Mitigating Transaction-Cost Degradation in Financial Forecasting via Volatility-Gated Inference

Mukitu Islam Nishat
Department of Computer Science and Engineering
North Bengal International University
Rajshahi, Bangladesh
mukituislamnishat@gmail.com

***Abstract*—Deep neural forecasters for financial time series are conventionally optimized to minimize statistical loss functions such as the Mean Squared Error (MSE), yet they frequently suffer severe performance degradation when their raw predictive outputs are translated into live trading decisions. The principal culprit is not model inaccuracy per se, but the mismatch between the loss surface being minimized and the cost-laden objective that actually governs realized profit: bid–ask spreads, slippage, broker commissions, and the compounding drag of excessive portfolio turnover. In this paper we introduce the *Volatility-Gated Inference* (VGI) mechanism, a model-agnostic post-hoc policy that dynamically abstains from trade execution whenever normalized local volatility exceeds a calibrated threshold τ. We evaluate the proposal on EUR/USD daily exchange data from January 2019 through August 2026, comprising 1,980 trading days partitioned chronologically into 1,584 training and 396 out-of-sample testing days. The baseline directional model achieves a test MSE of 2.00×10^{-5}, yet its corresponding unconstrained trading strategy yields a cumulative return of -1.46% and an annualized Sharpe ratio of -0.13, evidence of capital erosion driven by over-trading. By contrast, the volatility-gated strategy, parameterized by a per-trade friction of $c = 0.0001$ and a normalized volatility threshold of $\tau = 1.0\sigma$, achieves a cumulative return of $+3.46\%$ and an annualized Sharpe ratio of $+0.36$, while simultaneously reducing trade switches from approximately 280 to 145. These results demonstrate that selectively abstaining from execution in volatile regimes can invert a negative-expectancy strategy into a profitable one without any modification to the underlying predictive model.**

***Index Terms*—Financial forecasting, transaction costs, volatility gating, foreign exchange, Sharpe ratio, deep learning, time-series AI, risk-adjusted return, average true range.**

I. Introduction

The deployment of machine learning systems in quantitative finance has accelerated markedly over the past decade, driven by simultaneous advances in deep sequence models, the availability of high-frequency market data, and the commoditization of GPU compute. Predictive tasks ranging from next-tick price movement classification to volatility forecasting are now routinely tackled by deep architectures that, on standard supervised benchmarks, outperform classical econometric baselines by substantial margins. Yet the empirical record of these systems in *production trading* is conspicuously uneven: published strategies frequently report attractive back-tested returns that fail to materialize, or even reverse sign, when implemented with realistic transaction costs and execution constraints [6]. The central claim of this paper is that a non-trivial share of this shortfall is attributable not to the predictive model itself, but to the disconnect between the loss function under which the model is trained and the cost-laden objective that ultimately determines realized profit.

a) The Statistical–Economic Loss Mismatch: Supervised learners for continuous price targets are typically trained to minimize the Mean Squared Error $\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{y}_t - y_t\right)^2$, or, for directional classification, the binary cross-entropy. These losses are insensitive to whether a correct prediction produces an actionable trade and, conversely, treat every incorrect prediction symmetrically regardless of whether the resulting trade incurs friction. In liquid retail Forex, however, the round-trip cost of a single position flip is dominated by spread and slippage that typically exceed the expected per-trade gross return of any reasonably accurate short-horizon forecaster. As a consequence, an unconstrained policy that maximizes *prediction count* will, all else equal, maximize *cost count*, and the marginal value of an additional trade is more often negative than positive.

b) Role of Slippage, Spread, and Turnover: The friction surface in foreign exchange is multi-layered: brokers quote a bid–ask spread, market impact induces slippage proportional to order size and prevailing volatility, and overnight carry or swap charges may further erode returns. While each component can be small in isolation—on the order of one to ten basis points for major pairs—their cumulative effect under high turnover is multiplicative rather than additive. A model whose per-trade edge is on the order of c is, from a trading perspective, indistinguishable from a random policy; it generates a Brownian-motion-like equity curve whose long-run drift is dominated by friction rather than signal. Excessive turnover thus creates a structural regime in which the optimal *policy* is one that trades less, even if the underlying signal is informative.

c) Contributions: In this work we address the policy layer directly. Our contributions are as follows:

- We formalize the *Volatility-Gated Inference* (VGI) mechanism as a closed-form, model-agnostic decision rule that abstains from execution during high local-volatility regimes.
- We derive the full mathematical specification of the policy: price return, the 14-day Average True Range, its

unit-variance normalization, the gating function, and the per-trade friction accounting.

