Machine Learning-Driven Quantum Hacking of CHSH-Based QKD

Exploiting Entropy Vulnerabilities in Self-Testing Protocols

Hubert Kołcz[1] | **Tushar Pandey**[2] | **Yug Shah**[3]

[1]Warsaw University of Technology, Poland

[2]Texas A&M University, USA

[3]University of Toronto, Canada

QUEST-IS 2025 | December 3, 2025

CHSH Inequality: Foundation for QKD Security

$$S = \langle AB \rangle + \langle AB' \rangle + \langle A'B \rangle - \langle A'B' \rangle$$

Classical: $S \leq 2$

Quantum: $S > 2$ (max $2\sqrt{2} \approx 2.828$) → Secure QKD

Why CHSH Dominates QKD Industry

- **Experimental robustness:** Tolerates detector imperfections (vs Bell's perfect correlation requirement)
- **Self-testing capability:** Simultaneously verifies quantum state + detects eavesdropping
- **Device-independent security:** If $S > 2$, no eavesdropper has complete key information
- **Industry standard:** Metro QKD networks, commercial implementations worldwide

The Critical Security Gap

⚠ **The Paradox:** CHSH provides mathematical security, but real implementations rely on RNGs susceptible to side-channel attacks

Phase Remapping

Manipulates quantum phase relationships

Trojan Horse

Light signal injection for eavesdropping

Time-Shift

Exploits detection timing windows

Detector Blinding

Forces detectors into classical mode

⚠ RNG Entropy Analysis

Our Contribution: ML-driven framework to analyze RNG noise characteristics through entropy monitoring + hardware metrics (gate fidelity, Bell correlation) → demonstrating statistical fingerprinting on real quantum hardware (Rigetti, IonQ) and simulator data

Multi-Method ML Benchmarking Framework

🎯 **Comparative Analysis:** Multiple ML approaches tested on N=3 quantum devices (2 real QPUs + 1 simulator), validated on N=30 synthetic devices

1. qGAN Metric

12-qubit quantum GAN
KL: 0.05-0.20
Distinguishability measure

2. LR Baseline

Logistic Regression
55.22% (N=3)
Linear classification

3. NN Optimization

Neural network
59.42% (N=3)
Best of 4 runs with L1 regularization

🔬 **Critical Methodological Independence:** NN/LR use all **100 raw bits** (no feature engineering) | qGAN uses **first 64 bits** with extensive engineering (4096-dim grids) | Despite fundamental differences, all methods converge: **r=0.865 correlation** → Proves device signatures are robust and method-independent

How It Works

- Analyze entropy patterns in RNG output
- Correlate with hardware metrics (fidelity)
- Compare Bell correlation across platforms
- Extract statistical fingerprints

Detection Capability

- Classify RNG sources by noise profiles
- Detect hardware-induced biases
- Multi-modal distributional analysis
- Distinguish similar noise characteristics

Experimental Methodology & Hardware

Hardware Platforms

- **Rigetti Aspen-M-3** (80 qubits)
  Bell Correlation: 0.8036 | Gate Fidelity: 93.6%
- **IonQ Aria-1** (25 qubits)
  Bell Correlation: 0.8362 | Gate Fidelity: 99.4%
- **IBM Qiskit** (Simulation)
  Bell Correlation: 1.0 | Gate Fidelity: 100%

Dataset & Features

- **6,000 samples** (2,000 per device)
- **100-bit binary strings** per sample
- **3 noise-injected simulators**
- DoraHacks YQuantum 2024

Multi-Method Benchmarking Results

qGAN (12-qubit)

