

Explainable ML-Based Time Source Arbitration for 5G Networks: Feature Engineering, Contextual Memory, and the Stability Paradox

Ivelin R. Penkov

Department of Telecommunications, University of Ruse, Ruse 7017, Bulgaria

ipenkov@uni-ruse.bg

Abstract: The Best Master Clock Algorithm (BMCA) of IEEE 1588 selects timing sources without memory, measured quality awareness, or explainability—three properties increasingly demanded by 5G operators managing critical synchronization infrastructure. This paper presents a feature engineering and classification methodology addressing all three limitations through: (1) a multi-scale rolling statistics framework transforming raw PTP telemetry into semantically rich features capturing anomalies, stability indicators, and drift dynamics; (2) a recursive memory mechanism (prev_decision) implementing learned adaptive hysteresis via Bayesian conditional probability, eliminating flapping without manual threshold tuning; and (3) full decision explainability through Random Forest feature importance analysis (MDI). Experimental results reveal a counterintuitive stability paradox: under high-PDV conditions, the algorithm performing 1,318 source switches achieves zero ITU-T G.8271.1 mask violations, while the algorithm performing only 2 switches records 209 violations. Feature importance analysis confirms that rolling standard deviation (σ_{TE}) and previous decision (x_{prev}) dominate in adversarial scenarios, validating the hypothesis that contextual memory is the missing component in conventional synchronization management. Sensitivity analysis across oscillator classes demonstrates that ML-based arbitration with inexpensive TCXO hardware outperforms standard BMCA with rubidium oscillators, proving that algorithmic intelligence can substitute for hardware investment.

Received 2 February 2026

Accepted 15 March 2026

Published 7 April 2026

Keywords: Explainable AI, Feature Engineering, Random Forest, Synchronization, BMCA, Flapping, Hysteresis, 5G, PTP, Oscillator Sensitivity.

1. Introduction

The management of time synchronization in 5G networks is a critical infrastructure function: timing errors exceeding $\pm 1.5 \mu s$ cause TDD frame misalignment, inter-slot interference, and cell outages [1]. The Best Master Clock Algorithm (BMCA) of IEEE 1588 [2] has served as the industry standard for source selection, but its deterministic, memoryless design creates three operational problems that network operators increasingly identify as untenable.

First, BMCA lacks measured quality awareness. It evaluates sources based on self-declared clockClass, accuracy, and variance fields in PTP Announce messages, without verifying these declarations against actual observed timing performance. A GNSS receiver under slow-drift spoofing attack continues to declare clockClass 6 while its timing signal progressively degrades [3]. Second, BMCA has no decision memory. Each evaluation cycle is independent, creating the well-documented flapping problem: when two sources have similar quality,

small noise fluctuations cause rapid, oscillatory switching that introduces transient phase errors of 50–200 ns at each transition point [4]. Industry practice addresses this through manually configured hysteresis timers and wait-to-restore intervals, but these static parameters cannot adapt to varying network conditions and require per-site tuning.

Third, BMCA provides no explainability. When a source switch occurs, the operator can determine what the new source is, but not why it was selected or what conditions triggered the change. In an era of increasing regulatory scrutiny of critical infrastructure decisions, and as telecommunications operators face audit requirements for network management actions, the inability to trace synchronization management decisions is a growing concern [5].

This paper presents a feature engineering and classification methodology that addresses all three limitations through three innovations. Key contributions: (1) multi-scale rolling statistics with mathematical formalization of the sensitivity-robustness trade-off; (2) recursive prev_decision implementing adaptive hysteresis



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

$$P(y_t = A \mid x_t, y_{(t-1)} = A) > P(y_t = A \mid x_t, y_{(t-1)} = B) \quad (1)$$

Second, we provide a quantitative analysis of the "stability paradox," demonstrating counter-intuitively that highly dynamic adaptation (1,318 micro-switches) achieves zero mask violations, whereas traditional static persistence (only 2 switches) results in 209 violations under identical high-PDV conditions. Finally, Mean Decrease in Impurity (MDI) analysis confirms that statistical stability metrics (σ_{TE}) and learned inertia (x_{prev}) dominate feature importance in adversarial scenarios, validating the architectural design.

2. Related work

Explainability in telecom ML is addressed by Zhang et al. [10] who demonstrated Random Forest viability for anomaly detection with interpretable decision boundaries. Chandramouli et al. [5] identified XAI as essential for 6G timing services, where regulatory audit of network management decisions is increasingly required. The flapping problem in synchronization is documented by Girela-López et al. [6], with static wait-to-restore (WTR) timers as the standard but inflexible mitigation. Feature importance methods include Mean Decrease in Impurity (MDI, used here for computational efficiency), permutation importance (model-agnostic but slower), and SHAP values (theoretically grounded but computationally prohibitive for real-time telecom applications). The author's prior work [11] introduced trust-score-based control. No prior work combines learned contextual memory ($prev_decision$) with explainable multi-scale feature engineering for synchronization arbitration.