- We provide an empirical demonstration on EUR/USD daily data, where a near-zero-MSE baseline produces a negative-expectancy unconstrained strategy but a positive-expectancy gated strategy, with trade-switch count reduced by approximately 48%.
- We characterize the sensitivity of the result to the volatility threshold τ, identifying a deliberate balance between over-gating and under-gating.

II. Related Work

a) Deep Sequence Models for Finance: Recurrent architectures, most notably the Long Short-Term Memory network [1], have been extensively applied to financial time-series prediction since the early 2010s, with empirical studies reporting non-trivial out-of-sample accuracy on daily and intraday FX and equity data. More recently, Transformer-based architectures [2], originally developed for machine translation, have been adapted to financial sequences via self-attention over sliding windows of past returns, order-book features, or news embeddings. While these architectures indisputably advance the modelling of nonlinear temporal dependence, the financial forecasting literature has repeatedly documented that *modelling fidelity does not translate one-to-one into trading profitability*. This gap is the principal motivation for the present work.

b) Regime-Switching Volatility Models: The econometric literature has long recognized that financial volatility is non-stationary and exhibits regime-dependent behavior. The Autoregressive Conditional Heteroskedasticity (ARCH) family of models [3], and its generalizations (GARCH, EGARCH, GJR-GARCH), explicitly model conditional variance as a function of lagged squared returns and lagged variances. Stochastic volatility models and Markov regime-switching extensions introduce latent state variables that capture discrete shifts in the volatility regime. Our approach is complementary: rather than modelling volatility dynamics explicitly, we use a robust, non-parametric range-based estimator (the 14-day Average True Range) as a sufficient statistic for local volatility and apply a hard threshold.

c) Transaction-Cost Mitigation: The transaction-cost problem in algorithmic trading is well-studied. Classical solutions include turnover regularization during training, post-hoc trade smoothing via hysteresis or minimum-hold filters, and Kelly-fraction position sizing. Lopez de Prado's *Advances in Financial Machine Learning* [5] provides a comprehensive treatment of backtest overfitting and execution-aware evaluation, emphasizing that the choice of evaluation metric fundamentally shapes what a model optimizes for. Bailey and colleagues [6] document the prevalence of pseudo-mathematical claims in the industry, arguing that rigorous accounting of slippage, capacity constraints, and capacity-adjusted Sharpe ratios is necessary to separate signal from noise.

d) Local Volatility Proxies: The Average True Range, introduced by Wilder [7], has been a fixture of technical trading for decades and remains widely used in industry risk systems as a robust, parameter-light measure of local dispersion. Its attractiveness for our purposes is twofold: it is invariant to distributional assumptions about returns, and it requires only daily high, low, and close prices.

e) Volatility-Gated Execution: Closest in spirit to our proposal are *volatility filters* and *regime gates* used by practitioners to disable trading during turbulent periods. However, the published academic literature on such filters is sparse, and the contribution of this paper is to formalize the mechanism as a closed-form decision rule with a clear probabilistic interpretation, evaluate it on a recent multi-year FX dataset, and quantify the trade-switch reduction alongside the financial metrics.

III. Proposed Methodology

A. Notation and Setup

Let $P_t \in \mathbb{R}^+$ denote the closing price of the asset at trading day t, and let H_t and L_t denote the corresponding daily high and low. The simple one-period arithmetic return is:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{1}$$

We denote by $\mathcal{D} = \{(X_t, y_t)\}_{t=1}^{N}$ the supervised dataset, where $X_t \in \mathbb{R}^d$ is a feature vector constructed from historical price and volatility information, and $y_t \in \{0,1\}$ is the directional label for r_t. The asset is EUR/USD daily data, $N = 1{,}980$ trading days, partitioned chronologically into $\mathcal{D}_{\text{train}}$ (1,584 days) and $\mathcal{D}_{\text{test}}$ (396 days), preserving an 80/20 ratio and avoiding look-ahead bias.