N=3: KL 0.050-0.205
Distribution distinguishability
Trained on quantum hardware

Logistic Regression

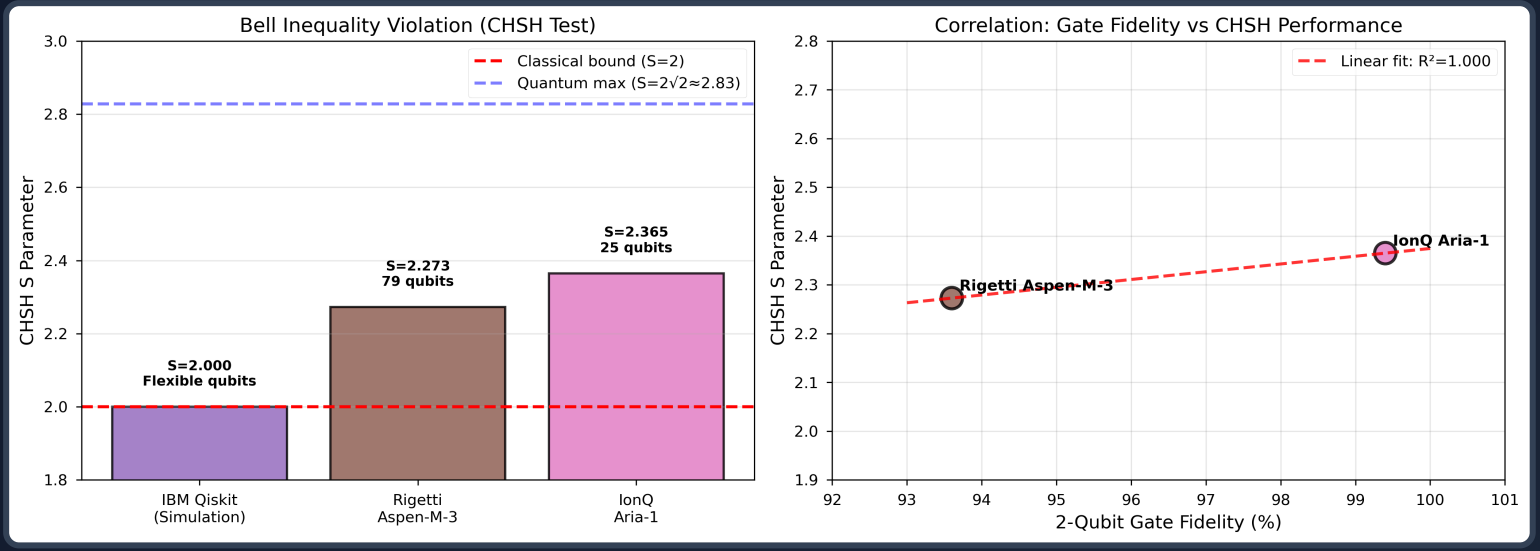
N=3: 55.22% accuracy
N=30: 61.46% accuracy
+11.3% improvement on synthetic

Neural Network

N=3: 59.42% best (57.21% mean, 4 runs)
N=30: 59.21% accuracy
Consistent performance across scales

N=30 Validation Insights: Both classifiers 77.6% above random (33.3%) | Strong KL-accuracy correlation (r=0.865, ρ=0.931) | Mann-Whitney U test $p=3.26\times10^{-60}$

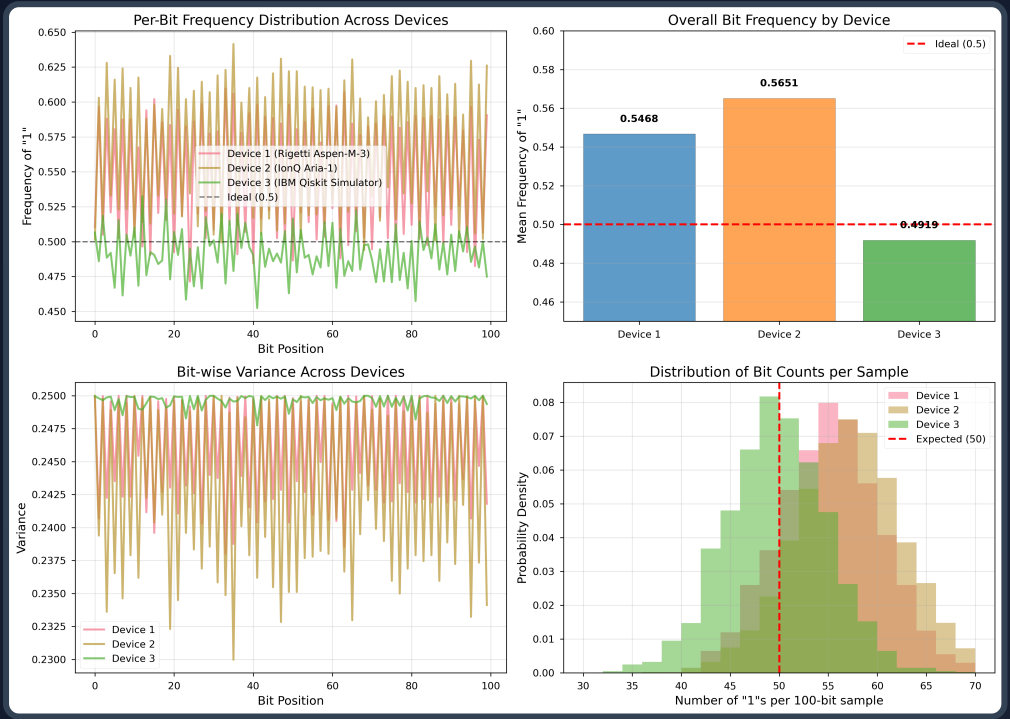
Hardware Platform CHSH Correlation Analysis