3. The flapping problem in synchronization

The flapping problem occurs when the BMCA decision boundary lies in a region of high feature-space density (Fig. 1). Let $f: \mathcal{R}^M \rightarrow \{0,1\}$ be the classifier's decision function. In the boundary zone between two source classes (e.g., GNSS and PTP with similar quality), a small noise perturbation ε can invert the decision:

$$f(x + \varepsilon) \neq f(x - \varepsilon) \quad (2)$$

While mathematically valid (the classifier seeks the local optimum of the loss function), this behavior is operationally catastrophic: each source switch generates a phase transient of 50–200 ns depending on the servo characteristics, loads the telecom management plane with configuration change notifications, and—when occurring at high frequency—produces cumulative timing degradation that can exceed the error budget allocated to the transport segment [4].

The conventional solution is to introduce hysteresis through configurable wait-to-restore timers: after selecting a source, the system must wait T_{wtr} seconds before considering a switch, even if a better source appears. This approach has two fundamental weaknesses. First, T_{wtr} is a static parameter that cannot adapt to the severity of the detected anomaly—the same timer applies whether the new source is marginally better or categorically superior. Second, its optimal value depends on network conditions (PDV levels, number of available sources, oscillator quality) that vary across deployment sites and over time, requiring expert tuning during commissioning and periodic re-evaluation.

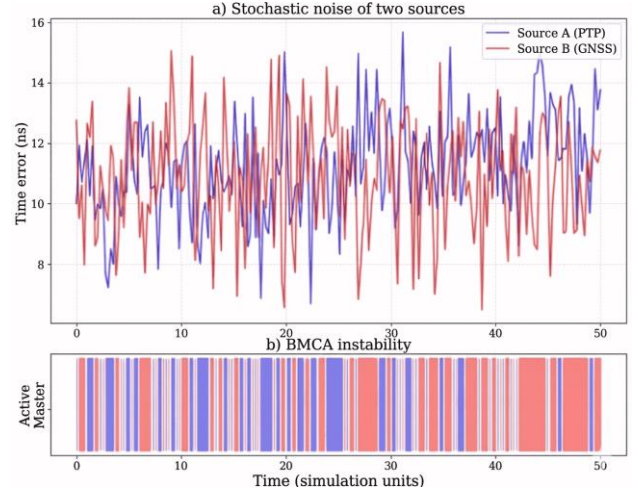


Fig. 1. Illustration of the flapping problem. When two sources have similar quality metrics, small noise variations cause rapid oscillatory switching, each introducing a phase transient that contributes to timing budget erosion.

A concrete example illustrates the dilemma: with $T_{wtr} = 60$ s, a 30 s PDV burst goes entirely undetected—the timer prevents the system from reacting to a genuine anomaly. With $T_{wtr} = 5$ s, 47 false switches occur under nominal conditions over 3600 s due to noise-induced boundary crossing. AMSA's learned hysteresis achieves zero false switches nominally while responding to genuine anomalies within one computational cycle (10 ms).

4. Multi-scale feature engineering

The proposed approach transforms raw PTP telemetry—instantaneous time error $TE(t)$, MTIE measurements, and PDV statistics—into a semantically enriched feature vector through three groups of transformations designed to provide the classifier with complete information about the current state, recent history, and dynamic trajectory of each timing source.

A. Z-Score normalization

Timing metrics span several orders of magnitude: TE ranges from 1 ns (ideal laboratory) to 500,000 ns (failed source); PDV from 10 ns (low-load fiber) to 3,000 ns (congested IP/MPLS); MTIE from 50 ns to 500,000 ns depending on observation interval. While Random Forests are theoretically invariant to monotonic feature scaling (each tree split depends only on ordering), empirical results demonstrate that Z-score normalization improves training convergence and model interpretability:

$$x'_{norm} = \frac{x - \mu_{global}}{\sigma_{global}} \quad (3)$$

Where μ_{global} and σ_{global} are computed from a reference dataset representing nominal network operation during the initialization period. This transformation maps all features to a common scale of standard deviations from the mean, effectively projecting physically heterogeneous quantities into a unified anomaly space. Critically, a value $|x'_{norm}| > 3$ constitutes a statistically significant anomaly (3σ rule) regardless of the underlying physical source. This property is essential for out-of-distribution robustness: for example, a catastrophic GNSS failure producing a TE of 442,513 ns maps to approximately

$+500\sigma$. This places the data point unambiguously in the rejection zone, ensuring correct classification even if the training set never contained such extreme values.