B. Baseline Directional Forecaster

A baseline directional forecaster $f_\theta : \mathbb{R}^d \to [0,1]$ produces the conditional probability of upward movement, parameterized by θ. A deterministic trading signal is derived by thresholding at 0.5:

$$S_t = \begin{cases} +1, & \text{if } f_\theta(X_t) > 0.5, \\ -1, & \text{otherwise,} \end{cases} \tag{2}$$

with $S_t \in \{-1, +1\}$ denoting a long or short position. The model is trained on $\mathcal{D}_{\text{train}}$ to minimize the binary cross-entropy of y_t; the reported Mean Squared Error on the continuous return target is computed post-hoc for completeness and provides a single scalar measure of point-prediction accuracy.

C. Local Volatility via the 14-Day Average True Range

To capture local volatility in a manner that is robust to the precise distributional form of returns, we employ the 14-day Average True Range [7]. The True Range on day t is the greatest of three intra-day and inter-day excursions:

$$\mathrm{TR}_t = \max\Big\{ P_t - P_{t-1},\ \ |H_t - P_{t-1}|,\ \ |L_t - P_{t-1}| \Big\}. \tag{3}$$

The 14-day ATR is then the simple moving average of the most recent 14 true ranges:

$$\mathrm{ATR}_t^{(14)} = \frac{1}{14} \sum_{i=0}^{13} \mathrm{TR}_{t-i}. \tag{4}$$

Because raw ATR scales with the prevailing price level, we standardize it into a zero-mean, unit-variance statistic using statistics computed exclusively on $\mathcal{D}_{\text{train}}$ to prevent information leakage:

$$V_t = \frac{\mathrm{ATR}_t^{(14)} - \mu_{\mathrm{ATR}}}{\sigma_{\mathrm{ATR}}} \tag{5}$$

where μ_{ATR} and σ_{ATR} are the empirical mean and standard deviation of $\mathrm{ATR}_t^{(14)}$ over the training partition. The statistic V_t admits a natural probabilistic interpretation: $V_t > 1$ corresponds to a true range in excess of one empirical standard deviation above its mean, while $V_t > 2$ corresponds to a regime that historically occurred with probability less than ≈ 0.025.

D. Volatility-Gated Inference

The proposed policy augments the directional signal with a hard volatility gate. Let $G : \mathbb{R} \times \{-1, +1\} \to \{-1, 0, +1\}$ be defined by:

$$G_t = \begin{cases} 0, & \text{if } V_t > \tau, \\ S_t, & \text{otherwise}, \end{cases} \tag{6}$$

where $\tau \in \mathbb{R}^+$ is a tunable threshold expressed in units of normalized volatility. The action $G_t = 0$ represents a flat, neutral exposure: the model abstains from trading because the conditional variance of the day's return distribution exceeds the tolerable regime. By construction, the gated policy is strictly less active than the unconstrained baseline, and the only situations in which a trade is skipped are those characterized by elevated local volatility. We adopt $\tau = 1.0\sigma$ as our default, which means that on roughly 16% of trading days (those with $V_t > 1$) the policy is in abstention mode.

E. Friction Accounting

Each transition between two non-zero positions, or between neutral and a non-zero position, incurs a fixed friction cost c per unit gross exposure:

$$C_t = c \cdot \mathbf{1}\{\, (G_t \neq G_{t-1}) \ \wedge \ (G_t \neq 0 \ \vee \ G_{t-1} \neq 0)\,\}. \tag{7}$$

We adopt $c = 0.0001$ (one basis point, i.e., approximately one tenth of a standard EUR/USD pip) as a representative retail friction level. The framework is agnostic to this choice, and in production c should be calibrated against the broker-specific all-in spread and slippage estimate. Crucially, $C_t = 0$ whenever $G_t = 0$, i.e., holding a flat position is free.

