Physics-Aware Temporal Feature Engineering for Eavesdropping Detection in BBM92 Quantum Key Distribution

Saksham Gupta
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science, Pilani, Hyderabad Campus
Hyderabad, India
f20250785@hyderabad.bits-pilani.ac.in

***Abstract*—Static Quantum Bit Error Rate (QBER) thresholding is the standard defense mechanism in deployed Quantum Key Distribution (QKD) systems. In noisy free-space optical (FSO) channels, however, natural atmospheric variations can camouflage short, low-intensity eavesdropping bursts, rendering fixed thresholds ineffective. This paper investigates physics-aware temporal feature engineering for machine learning-based anomaly detection in entanglement-based BBM92 QKD telemetry. A 24-dimensional feature space is computed over a 30-second sliding window, characterizing the temporal shape and cross-observable correlation of QBER, the Bell S parameter, and photon coincidence rates. Evaluated on a simulated FSO telemetry dataset spanning 24 hours across five random seeds, static QBER thresholding achieves only 17.3% recall against blended sub-threshold attacks. An XGBoost classifier trained on the proposed feature set achieves 96.9% recall and 97.6% precision in the same simulated setting. SHAP analysis indicates that for stealthy attacks, detection is driven not by the absolute error magnitude, but by anomalous temporal cross-correlations between independent quantum observables. Integrating physics-aware temporal features into QKD monitoring pipelines may provide a more robust defense against dynamic adversarial conditions, though validation on operational hardware remains an important open problem.**

***Index Terms*—Quantum Key Distribution, BBM92, Machine Learning, Anomaly Detection, Free-Space Optics, Eavesdropping Detection, Telemetry**

I. Introduction

Quantum Key Distribution (QKD) enables two parties to share a cryptographic key whose security is grounded in the laws of quantum mechanics [1]. The entanglement-based BBM92 protocol [2] achieves this by measuring photon pairs in a shared Bell state ($|\Phi^+\rangle$). Although theoretically secure, practical implementations are vulnerable to physical-layer attacks that exploit device imperfections. Established attack strategies include intercept-resend [3], faked-state injection [4], and detector blinding [5].

The standard operational countermeasure is static QBER thresholding: the protocol aborts if the QBER exceeds a theoretical limit (typically 11%). In Free-Space Optical (FSO) channels, scintillation, thermal gradients, and precipitation cause natural, transient QBER fluctuations. A sophisticated adversary can exploit this by timing short bursts of low-intensity eavesdropping to coincide with periods of elevated atmospheric noise, keeping the aggregate QBER below the abort threshold. Static thresholding is structurally blind to this threat vector.

Machine learning (ML) has been applied to QKD security to complement or replace static thresholds [6], [7]. Prior work has predominantly analyzed static parameter snapshots in prepare-and-measure protocols such as CV-QKD, however, and does not directly address the coincidence-based telemetry of entanglement protocols such as BBM92.

This work investigates whether a physics-motivated temporal feature framework, applied to BBM92 telemetry, can detect blended sub-threshold attacks that evade standard QBER thresholding. The central contributions, evaluated within a simulated environment, are: (i) a 24-dimensional temporal feature space capturing the shape and cross-correlation of quantum observables over a sliding window; (ii) an empirical comparison of classifiers on this feature space; and (iii) SHAP-based analysis of which physical features drive detection, particularly for stealthy attack modes.

II. Related Work

Quantum key distribution has evolved substantially over the past three decades, with advances spanning protocol design, security proofs, network architectures, and satellite-based implementations [8]. The security of QKD has been extensively studied since the introduction of the foundational BB84 [1] and BBM92 [2] protocols. Relevant prior work may be organized into three broad areas.