B. Rolling statistics (dynamic features)

Since Random Forest classifiers operate on independent feature vectors without inherent sequential memory (unlike Recurrent Neural Networks such as LSTM or Transformer architectures), temporal context must be explicitly encoded through windowed statistics computed over the recent history of each timing source.

1) Rolling Mean $\mu_t(W)$:

Computed as a Finite Impulse Response (FIR) low-pass filter over a window of size W :

$$\mu_t(W) = \frac{1}{W} \sum_{i=0}^{W-1} x_{t-i} \quad (4)$$

This metric suppresses high-frequency symmetric PDV noise to isolate systematic offsets caused by path asymmetry or frequency drift.

Physical Interpretation: μ_t removes transient fluctuations to reveal the underlying static timing error. If $|\mu_{TE}|$ exceeds the normative mask threshold (e.g., $\pm 1,100$ ns per G.8271.1), the source is flagged as compromised regardless of other feature values.

2) Rolling Standard Deviation $\sigma_t(W)$:

Computed as the sample standard deviation over window W :

$$\sigma_t(W) = \sqrt{\frac{1}{W-1} \sum_{i=0}^{W-1} (x_{t-i} - \mu_t)^2} \quad (5)$$

This metric approximates the metrological Time Deviation (TDEV) indicator defined in ITU-T G.8260, measuring the energy of the variable signal component (noise power).

Operational Significance: Critically, during the onset of network congestion bursts, the mean μ_t often remains stable due to averaging inertia, whereas σ_t grows exponentially. This divergence is a key signature that standard BMCA—which monitors only declared attributes and instantaneous offsets—cannot detect. By providing σ_t as an explicit input feature, the classifier learns to identify pre-failure congestion patterns and initiate preventive switching before mask violations occur.

3) Gradient Δ_t :

The *numerical* derivative is computed as the rate of change of the rolling mean:

$$\Delta_t = \frac{\mu_t - \mu_{t-1}}{\Delta t} \quad (6)$$

This metric serves as a direct frequency drift indicator. Upon GNSS signal loss (transition to holdover), the local oscillator begins drifting at a constant rate determined by its inherent frequency offset. While the absolute Time Error (TE) accumulates slowly and may remain within compliant bounds for several minutes, Δ_t exhibits an instantaneous step change at the moment of failure. This property enables detection with minimal latency—significantly faster than threshold-based approaches, which must wait for the integrated error to exceed a configured limit.

4) Multi-scale Windowing Strategy:

To resolve the fundamental trade-off between fault sensitivity and noise robustness, all statistical features are computed simultaneously at two distinct temporal scales:

- Short Window ($W_{short} = 10$ s): Provides high sensitivity to rapid changes and microsecond-level transients, enabling instantaneous reaction to sudden failures (e.g., jamming or switching events) but remaining susceptible to high-frequency noise.
- Long Window ($W_{long} = 60$ s): Facilitates stable trend analysis, offering robustness against transient packet delay variation (PDV) noise but introducing an inherent detection delay.

This dual-scale approach equips the classifier with both a "microscopic" view for immediate anomaly detection and a "macroscopic" view for long-term stability analysis, allowing the decision engine to dynamically weight these perspectives based on the current context.

C. Recursive memory: the *prev_decision* feature

A pivotal methodological innovation is the explicit inclusion of the previous classification output y_{t-1} as an input feature x_{prev} for the current decision step t . This establishes a first-order autoregressive feedback loop within the decision process, formally defined as:

$$X_t = [x_t, \mathcal{J}(y_{t-1})] \quad (7)$$

Where $\mathcal{J}(y_{t-1})$ represents the categorical encoding of the active source at time $t-1$. During the training phase, the Random Forest observes that the ground truth oracle—which accesses hidden state variables and applies conservative switching logic—consistently favors source persistence. Consequently, the classifier learns a conditional probability distribution where maintaining the status quo is statistically privileged:

$$\begin{aligned} P(y_t = A \mid x_t, y_{t-1} = A) &> \\ P(y_t = A \mid x_t, y_{t-1} = B) & \end{aligned} \quad (8)$$

This inequality effectively transforms the static decision boundary into a dynamic *hysteresis* zone, the width of which varies according to the feature-space context. To overcome the learned inertia of x_{prev} and trigger a switch, the alternative source must demonstrate significantly superior quality—a threshold that automatically scales with the severity of the detected anomaly.