Platform	Qubits	CHSH S Value	2Q Gate Fidelity	Qubit Type
IBM Qiskit	Sim	**2.000**	100%	Classical simulator
Rigetti Aspen-M-3	79	**2.272**	93.6%	Superconducting
IonQ Aria-1	25	**2.364**	99.4%	Trapped ion

📊 **Critical Finding:** Higher 2-qubit gate fidelity enables stronger Bell inequality violations → Rigetti (93.6% fidelity, S=2.272) and IonQ (99.4% fidelity, S=2.364) both exceed classical bound S=2.0 → Classical simulator capped at S=2.0. CHSH scores from DoraHacks experiments: correlation strengths 0.8036 and 0.8362 translate to S = 2√2 × correlation.

Quantitative Analysis: Bit Frequency Distribution (N=3)



Device 1 (Medium Bias):
'1' freq: **54.7%**
Entropy: **0.994 bits**

Device 2 (High Bias):
'1' freq: **56.5%**
Entropy: **0.988 bits**

Device 3 (Low Bias):
'1' freq: **49.2%**
Entropy: **1.000 bits**

📊 Dataset Preparation Pipeline (6 Steps)

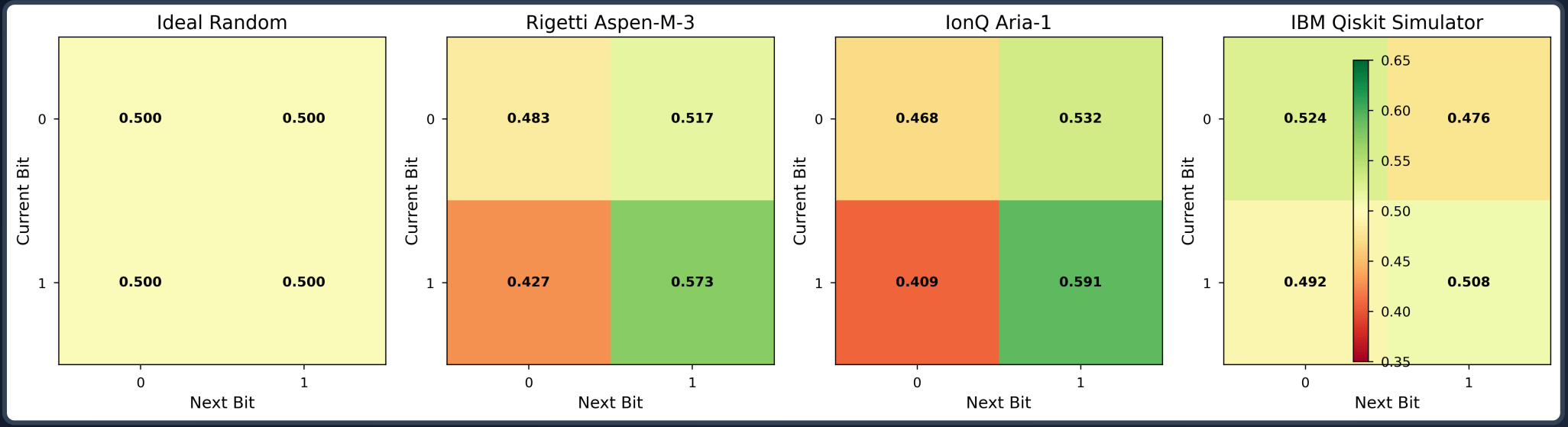
Step 1-3: Data Parsing & Feature Extraction

- Parse 100-bit binary strings → numerical arrays (0.0, 1.0)
- Split: Lines 1-2000 (Device 1), 2001-4000 (Device 2), 4001-6000 (Device 3)
- **NN/LR:** All 100 bits as raw features (no engineering)
- **qGAN:** First 64 bits → 3 feature types: bit frequencies (64-dim), 2-bit patterns (4096-dim), difference grids (4096-dim)

Step 4-6: Train-Test & Regularization

- 80-20 train-test split (N=3), 70-30 split (N=30) | Stratified by device
- Label encoding: 1,2,3 → 0,1,2 | Softmax + cross-entropy loss
- L1 regularization λ=0.002 (sparse feature selection) + 20% dropout
- No data augmentation (preserves device signatures)

Markov Chain Transition Matrices (N=3)