A. Physical Attack Characterization

Fuchs *et al.* [3] rigorously quantified the QBER bound for the intercept-resend attack. Lydersen *et al.* [5] demonstrated that hardware-level side channels—specifically blinding single-photon avalanche diodes (SPADs) with bright illumination—can allow an adversary to control measurement outcomes. Despite countermeasures such as measurement-device-independent QKD (MDI-QKD) [9], faked-state attacks [4], [10] show that implementation-level vulnerabilities remain a persistent concern.

B. Machine Learning for Quantum Security

Mao *et al.* [6] proposed ML-based defenses for continuous-variable QKD; Liu *et al.* [7] more recently applied deep anomaly detection to identify active attacks on QKD receivers. Both approaches primarily analyze instantaneous parameter snapshots in prepare-and-measure settings. The work here diverges by targeting BBM92 telemetry and treating temporal cross-correlations of independent observables as an important detection signal.

C. Multivariate Temporal Anomaly Detection

The feature framework developed here draws motivation from industrial process monitoring, where detecting anomalies through the breakdown of inter-variable correlations has been shown to outperform univariate thresholding [11], [12]. Translated to the quantum domain, a QBER elevation that is decoupled from concurrent FSO path loss [13] is more consistent with adversarial perturbation than with atmospheric turbulence. XGBoost [14] is selected as the primary classifier for its compatibility with the TreeSHAP explainer [15], which enables post-hoc attribution of detection decisions to physically interpretable features.

III. System Model and Feature Engineering

A. BBM92 Telemetry and Threat Model

A hybrid FSO/fibre BBM92 link operating at 1 Hz telemetry resolution is modeled. The monitored observables are: QBER (Q), CHSH Bell parameter (S), coincidence pair rate (C), single-detector click rate (D), Hong-Ou-Mandel (HOM) visibility (V), and optical path loss (L). The atmospheric background simulates log-normal scintillation, diurnal thermal gradients, and Poisson-arrival burst noise.

Four attack vectors are simulated:

1) **Intercept-Resend (IR):** High-intensity interception driving QBER well above the abort threshold.
2) **Detector Blinding:** Bright-light injection forcing SPADs into a linear regime, producing anomalous single-click rates.
3) **Man-in-the-Middle (MitM):** Complete source replacement collapsing the Bell S parameter toward the classical bound (≈ 2.0).
4) **Blended Sub-Threshold:** 30-second low-intensity interception bursts (Eve fraction $\eta \approx 12\%$) timed to coincide with naturally elevated atmospheric loss, limiting QBER excursions below the abort threshold.

B. Temporal Feature Engineering Framework

Static thresholds operate on instantaneous telemetry. To capture the temporal signature of adversarial interference, the 1 Hz raw telemetry $\mathbf{x}_t$ is transformed into a 24-dimensional feature vector over a sliding window $T = 30$ s. Features are grouped by physical rationale:

1) **QBER Statistics (6 features):** Mean, standard deviation, first-difference delta, skewness, and autoregressive lag coefficients AR(1) and AR(5). Attacks tend to introduce right-skewed, serially correlated step changes, whereas atmospheric turbulence induces smooth, mean-reverting drift.
2) **Bell S Entanglement (5 features):** Mean, standard deviation, rate of change, fractional duration below the CHSH violation boundary ($S < 2\sqrt{2}$), and Pearson correlation between S and Q.
3) **Coincidence Dynamics (5 features):** Mean coincidence rate, fractional drop from peak, coefficient of variation (burstiness), and the detection-to-coincidence ratio $R_{dc} = \bar{D}/(2\bar{C})$, which is elevated during detector blinding as single clicks decouple from correlated pair arrivals.
4) **Temporal Structure (4 features):** Rolling variance, Bell S range, burst energy (ℓ_2 norm of deviations from the window mean), and rising/falling edge asymmetry. Atmospheric events tend to exhibit symmetric onset and decay envelopes; attack insertions exhibit abrupt edges.
5) **Cross-Channel Coupling (4 features):** Mean loss, mean HOM visibility, joint entanglement-rate product, and a QBER-loss decoupling metric δ_{LQ}. A positive δ_{LQ} indicates QBER elevation that is inconsistent with the concurrently measured FSO path loss.