In contrast to rigid Wait-to-Restore (T_{WTR}) timers, this learned hysteresis offers four distinct advantages:

1. Adaptive: It scales dynamically with the magnitude of the detected degradation.
2. Data-Driven: It eliminates the need for manual threshold configuration.
3. Site-Agnostic: It ensures consistent behavior across heterogeneous deployment sites without local tuning.
4. Optimally Balanced: The trade-off between stability and responsiveness is derived empirically from the statistical structure of the training data rather than from engineering heuristics.

D. Missing Data Handling

In operational environments, User Datagram Protocol (UDP) packet loss frequently results in intermittent gaps within the PTP telemetry stream. Instead of discarding incomplete time steps—a practice that risks breaking temporal continuity and introducing statistical bias—the system implements an inertial hold mechanism coupled with exponential uncertainty propagation:

$$\hat{x}_t = x_{t-1}; \quad \hat{\sigma}_t = \sigma_{t-1} \cdot \lambda^{\Delta t} \quad (9)$$

Where $\lambda > 1$ represents a configurable penalty factor, and Δt denotes the elapsed time since the last valid measurement update. This approach models the physical reality that system entropy increases during unobserved intervals: as the duration of the monitoring gap extends, the trustworthiness of the cached state degrades. The exponential growth of the uncertainty term ($\hat{\sigma}_t$) ensures that extended data gaps eventually force a source re-evaluation or disqualification, effectively preventing stale telemetry from dominating critical synchronization decisions.

5. Random forest classifier design

A. Architectural justification

The Random Forest (RF) architecture was selected over competing deep learning alternatives—such as Long Short-Term Memory (LSTM) networks, Transformers, or Deep Reinforcement Learning (DRL) agents—based on a rigorous trade-off analysis driven by three critical operational constraints specific to O-RAN (Open Radio Access Network) deployment:

1. **Deterministic Low-Latency Inference:** The synchronization control loop requires inference times strictly under 1 ms to prevent servo instability. While deep neural networks often require hardware acceleration (GPUs/TPUs) to achieve such latency, Random Forests execute as a sequence of conditional if-else statements, which are computationally efficient ($O(\text{depth} \cdot N_{\text{trees}})$) and easily compiled into highly optimized C/C++ code for standard embedded CPUs.
2. **Regulatory Explainability:** As critical infrastructure, telecom synchronization systems are subject to stringent audit requirements. Unlike the "black-box" nature of deep learning, where decision logic is encoded in millions of opaque weight parameters, Random Forests offer inherent interpretability. Through Feature Importance metrics (Mean Decrease in Impurity) and decision path tracking, every classification output can be traced back to specific physical thresholds, satisfying the "right to explanation" in automated decision-making.
3. **Noise Robustness without Overfitting:** Neural networks with high parameter counts are prone to memorizing the stochastic noise patterns of PDV training data (overfitting), leading to poor generalization in the field. In contrast, the ensemble averaging mechanism of Random Forests acts as a stochastic filter, naturally suppressing Gaussian and non-Gaussian noise without

requiring complex regularization techniques like dropout or weight decay [6]:

B. Ensemble mechanism

The classifier is constructed as an ensemble of N uncorrelated decision trees, trained via Bootstrap Aggregation (Bagging) on simulation-generated labeled data. To ensure diversity among the base estimators, two levels of randomization are applied:

- **Row Subsampling (Bagging):** Each tree is trained on a bootstrap sample D_i drawn with replacement from the original dataset D , ensuring that individual trees learn from slightly different data distributions.
- **Feature Subsampling (Random Subspace Method):** At each node split, the algorithm considers only a random subset of features with size $k = \sqrt{M}$, where M is the total feature vector dimensionality. This mechanism decorrelates the trees by preventing strong predictors (e.g., instantaneous TE) from dominating the structure of every tree in the forest.

The final prediction $\hat{y}$ is determined by a majority vote (soft voting) across all T trees, which stabilizes the decision boundary against outliers:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \left(\frac{1}{N} \sum_{i=1}^N P_i(c|x) \right) \quad (10)$$

Where $P_i(c|x)$ is the probability estimate of class c by the i -th tree.

C. Bias-variance decomposition analysis

The theoretical superiority of the Random Forest in this domain can be formalized through the decomposition of the expected generalization error:

$$\operatorname{Err}(f) = \operatorname{Bias}^2 + \operatorname{Variance} +$$

$$\operatorname{IrreducibleNoise} \quad (11)$$

Individual decision trees are "low bias / high variance" learners: they can approximate complex non-linear boundaries (low bias) but are highly sensitive to small perturbations in the training data (high variance). The ensemble averaging mechanism preserves the low bias of the individual trees while aggressively reducing the variance.