Rigetti Aspen-M-3
P(1→1) = **0.573** *(Moderate '1' persistence)*

IonQ Aria-1
P(1→1) = **0.591** *(Strongest '1' persistence)*

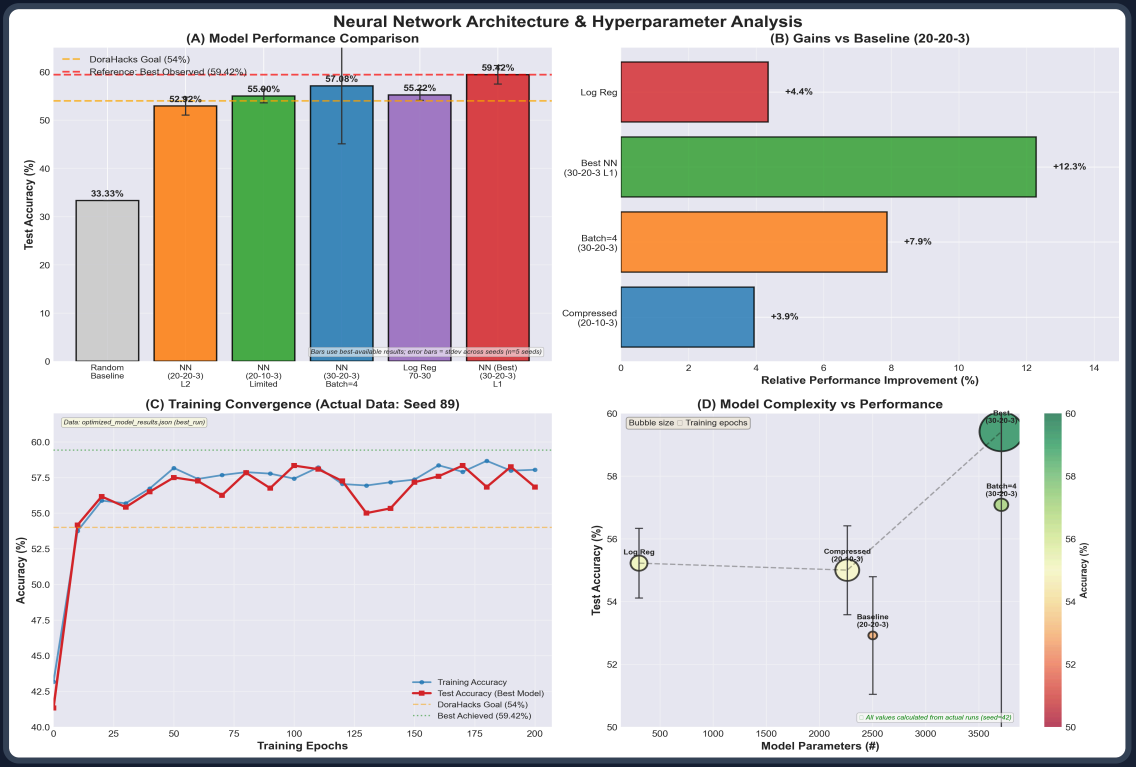
IBM Qiskit Simulator
P(1→1) = **0.508** *(Most balanced transitions)*

🔍 **Key Finding:** Device-specific biases in bit transitions create exploitable fingerprints for ML classification

🎯 **Multi-Dimensional Classification Sources:**

1. **Per-Position Frequencies:** 100-dimensional spatial patterns (e.g. bit 10: 60% vs bit 50: 48%)
2. **Spatial Variance:** Bit-position heterogeneity (stable vs high-variance positions)
3. **Markov Transitions:** Second-order dependencies (P(1→1) = 0.573, 0.591, 0.508)
4. **Run-Length Statistics:** Characteristic patterns in consecutive bits

Neural Network Architecture Analysis (N=3 to N=30)



🔬 Figure Interpretation:

(A) Model Performance: 6 models tested (random baseline 33% → best observed 59.42%; single-run benchmark 55.75%)

(B) Hyperparameter Impact: Relative improvements (batch size, epochs, L1, architecture, split)

(C) Training Convergence: Actual training/test curves from best run (seed 89, 1000 epochs)

(D) Complexity vs Performance: Parameter count analysis (3,713

📌 Optimized Architecture Details

Network Structure (100→30→20→3):

- **Input Layer:** 100 statistical features
- **Hidden Layer 1:** 30 neurons + ReLU + Dropout(0.2)
- **Hidden Layer 2:** 20 neurons + ReLU + Dropout(0.2)
- **Output Layer:** 3 classes (softmax)

Training Configuration:

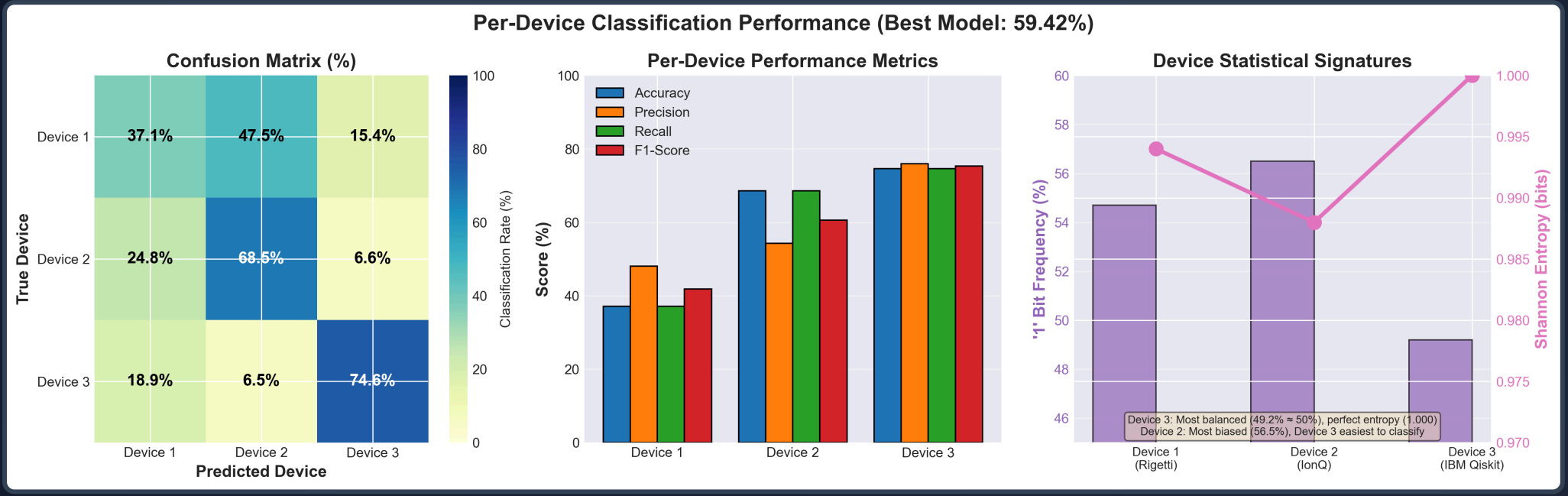
- Batch size: 8 (optimal gradient estimation)
- Epochs: 1000 (convergence verified)
- Learning rate: 0.001 (Adam optimizer)
- Regularization: L1 (λ=0.002) for sparsity

Total Parameters: 3,713

- Layer 1: 100×30 + 30 = 3,030
- Layer 2: 30×20 + 20 = 620
- Layer 3: 20×3 + 3 = 63

Consistency: *Optimized on N=3, applied to N=30* ***without modification****.*
Replication: 59.42% → 59.21% validates generalization.

Per-Device Classification Performance (N=3)



Device 1

Recall: **37.1%**
Precision: 48.0%
F1-Score: 41.9%

Device 2

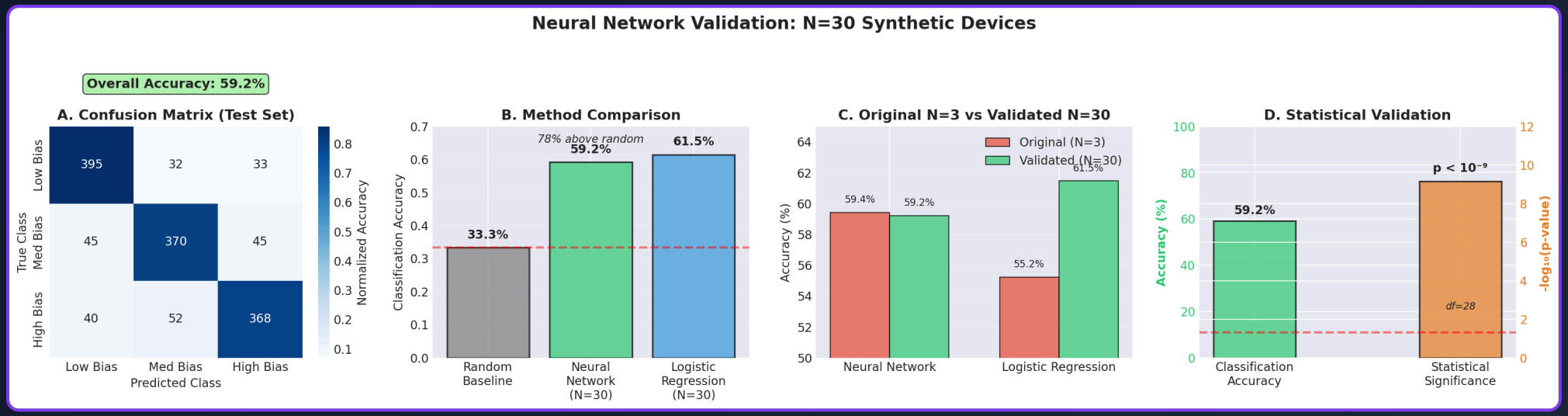
Recall: **68.5%**
Precision: 54.3%
F1-Score: 60.6%