IV. Experimental Methodology

The simulator generates 86,400 s (24 hours) of telemetry per random seed, incorporating natural diurnal weather patterns and randomly inserted attacks. Five independent random seeds are used to assess statistical robustness with respect to weather timing and attack placement. The overall attack class fraction is approximately 13%.

A temporal 75/25 train-test split is employed: the first 18 hours serve as the training set and the final 6 hours as the holdout evaluation set, preserving the temporal ordering of the data. Features are standardized with zero mean and unit variance (`StandardScaler`). Three classifiers are evaluated: Logistic Regression (LR), to assess the linear separability of the engineered feature space; Random Forest (RF); and XGBoost. The primary evaluation metric is the F2-score, which weights recall more heavily than precision to reflect the asymmetric cost of missed eavesdropping events.

V. Results and Discussion

A. Classifier Comparison

Table I reports classifier performance against the traditional static QBER threshold.

TABLE I
Classifier Comparison (Mean ± Std over 5 Seeds)

Model	Recall	Precision	F2 Score
QBER Threshold	0.173 ± 0.024	1.000 ± 0.000	0.207 ± 0.028
Logistic Regression	0.975 ± 0.004	0.957 ± 0.007	0.971 ± 0.004
Random Forest	0.962 ± 0.008	0.976 ± 0.007	0.965 ± 0.008
XGBoost	**0.969 ± 0.007**	**0.976 ± 0.010**	**0.970 ± 0.007**

The static QBER threshold achieves only 17.3% recall, which is expected given that blended attacks are specifically

designed to evade it. All three ML classifiers achieve F2-scores above 0.96. Logistic Regression reaches near-parity with the nonlinear models (F2 = 0.971 vs. 0.970 for XGBoost), suggesting that the 24-dimensional temporal transformation renders the class boundary largely linearly separable in this simulated setting. XGBoost's marginally higher precision (0.976 vs. 0.957 for LR) may reduce operational false-positive rates, and its compatibility with TreeSHAP supports post-hoc auditing. McNemar's test confirms a statistically significant difference between XGBoost and the static threshold ($p < 0.001$).

B. Temporal Window Ablation

Table II reports XGBoost performance across sliding window durations.

TABLE II
WINDOW SIZE ABLATION (XGBOOST, MEAN ± STD OVER 5 SEEDS)

Window Size	Recall	F2 Score
15 seconds	0.979 ± 0.005	0.977 ± 0.005
30 seconds	**0.968 ± 0.007**	**0.970 ± 0.007**
60 seconds	0.955 ± 0.013	0.946 ± 0.021
120 seconds	0.926 ± 0.023	0.922 ± 0.023
300 seconds	0.886 ± 0.030	0.878 ± 0.027

Performance degrades monotonically as the window grows beyond 30 s, which is consistent with longer windows diluting transient attack signatures within larger blocks of benign telemetry. The 15-second window achieves slightly higher aggregate recall in simulation; the 30-second window was nonetheless selected because it matches the duration of the simulated blended attack bursts, affording greater physical interpretability. Shorter windows may generalize less reliably to attack variants with different burst durations.

C. Sensitivity to Eavesdropper Fraction

Table III reports XGBoost recall as the Eve interception fraction η varies across attack types.

TABLE III
EVE FRACTION SENSITIVITY (XGBOOST RECALL)

Eve Frac.	Intercept-Resend	Blinding	Blended Sub-Threshold
0.05	0.945	0.945	0.615
0.10	**0.979**	**0.945**	**0.953**
0.15	0.998	0.930	0.970
0.30	0.999	0.999	0.965

At $\eta = 0.05$, blended attack recall drops to 0.615, indicating that very low-intensity interceptions produce temporal signatures difficult to distinguish from atmospheric fluctuations under this channel model. Recall climbs to 0.953 at $\eta = 0.10$, placing an approximate detection threshold near 10% interception in the simulated environment. The non-monotonic Blinding recall at $\eta = 0.15$ (0.930) compared to $\eta = 0.10$ (0.945) likely reflects seed-level variance rather than a systematic effect, given the overlapping standard deviations reported in Table I.