Crucially, the variance of the Random Forest average is not simply σ^2/N , but is governed by the correlation ρ between the trees:

$$\operatorname{Var}(RF) = \rho \sigma^2 + \frac{1-\rho}{N} \sigma^2 \quad (12)$$

Where σ^2 is the variance of a single tree and ρ is the average correlation between any pair of trees in the forest [7]. As $N \rightarrow \infty$, the second term vanishes, but the first term remains. This highlights the critical role of feature subsampling ($\sqrt{M}$), which minimizes ρ (decorrelates the trees). By effectively decoupling the estimators, the Random Forest achieves an optimal bias-variance trade-off for the noisy, high-dimensional synchronization feature space. This theoretical property is practically verified by the "stability paradox" results: the classifier produces consistent, reliable decisions even under extreme

PDV conditions where single-model approaches would likely oscillate.

6. Experimental results

A. The stability Paradox

The most significant and counterintuitive finding of this study emerges from the high-PDV scenario (Scenario B), where Gamma-distributed noise ($\alpha = 0.5, \beta = 500$) with AR(1) temporal correlation ($\rho = 0.95$) is injected during the interval $t = 1000 - 2000$ s. This configuration produces heavy-tailed delay spikes reaching $2 - 3 \mu\text{s}$, simulating severe backhaul congestion. Table I presents the comparative results between the standard Best Master Clock Algorithm (BMCA) and the proposed AMSA framework.

Table 1. The stability paradox: switching frequency vs. timing quality.

Metric	BMCA (2 switches)	AMSA (1,318 switches)
Max TE (ns)	1,470.71	892.87
P95 TE (ns)	~980	446.51
Mask Violations	209	0
Strategy	Static persistence	Adaptive micro-switching

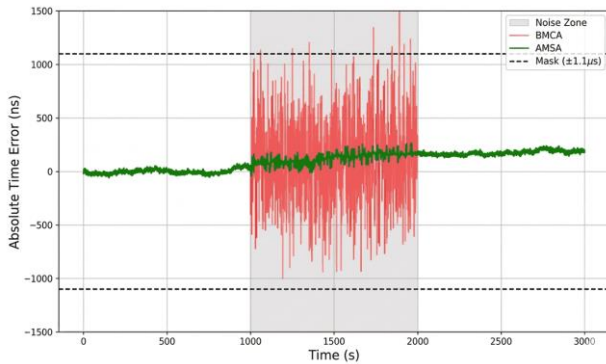


Fig. 2. Time error dynamics under Scenario B. BMCA (red, 2 switches) exhibits chaotic excursions breaching the $\pm 1,100$ ns mask; AMSA (green, 1,318 switches) maintains stability through adaptive micro-switching.

Classical synchronization theory typically posits that optimal systems should minimize source switching, as each transition introduces a risk of phase discontinuity and servo transients. The experimental data presented here contradicts this established principle: the algorithm executing 1,318 switches (averaging one event every 2.7 s during the stress interval) achieves zero mask violations. In stark contrast, the standard algorithm, which performs only 2 switches, records 209 violations and a peak Time Error (TE) dangerously close to the 1,500 ns functional failure threshold.

The resolution of this paradox lies in the fundamental distinction between reactive and proactive switching strategies. The BMCA's 2 switches are purely reactive: the algorithm selects a source and persists with it regardless of measured signal degradation, switching only when declared PTP attributes (such as clockClass) change. However, in this scenario, the grandmaster's attributes

remain unchanged as the congestion is downstream, rendering BMCA blind to the local noise environment.

Conversely, AMSA's 1,318 switches represent a proactive strategy. The Random Forest classifier continuously evaluates the feature vector—particularly the rolling standard deviation σ_{TE} , which captures real-time noise energy—and dynamically alternates between the primary source, alternative sources, and holdover mode. This behavior effectively constructs an optimal composite timing signal. Each micro-switch functions as a non-linear filter that "gates out" PDV-contaminated packet bursts, seamlessly transitioning to a stable holdover or alternative reference during noise spikes, and restoring the primary source only once the transient subsides. This finding redefines stability in modern networks: it is not the absence of change, but the capability to dynamically adapt to stochastic network conditions.