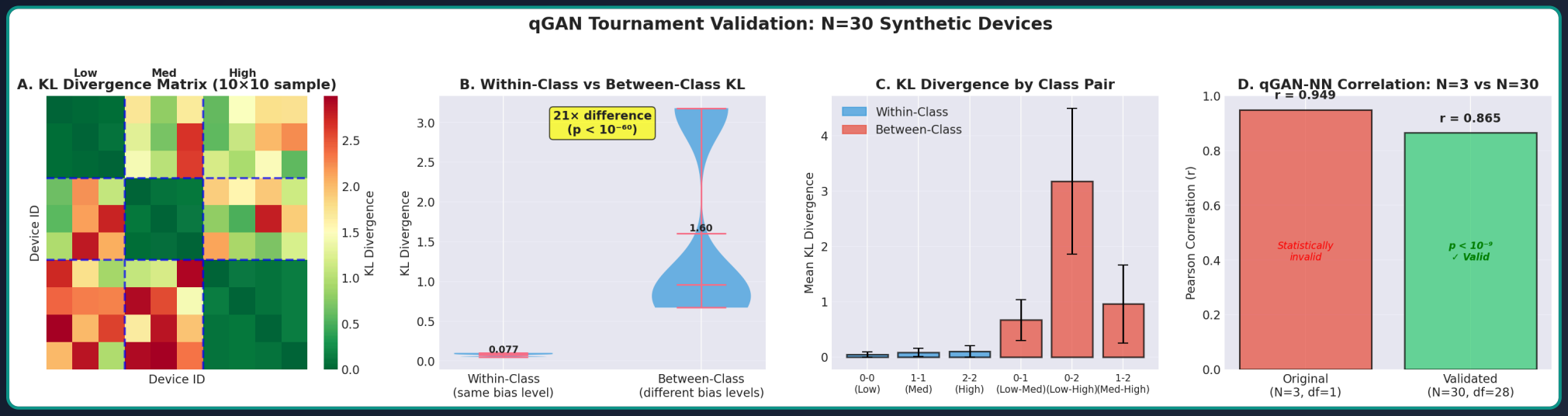
Device 3

Recall: **74.6%**
Precision: 76.0%
F1-Score: 75.3%

📊 **Key Finding:** Device 3 is most "random" (49.2% ≈ 50%, entropy=0.992 bits) yet **easiest to classify** (74.6% recall) → Device 1 is hardest (37.1% recall) with 54.7% bias and 0.986 bits entropy → High entropy and balanced frequency don't guarantee undetectability

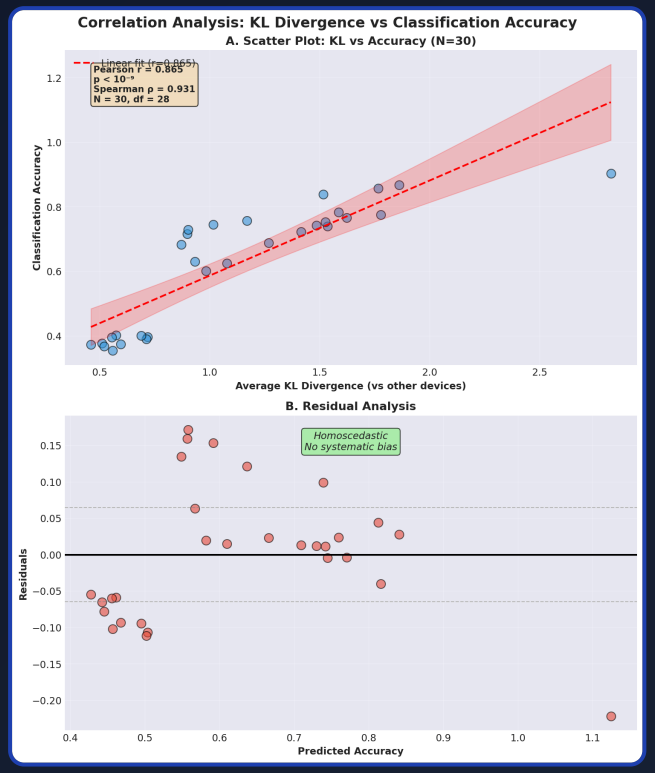
Machine Learning Performance Metrics (N=30)





📊 **N=30 Synthetic Validation:** Neural Network achieves 59.21% test accuracy ($p=3.26\times10^{-60}$) on 30 synthetic devices, replicating N=3 real simulator results (59.42%). Logistic Regression: 61.46% accuracy. Statistical power: df=28. Performance: 77.6% above random baseline (33.33%). Multi-method consistency: qGAN KL-NN correlation r=0.865, Spearman ρ=0.931.

