not settle it.

4.6 A Methodological Warning: Under-Trained Attackers Can Fake a Defence

The Transformer, evaluated at a comparatively small configuration under the same fixed schedule as the other five attackers, consistently underfit classical CRPs, crawling in the 0.50–0.60 range across every mix and budget tested. Because a BB84-encoded accuracy ceiling measured elsewhere in this research programme sits near 0.85 [7], plotting Transformer accuracy against CRPs shows the encoded curve sitting above the classical one — the opposite of every other attacker here, and the opposite of what a throttling protocol should produce. This is not evidence that encoding helped the Transformer; it is an artifact of the classical model never reaching convergence at this configuration. A milder version appeared in the autoencoder, whose classical accuracy peaked early on one mix then declined at larger budgets under the fixed-epoch schedule — under-training, not lost signal. A benchmark comparing a defended condition against an undefended baseline can misreport effectiveness whenever the baseline is not trained to its own ceiling first; this is caught here only because the other five attackers, under the same schedule, do reach a stable classical ceiling above the encoded one. BB84-style quantum encoding is not the only defence this pitfall threatens [7]: Zhuang et al.'s [11] purely classical challenge-obfuscation interface for Arbiter-PUF variants, evaluated the same way (defended curve against undefended baseline), is exactly as exposed. A natural extension would evaluate it with this benchmark's six architectures, trained to a verified ceiling first, to see whether a classical and a quantum-protocol defence degrade the same attackers by comparable amounts.

4.8 Practical Recommendations and Limitations

Two recommendations follow. Anyone selecting a machine-learning baseline to evaluate a PUF's classical vulnerability should start with an MLP or autoencoder, not a linear baseline alone, and should report LMN results only alongside the specific mix tested, given how sharply its performance bifurcates. Anyone evaluating a defence by comparing a defended condition against an undefended baseline should first verify, with a known-good attacker such as LR or MLP, that the baseline has been trained to its own ceiling before concluding that a gap between curves reflects the defence rather than an under-trained comparison point.

These results come from simulated CRPs on a software AOX-PUF model, not fabricated silicon, and the twelve mixes tested do not exhaust the configuration space; the LMN bifurcation may not generalise to untested mixes. All six attackers trained under a single, fixed-epoch schedule not tuned per architecture, so the Transformer and autoencoder results of [Section 4.6](#) are characteristic of that schedule, not an inherent ceiling. No reliability data were collected, so [Sections 4.2 and 4.4](#) remain qualitative. Four extensions follow: re-evaluate the autoencoder and Transformer with per-architecture tuning and early stopping; sweep (x,y,z) mixes more finely around the LMN bifurcation, guided by Fei et al.'s [\[10\]](#) Transition Theorem; repeat the benchmark on physical silicon CRPs; and implement a reliability-based attacker [\[5\]](#) against the same twelve mixes as an eighth row in [Table 1](#). A reporting checklist requiring a demonstrated classical ceiling before any defended-versus-undefended comparison would benefit the wider field.

5. CONCLUSION

This article asked whether any single machine-learning attacker family dominates against an AOX-PUF independent of circuit composition. Across six architectures run under identical conditions against the same substrate, the answer is no: an autoencoder matches a purpose-built MLP, a feature-lifted linear model bifurcates sharply by circuit configuration, and a Transformer with an inadequate training schedule can underperform every other attacker so badly that an unrelated protocol-level defence appears, misleadingly, to outperform the classical baseline it should be compared against. Benchmarked against the wider literature, these results corroborate and extend the original OAX/AOX evaluation [\[1\]](#) and the hybrid-PUF Transition Theorem result [\[10\]](#), complement the deep-learning attack literature [\[8\]](#), [\[9\]](#) while showing that a fixed architecture's failure against a sufficiently deep composition is not the same as general security ([Section 4.5](#)), and leave open concrete questions about reliability-based attacks [\[3\]](#), [\[5\]](#), residual bias on even-chain mixes [\[4\]](#), and classical obfuscation defences [\[11\]](#). The benchmark result and the under-training pitfall belong together: attacker comparisons in PUF security research are only as informative as the training discipline applied to every attacker in the comparison, not only the one a study means to showcase, and a defence is shown secure only against the attackers actually tested against it, not against the field's best available attacker in general.

REFERENCES

1. J. Yao, L. Pang, Y. Su, Z. Zhang, W. Yang, A. Fu, and Y. Gao, "Design and evaluate recomposited OR-AND-XOR-PUF," IEEE

- Transactions on Emerging Topics in Computing, vol. 10, no. 2, pp. 662–677, 2022.
<https://doi.org/10.1109/TETC.2022.3170320>
2. H. Fei, O. Millwood, P. Gope, J. Miskelly, and B. Sikdar, “Attacking delay-based PUFs with minimal adversarial knowledge,” IEEE Transactions on Information Forensics and Security, 2024.
<https://arxiv.org/abs/2403.00464>
 3. P. H. Nguyen, D. P. Sahoo, C. Jin, K. Mahmood, U. Rührmair, and M. van Dijk, “The Interpose PUF: Secure PUF design against state-of-the-art machine learning attacks,” IACR Transactions on Cryptographic Hardware and Embedded Systems, vol. 2019, no. 4, pp. 243–290, 2019.
<https://doi.org/10.13154/tches.v2019.i4.243-290>
 4. N. Wisiol and N. Pirnay, “XOR Arbiter PUFs have systematic response bias,” in Financial Cryptography and Data Security (FC 2020), LNCS vol. 12059, pp. 50–57, 2020.
https://doi.org/10.1007/978-3-030-51280-4_4
 5. G. T. Becker, “The gap between promise and reality: On the insecurity of XOR Arbiter PUFs,” in Cryptographic Hardware and Embedded Systems (CHES 2015), LNCS vol. 9293, pp. 535–555, 2015.
https://doi.org/10.1007/978-3-662-48324-4_27
 6. N. Wisiol, C. Gräbnitz, C. Mühl, B. Zengin, T. Soroceanu, N. Pirnay, K. T. Mursi, and A. Baliuka, “pypuf: Cryptanalysis of physically unclonable functions,” [Computer software], 2021. <https://github.com/nils-wisiol/pypuf>
 7. I. Makhamatdjonov, “Quantum-inspired PUFs: Advantages, challenges, and applications,” Unpublished MSc dissertation, University of Sheffield, 2025.
 8. H. Wang, W. Hao, Y. Tang, B. Zhu, W. Dong, and W. Liu, “Deep neural network modeling attacks on arbiter-PUF-based designs,” Cybersecurity, vol. 8, no. 11, 2025.
<https://doi.org/10.1186/s42400-024-00308-7>
 9. H. Wang, W. Liu, W. Cai, Y. Lu, and C. Wan, “Efficient attacks on strong PUFs via covariance and Boolean modeling,” ACM Transactions on Design Automation of Electronic Systems, vol. 29, no. 6, 2024.
<https://doi.org/10.1145/3687469>
 10. H. Fei, P. Gope, O. Millwood, and B. Sikdar, “Optimal machine-learning attacks on hybrid PUFs,” in European Symposium on Research in Computer Security (ESORICS 2024), LNCS vol. 14982, pp. 251–270, 2024.
https://doi.org/10.1007/978-3-031-70879-4_13
 11. Y. Zhuang, K. T. Mursi, and L. Gaoxiang, “A challenge obfuscating interface for Arbiter PUF variants against machine learning attacks,” arXiv:2103.12935, 2021.
<https://arxiv.org/abs/2103.12935>