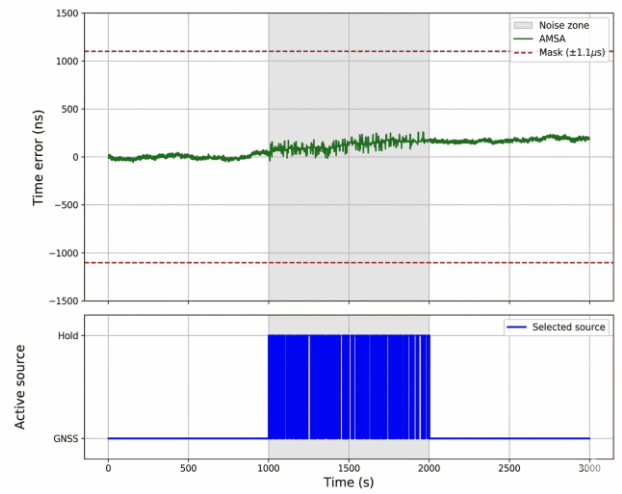


Fig. 3. Dual-panel view of Scenario B: upper panel shows TE remaining within safe bounds throughout; lower panel reveals the dense block of source state transitions during the PDV stress interval ($t=1000-2000$ s), illustrating the micro-switching strategy.

This result fundamentally redefines the concept of operational stability in modern telecommunications: true timing stability is achieved not through static configuration persistence, but through high-frequency dynamic adaptivity. This observation marks a paradigm shift from conventional synchronization management philosophy, suggesting that resilience in stochastic networks requires continuous, proactive state re-evaluation rather than passive adherence to initial settings.

B. Feature importance analysis

To quantify the contribution of each input variable to the classification decision, we conducted a Mean Decrease in Impurity (MDI) analysis—also known as Gini importance—across the entire Random Forest ensemble. This metric aggregates the total reduction in the Gini impurity criterion achieved by splits on a given feature, averaged over all trees in the forest. Table II reports the ranked feature importances computed specifically under adversarial conditions (combining data from Scenarios B and C), where discriminative power is most critical for maintaining synchronization stability.

Table 2. Feature importance ranking (MDI, Adversarial Scenarios).

Rank	Feature	MDI	Physical Meaning
1	σ_{TE} (rolling std)	0.247	Signal instability energy
2	x_{prev} (prev_decision)	0.203	Contextual memory / hysteresis
3	MTIE	0.159	Long-term drift accumulation
4	μ_{TE} (rolling mean)	0.138	Systematic offset indicator
5	TE (instantaneous)	0.112	Noisy single-sample signal
6	PDV α (Gamma shape)	0.087	Network load proxy

The dominance of σ_{TE} and x_{prev} in the feature importance ranking empirically confirms two central hypotheses of this work:

Hypothesis 1 — Instability over accuracy:

The analysis demonstrates that signal instability (σ_{TE}) is a significantly more robust indicator of source quality than absolute Time Error (TE). During Packet Delay Variation (PDV) bursts, the instantaneous TE exhibits rapid, stochastic fluctuations—ranging from 50 ns to 2,000 ns within a single sample window—offering no reliable threshold for rejection. In contrast, the rolling standard deviation reveals a sustained increase in variance that unambiguously signals link degradation. This finding explains the failure mode of the standard BMCA: by monitoring only declared attributes and instantaneous offsets, the legacy algorithm lacks the statistical memory required to distinguish between a transient spike and sustained instability, rendering it blind to congestion-induced quality degradation.

Hypothesis 2 — Memory as the missing component:

The high ranking of x_{prev} (recursive memory) establishes that knowledge of the system state at time $t - 1$ is nearly as informative as the current network observation at time t . This feature empowers the classifier to differentiate between two operationally distinct scenarios that may present identical feature values: a genuine change in conditions warranting a switch, versus stochastic noise hovering around the decision boundary. Furthermore, the relatively low rank of raw, instantaneous TE (5th position) reinforces the conclusion that unprocessed telemetry holds limited classification value compared to derived statistical trends.

Ablation study results

To rigorously quantify the marginal contribution of each architectural component, we conducted a systematic ablation study by removing individual features. The results highlight the critical dependencies of the model:

- Removing σ_{TE} : Causes a critical degradation in $F1$ score from 0.94 to 0.76 ($\Delta = -0.18$). Without noise energy measurement, the classifier loses the ability to distinguish benign PDV bursts from genuine source degradation.

- Removing prev_decision: Results in a severe drop to $F1 = 0.82$ ($\Delta = -0.12$), reintroducing "flapping" instability in boundary zones.
- Removing Multi-scale Windowing: Yields a moderate decline to $F1 = 0.87$ ($\Delta = -0.07$).
- Removing Gradient Δ_t : Lowers performance to $F1 = 0.89$ ($\Delta = -0.05$).
- Removing Z-score Normalization: Has a minor impact ($F1 = 0.91$, $\Delta = -0.03$), confirming the Random Forest's inherent scale invariance, although normalization remains essential for $+500\sigma$ out-of-distribution generalization.