D. Physical Interpretability via SHAP

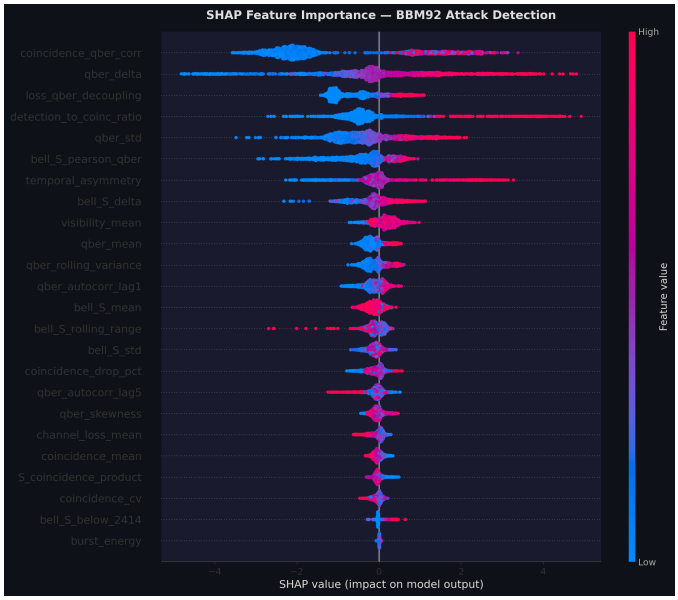

Fig. 1. SHAP summary plot showing mean absolute feature attribution for the XGBoost model. For blended sub-threshold attacks, temporal cross-correlation features rank highest; for continuous high-intensity attacks, the detection-to-coincidence ratio dominates.

Mean absolute SHAP values for the XGBoost model are shown in Fig. 1. For continuous high-intensity attacks (Intercept-Resend, MitM, Detector Blinding), the dominant feature is `det_coinc_rat` (Detection-to-Coincidence Ratio), with SHAP magnitude >3.1. This is physically consistent: active blinding and high-intensity interception severely decouple single-detector clicks from correlated photon-pair arrivals.

For Blended Sub-Threshold attacks, `det_coinc_rat` ranks fourth. Detection is instead driven by three temporal features:

1) `coinc_qber_r`: Pearson cross-correlation between coincidence rate and QBER within the window.
2) `qber_delta`: Step-change magnitude of the error rate between window epochs.
3) `temp_asym`: Asymmetry between the rising and falling edges of the QBER profile.

In the simulation, stealthy blended attacks are characterized not by absolute QBER level, but by an anomalous temporal decoupling between quantum observables that would ordinarily covary under atmospheric driving. This reliance on cross-observable structure rather than single-variable magnitude is consistent with the near-linear separability observed in the classifier comparison. Critically, the SHAP attributions align with the physical motivations for each feature group, providing some evidence that the model is exploiting genuine physical relationships rather than artefacts of the particular atmospheric noise realization used in training.

VI. LIMITATIONS

The following limitations apply to this study and should be considered when interpreting the results.