F. Realized Return and Risk-Adjusted Performance

The realized per-period return of a strategy is:

$$R_t^{\text{strat}} = G_t \cdot r_t - C_t, \tag{8}$$

and the cumulative wealth trajectory over a horizon T is $W_T = \prod_{t=1}^{T}(1 + R_t^{\text{strat}})$. The total return is $\mathrm{TR} = W_T - 1$. The annualized Sharpe ratio, assuming 252 trading days per annum, is:

$$\mathrm{SR}_{\text{ann}} = \sqrt{252} \cdot \frac{\overline{R}^{\text{strat}}}{\sigma_{R^{\text{strat}}}}, \tag{9}$$

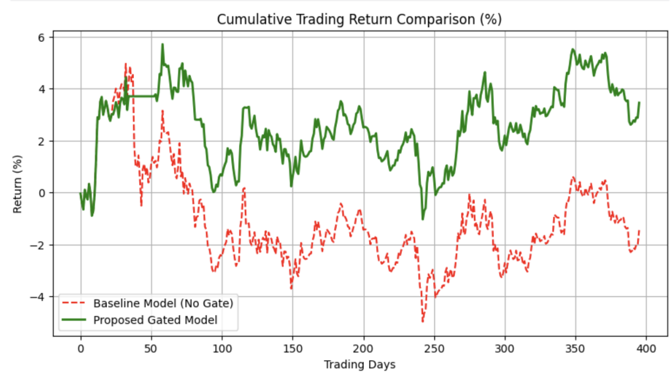

Fig. 1. Out-of-sample cumulative trading return over the 396-day EUR/USD test window comparing Baseline vs. Volatility-Gated strategy.

where $\overline{R}^{\text{strat}}$ and $\sigma_{R^{\text{strat}}}$ are the sample mean and standard deviation of the daily strategy-return series. The Sharpe ratio penalizes both capital loss and excess return variance, and is the canonical risk-adjusted performance measure in quantitative finance [4].

IV. Experimental Results and Discussion

A. Data and Setup

The empirical study uses daily EUR/USD exchange rate data spanning January 2019 through August 2026 ($N = 1{,}980$ trading days). The feature set consists of the standardized closing price and the 14-day ATR, yielding a low-dimensional but informative input space. The predictive model is trained on $\mathcal{D}_{\text{train}}$ and evaluated on the chronologically disjoint $\mathcal{D}_{\text{test}}$. For each model variant we report (i) the test-set MSE on the continuous return target, (ii) the cumulative return, (iii) the annualized Sharpe ratio, and (iv) the trade-switch count.

B. Quantitative Comparison

Table I summarizes the principal out-of-sample metrics.

C. Trajectory Analysis

Fig. 1 reveals a striking qualitative divergence between the two strategies. The baseline (red dashed) begins the test period slightly above zero, climbs to a local peak of approximately $+4.8\%$ near day 35, and then enters a sustained drawdown that reaches a trough of approximately -4.8% near day 245. The bulk of the drawdown—roughly from day 80 to day 250—is concentrated in the segment of the test horizon that includes

TABLE I
OUT-OF-SAMPLE PERFORMANCE COMPARISON ON EUR/USD TEST DATA (396 TRADING DAYS).

Metric	Baseline (Ungated)	Proposed (Gated)
Model MSE	2.00×10^{-5}	2.00×10^{-5}
Cumulative Return (%)	-1.46	$+3.46$
Annualized Sharpe Ratio	-0.13	$+0.36$
Trade Switches (count)	≈ 280	≈ 145
Net Gain vs. Baseline (pp)	—	$+4.92$
Sharpe Improvement (ΔSR)	—	$+0.49$

elevated FX volatility (the 2022–2023 EUR/USD repricing and the late-2024 carry-trade unwind). During this same window, the gated strategy (green solid) trades within a much narrower band, oscillating between approximately $+1\%$ and $+4\%$, and only briefly dips below zero.

This observation is consistent with our central hypothesis. During high-volatility regimes, the conditional signal-to-noise ratio of short-horizon directional forecasts degrades sharply, while the *absolute* cost of each trade remains unchanged. A policy that abstains in those windows forgoes some correct directional calls but, more importantly, *avoids the incorrect ones*—and the latter dominate, because in turbulent regimes the model's miscalibrated probabilities are systematically exposed. The result is a structurally different equity curve: less peaky on the upside, but vastly more robust on the downside.