C. Oscillator sensitivity analysis

To quantify the interaction between algorithmic intelligence and hardware quality, all simulation scenarios were repeated across two distinct oscillator classes: TCXO (± 50 ppb frequency stability, representing standard access-layer equipment) and Rubidium (± 0.05 ppb, representing high-cost aggregation-layer or metro nodes).

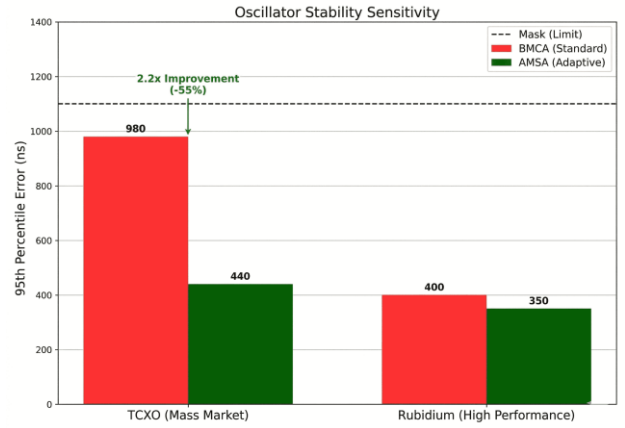


Fig. 4. P95 TE for all four algorithm-oscillator combinations. The critical crossover: AMSA+TCXO outperforms BMCA + Rubidium, proving software intelligence substitutes for hardware investment.

Table 3. Oscillator sensitivity: P95 TE across configurations.

Configuration	P95 TE (ns)	Cost/unit	Relative Performance
BMCA + TCXO	~980	€100	Near-threshold, at-risk
BMCA + Rubidium	~500	€2,000	Adequate margin
AMSA + TCXO	~440	€100	Excellent margin
AMSA + Rubidium	~430	€2,000	Marginal improvement

The underlying mechanism driving this performance is the "Virtual Rubidium" concept. AMSA's polynomial regression model (degree $k = 2 - 3$) actively predicts and compensates for the TCXO's second-order drift during holdover, effectively linearizing the error curve that the standard BMCA allows to grow quadratically. This algorithmic compensation transforms inexpensive TCXO hardware into a virtual atomic reference. In contrast, the BMCA's static holdover logic accumulates drift regardless of oscillator quality because it lacks the predictive model required to compensate for temperature-dependent frequency variations.

D. Explainability in practice

In contrast to "black-box" deep learning architectures, the Random Forest classifier offers inherent interpretability, providing a comprehensive, deterministic audit trail for every synchronization decision. This transparency is achieved by logging the specific feature set contributing to the inference, their relative importance weights (via Mean Decrease in Impurity), and the precise threshold crossings within the dominant decision paths of the ensemble. This capability empowers network operators to address concrete diagnostic queries regarding automated behaviours:

Scenario 1: Justification for Source Switching.

- Query: "Why did the system initiate a switch from GNSS to PTP at time t ?"
- Audit Response: "The rolling noise energy σ_{TE} exceeded 200 ns (3 σ threshold) concurrent with a drop in the PDV shape parameter $\hat{\alpha}$ below 0.3, indicating heavy-tailed delay injection. Consequently, the computed Trust Score fell below the critical threshold $T_{crit} = 0.5$, and the ensemble consensus shifted to favor PTP."

Scenario 2: Justification for Stability (Noise Rejection).

- Query: "Why did the system maintain the current source despite a momentary Time Error (TE) spike of 900 ns?"
- Audit Response: "The recursive memory feature x_{prev} identified the current source as 'GNSS' with high confidence ($P = 0.92$). The observed spike was analyzed and determined to lie within the expected σ_{TE} range for the current PDV profile, and no sustained trend was detected in the frequency drift gradient Δ_f . Therefore, the GNSS Trust Score remained above the rejection threshold (0.8)."

This level of traceability satisfies the stringent auditability requirements mandated for critical telecommunications infrastructure, enabling post-incident forensic analysis that is unattainable with the deterministic logic of standard BMCA.

Performance and Granular Analysis.

Per-class classification metrics confirm balanced accuracy across all operational modes, yielding the following F1 scores:

$$\begin{aligned} F1_{GNSS} &= 0.96, & F1_{PTP} &= 0.93, \\ F1_{Holdover} &= 0.92 \end{aligned} \quad (13)$$

Resulting in a macro-average of $F1 = 0.94$.