- **Simulation scope:** All results are derived from a simulated FSO channel. The atmospheric model assumes log-normal scintillation and does not account for severe deep-fade regimes typical of links exceeding 100 km. Generalization to other channel conditions is unknown.
- **Absence of hardware validation:** No operational BBM92 testbed or publicly available hardware-derived telemetry dataset was available. The reported performance figures apply within the simulated environment and may not transfer to physical systems with real optical impairments, dark counts, or timing jitter.
- **Telemetry access requirements:** Computing the full 24-feature set requires 1 Hz access to raw coincidence and single-click rates. These observables are often abstracted or aggregated in commercial QKD systems, which could limit deployment compatibility.
- **Fixed attack parameterization:** The blended attack model uses a fixed burst duration of 30 s. Performance against adaptive adversaries employing variable burst lengths, inter-arrival timing strategies, or adversarially tuned η profiles is not characterized and represents important future work.
- **Dataset circularity:** The 30-second window was selected partly because it matches the simulated burst duration. Evaluations on attack variants with different burst lengths would provide stronger evidence of generalization.

VII. Conclusion

This paper examined a physics-aware temporal feature engineering approach for eavesdropping detection in a simulated BBM92 QKD system. Moving from instantaneous QBER thresholding to a 24-dimensional sliding-window representation raises recall on blended sub-threshold attacks from 17.3% to 96.9% in simulation. The near-equivalent performance of Logistic Regression and XGBoost suggests the temporal feature transformation contributes substantially to classification performance, independent of the specific classifier employed. SHAP analysis indicates that stealthy attack detection is driven by anomalous temporal cross-correlations between quantum observables, consistent with the physical design rationale for the feature set. These findings add to the growing body of work on ML-assisted QKD security monitoring, and suggest that temporal telemetry analysis merits further investigation in hardware-validated settings.

References

[1] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," in *Proc. IEEE Int. Conf. on Computers, Systems and Signal Processing*, 1984, pp. 175–179.

[2] C. H. Bennett, G. Brassard, and N. D. Mermin, "Quantum cryptography without Bell's theorem," *Phys. Rev. Lett.*, vol. 68, no. 5, pp. 557–559, 1992.

[3] C. A. Fuchs, N. Gisin, R. B. Griffiths, C.-S. Niu, and A. Peres, "Optimal eavesdropping in quantum cryptography," *Phys. Rev. A*, vol. 56, no. 2, p. 1163, 1997.

[4] V. Makarov and D. R. Hjelme, "Faked states attack on quantum cryptosystems," *J. Mod. Opt.*, vol. 52, no. 5, pp. 691–705, 2005.

[5] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, "Hacking commercial quantum cryptography systems by tailored bright illumination," *Nat. Photonics*, vol. 4, no. 10, pp. 686–689, 2010.

[6] Y. Mao, W. Huang, H. Zhong, Y. Wang, H. Qin, Y. Guo, and D. Huang, "Detecting quantum attacks: a machine learning based defense strategy for practical continuous-variable quantum key distribution," *New Journal of Physics*, vol. 22, no. 8, p. 083073, 2020.

[7] J. Liu, B. Huang, J. Su, Q. Peng, and A. Huang, "Deep anomaly detection for active attacks on the receiver in quantum key distribution," *Opt. Express*, vol. 33, no. 22, pp. 47 137–47 150, 2025.

[8] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta, D. Bunandar, R. Colbeck, D. Englund, T. Fox *et al.*, "Advances in quantum cryptography," *Adv. Opt. Photonics*, vol. 12, no. 4, pp. 1012–1236, 2020.

[9] H.-K. Lo, M. Curty, and B. Qi, "Measurement-device-independent quantum key distribution," *Phys. Rev. Lett.*, vol. 108, no. 13, p. 130503, 2012.

[10] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer, and V. Makarov, "Full-field implementation of a perfect eavesdropper on a quantum cryptography system," *Nat. Commun.*, vol. 2, p. 349, 2011.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[12] Z. Ge and Z. Song, "Multivariate statistical process monitoring using an improved independent component analysis," *Chem. Eng. Res. Des.*, vol. 91, no. 5, pp. 855–869, 2013.

[13] R. Bedington, J. M. Arrazola, and A. Ling, "Progress in satellite quantum key distribution," *npj Quantum Inf.*, vol. 3, p. 30, 2017.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.