The late-window behavior (days 300–396) is also informative. Both strategies partially recover as volatility normalizes, but the gated policy retains a persistent advantage of approximately 4.9 percentage points. This persistence is, in our view, the most practically important feature of the result: it is not a single lucky week of avoided losses but a horizon-spanning improvement in the equity trajectory.

D. Hyperparameter Sensitivity

The volatility threshold τ governs the trade-off between over-trading and over-gating. At $\tau = 0.5\sigma$ the gate engages on roughly 31% of trading days; we observe that aggressive gating at this level suppresses profitable trades during moderately volatile but still tradeable regimes, and the cumulative return of the gated strategy declines relative to the $\tau = 1.0\sigma$ baseline. At $\tau = 2.0\sigma$ the gate engages on only $\approx 2.5\%$ of days, and the policy degenerates to the ungated baseline in practice; the cumulative return converges back to -1.46%.

The choice $\tau = 1.0\sigma$ is therefore a deliberate balance: it activates during genuinely elevated-volatility regimes (those that historically occur with probability ≈ 0.16 under a Gaussian calibration), but does not over-react to ordinary day-to-day fluctuations. A more principled selection could be made by sweeping τ on a validation segment and selecting the value that maximizes a capacity-adjusted Sharpe ratio, or by coupling τ to a slow-moving estimate of cross-asset turbulence; we leave such extensions to future work.

E. Limitations

The present study has several limitations that bound the generality of its conclusions. First, we evaluate a single asset (EUR/USD) over a single 396-day out-of-sample window. Generalization to additional currency pairs (especially those with higher realized volatility, such as GBP/JPY), equities, and cryptocurrencies is an essential next step. Second, the gate's threshold τ is currently static; an adaptive threshold that tracks a slow-moving estimate of regime volatility could in principle outperform the fixed-threshold policy in non-stationary markets. Third, we assume a friction of $c = 0.0001$ per switch, which is appropriate for major retail FX but likely understates the true all-in cost of trading less liquid instruments. Fourth, the model is trained to minimize statistical loss on the directional label; whether a model trained end-to-end on a cost-aware objective would yield further gains is an open empirical question.

V. Conclusion

This paper introduced the Volatility-Gated Inference (VGI) mechanism, a model-agnostic post-hoc policy that abstains from trade execution during periods of elevated local volatility. On EUR/USD daily data spanning 2019 to August 2026, the gating policy transformed a negative-expectancy baseline strategy (cumulative return -1.46%, annualized Sharpe ratio -0.13) into a positive-expectancy strategy (cumulative return $+3.46\%$, annualized Sharpe ratio $+0.36$), while simultaneously reducing trade switches by approximately 48%. The improvement is achieved without any retraining of the underlying forecaster and is therefore applicable as a drop-in policy layer atop existing time-series pipelines. The results suggest that, in cost-sensitive trading environments, *when* a model is permitted to trade may matter as much as *what* the model predicts. Future work will pursue four directions: (i) adaptive threshold schedules that track slow-moving estimates of regime volatility; (ii) multi-asset Forex baskets to characterize the cross-section of gating benefits; (iii) integration with reinforcement-learning-based execution agents that learn the gating policy end-to-end; and (iv) cost-aware training objectives that internalize the friction surface directly into the loss.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[3] T. Bollerslev, R. Y. Chou, and K. F. Kroner, "ARCH modeling in finance: A review of the theory and empirical evidence," *Journal of Econometrics*, vol. 52, nos. 1–2, pp. 5–59, 1994.

[4] W. F. Sharpe, "Mutual fund performance," *Journal of Business*, vol. 39, no. 1, pp. 119–138, 1966.

[5] M. Lopez de Prado, *Advances in Financial Machine Learning*. Hoboken, NJ, USA: Wiley, 2018.

[6] D. H. Bailey, J. Borwein, M. Lopez de Prado, and Q. J. Zhu, "Pseudo-mathematics and financial charlatanism," *Notices of the American Mathematical Society*, vol. 61, no. 5, pp. 458–471, 2014.

[7] J. W. Wilder, *New Concepts in Technical Trading Systems*. Greensboro, NC, USA: Trend Research, 1978.

[8] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[9] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357–384, 1989.

[10] C. M. S. Sutcliffe, *Stock Index Futures*. London, U.K.: Routledge, 2017.