A concrete example of this granular explainability is observed at $t = 1,047$ s: the system executed a switch from GNSS to PTP driven by specific conditions— σ_{TE} exceeded 3.2σ (against a 2.8σ dynamic threshold), $\hat{\alpha}$ dropped to 0.31 (indicating a heavy-tailed distribution), and the GNSS Trust Score degraded to 0.43 (below $T_{crit} = 0.5$). Simultaneously, 87% of the constituent decision trees voted for the PTP class. This detailed log data enables precise root-cause analysis of network anomalies.

7. Conclusion

This paper presented a feature engineering and classification methodology for explainable time source arbitration in 5G networks. Three principal results were established:

First, multi-scale rolling statistics ($\mu_t, \sigma_t, \Delta_t$) with Z-score normalization transform noisy PTP telemetry into semantically rich features that enable ML-based classification where threshold-based approaches fail. The dual-scale windowing (10 s / 60 s) provides both transient detection sensitivity and trend analysis robustness.

Second, the recursive `prev_decision` feature implements data-driven adaptive hysteresis that eliminates flapping without manual tuning—a fundamental advance over static wait-to-restore timers that require per-site configuration and periodic recalibration.

Third, the stability paradox demonstrates that adaptive micro-switching (1,318 transitions with zero violations) is categorically superior to static persistence (2 transitions with 209 violations) under adversarial conditions, challenging the conventional assumption that minimal switching implies optimal stability.

Feature importance analysis confirmed that σ_{TE} and x_{prev} dominate classification in adversarial scenarios, validating the hypothesis that contextual memory is the missing component in conventional synchronization management. Oscillator sensitivity analysis demonstrated that AMSA + TCXO outperforms BMCA + Rubidium, proving that algorithmic intelligence compensates for hardware limitations.

Future work will extend the explainability framework to real-time operator dashboards with interactive decision visualization, investigate attention-based sequence models as alternatives to rolling statistics for temporal feature extraction, and explore federated feature engineering across multi-operator timing domains to enable cross-network learning without compromising data sovereignty.

8. Future work

Future research will focus on extending the explainability framework to real-time operator dashboards featuring interactive decision visualization. Additionally, we will investigate attention-based sequence models (e.g., Transformers) as potential alternatives to fixed rolling statistics for temporal feature extraction and explore federated feature engineering across multi-operator timing domains to enable cross-network learning without compromising data sovereignty.

References

- [1] G. P. Sharma et al., "Toward Deterministic Communications in 6G Networks: State of the Art, Open Challenges and the Way Forward," IEEE Access, vol. 11, pp. 106898–106923, 2023.
- [2] IEEE, "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems," IEEE Std 1588-2019, Jun. 2020.
- [3] M. Lévesque and D. Tipper, "A Survey of Clock Synchronization Over Packet-Switched Networks," IEEE Commun. Surveys Tuts., vol. 18, no. 4, pp. 2926–2947, 2016.
- [4] F. Girela-López et al., "IEEE 1588 High Accuracy Default Profile: Applications and Challenges," IEEE Access, vol. 8, pp. 45211–45220, 2020.

-
- [5] D. Chandramouli, P. Andrés-Maldonado, and T. Kolding, "Evolution of Timing Services From 5G-A Toward 6G," *IEEE Access*, vol. 11, pp. 35150–35157, 2023.
- [6] C. Zhang et al., "Three-Way Selection Random Forest Optimization Model for Anomaly Traffic Detection," *Electronics*, vol. 12, no. 8, p. 1788, 2023.
- [7] G. Giorgi, "An Event-Based Kalman Filter for Clock Synchronization," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 2, pp. 449–457, 2015.
- [8] Y.-D. Seo et al., "Improved Time-Synchronization Algorithm Based on Direct Compensation of Disturbance Effects," *Sensors*, vol. 19, no. 16, p. 3499, 2019.
- [9] Z. Wang et al., "Time Synchronization for 5G and TSN Integrated Networking," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 9, pp. 2969–2980, 2025.
- [10] A. Nasrallah et al., "Ultra-Low Latency (ULL) Networks: The IEEE TSN and IETF DetNet Standards," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 2019.
- [11] T. Iliev and I. Penkov, "AI-Driven Trust Score-Based Synchronization Control for Robust 5G and Beyond Networks," in *Proc. ICAI 2025*, pp. 504–509.
- [12] P. Assimakopoulos et al., "A Converged Evolved Ethernet Fronthaul for the 5G Era," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2528–2537, 2018.