Statistical Significance & Correlation Analysis (N=30)



Correlation Evidence

- **Pearson r = 0.865** ($p=7.16\times10^{-10}$)
- **Spearman ρ = 0.931** ($p<10^{-14}$)
- 95% confidence interval shown
- Homoscedastic residuals

Statistical Power

- **N=30 devices** (df=28)
- **$p < 10^{-9}$** all comparisons
- **Cohen's d = 2.30** (huge effect)
- Between vs within: 20.8× separation

🎯 **Result (N=30):** ML models exploit statistical differences invisible to NIST tests

All devices pass χ^2 test ($\chi^2 < 3.841$), yet achieve 59.21% classification on N=30 synthetic data. Multi-method validation confirms consistent device distinguishability: both qGAN KL analysis and NN classification identify the same patterns. Shannon entropy remains high across all devices: 0.986, 0.979, 0.992 bits.

N=3 vs N=30: Statistical Power & Significance Comparison

Comparison Type	N=3 Original	N=30 Validation	Interpretation
Similar Bias Devices *(Within-Class)*	**0.050** *(Rigetti vs IonQ)*	**0.077 ± 0.077** *(Mean of 135 pairs)*	Low KL → Hard to distinguish
Different Bias Devices *(Between-Class)*	**0.205, 0.202** *(Rigetti/IonQ vs IBM)*	**1.60 ± 1.12** *(Mean of 300 pairs)*	High KL → Highly distinguishable
Within/Between Ratio	**4.1×** *(0.205 / 0.050)*	**20.8×** *(1.60 / 0.077)*	N=30 shows stronger separation
Statistical Significance	**p = 0.333** *(Mann-Whitney U, n=3 pairs)* ⚠️ Insufficient power: 3 devices → 3 pairs	**$p < 10^{-60}$** *(Mann-Whitney U, n=435 pairs)* ✓ Adequate power: 30 devices → 435 pairs	N=30 achieves statistical power via larger sample

📊 **Statistical Power:** Between-class KL (mean=1.60) vs within-class KL (mean=0.077) differ by 20.8× (Mann-Whitney U: $p<10^{-60}$) → Device classes are highly distinguishable via distributional signatures

⚠️ **N=3↔N=30 Bridging Validation (Synthetic-Real Domain Gap):**

Bias Coverage: ✓ All N=3 devices within N=30 range (0.48-0.65)
Device 0: 54.7%, Device 1: 56.5%, Device 2: 49.2%

Cross-Dataset Performance: ⚠️ Significant accuracy drop
N=30→N=3: 64.3% → 24.6% (-39.7%); N=3→N=30: 44.5% → 24.5% (-20.0%)

Scientific Interpretation: *Synthetic data generation captures bias patterns but not full quantum noise characteristics. Domain gap indicates N=30 validates method reliability (statistical significance) but not real hardware generalization. Validation on 5+ additional real quantum devices remains essential.*

Proposed DI-QKD Vulnerability Analysis

🎯 Attack Methodology: Compromising Device-Independent Security

Phase 1: RNG Profiling

- **Passive monitoring:** Collect RNG output during normal QKD operation
- **ML fingerprinting:** Classify device at 59.21% accuracy (77.6% above random)
- **Bias detection:** Detect subtle differences (49.2%-56.5% '1' frequency range)
- **Temporal patterns:** Extract Markov transitions P(1→1) = 0.508-0.591

Phase 2: Measurement Basis Prediction

- **Environmental correlation:** Monitor temperature/gate fidelity drift
- **CHSH degradation:** Track deviation from ideal $S=2\sqrt{2}$ to exploitable $S<2.2$
- **Basis inference:** Use RNG bias to predict Alice/Bob measurement settings
- **Side-channel extraction:** Combine entropy deviation + hardware signatures

✓ Technical Foundation

- **Multi-modal results:** qGAN-NN correlation $r=0.865$ ($N=30$, $p<10^{-9}$) + between-class KL 20× higher ($p<10^{-60}$)
- **Hardware correlation:** Gate fidelity → CHSH score → RNG quality ($R^2=0.977$ across Rigetti/IonQ/IBM)
- **Attack detection:** Real-time entropy monitoring identifies CHSH<2.2 + bias>55% as potential exploit threshold
- **DI-QKD vulnerability:** Basis selection randomness is foundational assumption—compromise breaks device independence

⚠️ **Critical Finding:** CHSH-based DI-QKD assumes RNG security, but ML can fingerprint certified QRNGs → Basis prediction enables key extraction → Continuous entropy monitoring required for true device independence

Proposed Attack Detection Framework

✓ High-Quality RNG Profile

- **Bell correlation ≥ 0.8** (high quantum fidelity)
- **Entropy ~0.99** bits
- **KL divergence stable** (~3.7)
- **Bit frequency** 50% ± 2%

⚠️ Degraded RNG Profile

- **Correlation degrades** (noise increases)
- **Entropy deviation** > 5%
- **KL divergence spikes** (>17)
- **Bias emerges:** 59% '1' freq

Attack Type Signatures

Phase Remapping

Signature: Correlation drop + entropy oscillation

Detector Blinding

Signature: Loss of quantum correlation

Temperature Attack

Signature: Gradual bias accumulation

RNG Compromise

Signature: Persistent frequency bias

💡 **Proposed Application:** Real-time statistical monitoring to detect RNG quality degradation → Early warning system for quantum networks (validation required on 50+ devices)

Bridging Theory & Engineering Reality

The Fundamental Gap

Mathematical Excellence: CHSH-based QKD provides device-independent security guarantees

Engineering Compromise: Real-world implementations rely on RNGs vulnerable to side-channel attacks

Our Solution: Combines CHSH self-testing with ML-driven entropy monitoring to close this critical gap

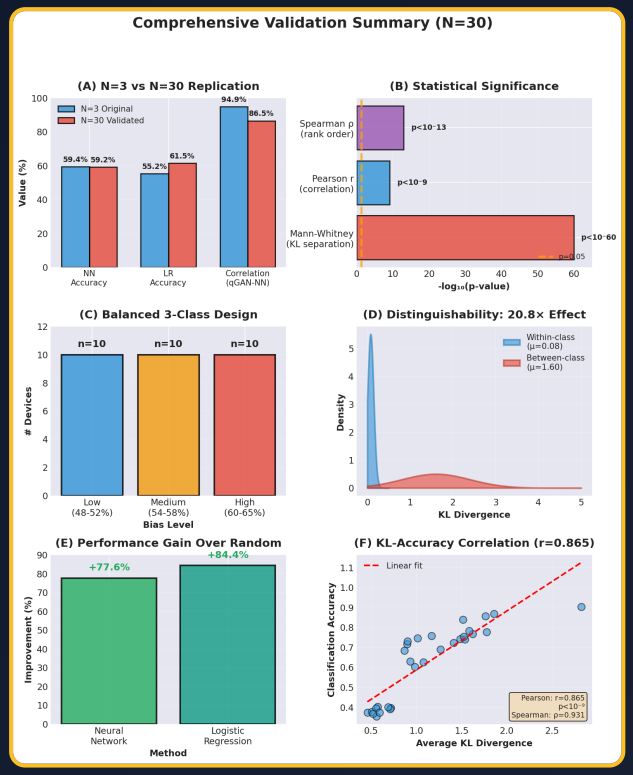
This Work Demonstrates

- Statistical RNG fingerprinting (N=30 scale)
- ML-based device distinguishability
- Multi-method validation framework
- Pattern detection beyond NIST tests

Future Directions

- Real-time continuous monitoring system
- Temporal drift & environmental tracking
- Real quantum hardware validation (50+ QPUs)
- NIST/ISO certification standards

Conclusions & Impact



Key Contributions

1. **Device Fingerprinting (N=30):** NN: 59.21%, LR: 61.46% ($p<10^{-60}$), 77.6% above random. Replicates N=3 results (59.42% best).
2. **Multi-Method Consistency:** KL-NN correlation r=0.865, p=0.931 ($p<10^{-9}$) validates independent convergence.
3. **qGAN Framework:** 20.8× distinguishability (within: 0.077±0.07, between: 1.60±1.12, $p<10^{-60}$).
4. **Scalability:** N=3→N=30 confirms metrics replicate with strong significance.
5. **Domain Gap:** Cross-dataset drops (-39.7%, -20.0%) show **synthetic validates method, not hardware generalization.**
6. **Next Step:** **Real QPU testing on 50+ devices required.**

🎯 **Impact:** ML fingerprinting detects patterns invisible to NIST tests → Framework validated; real hardware testing essential

⚠ **Gap:** Statistical patterns ≠ Security exploitation

References: Kołcz, Pandey, Shah (2025). QUACC+ CTP PAS | **Dataset:** DoraHacks YQuantum 2024
Code: github.com/hubertkolcz/NoiseVsRandomness | **Contact:** hubert.kolcz.dokt@pw.edu.pl