

Nima HEIDARI, PhD Candidate

n_heidari@ind.iust.ac.ir

Iran University of Science and Technology, Tehran, Iran

Saeed MIRZAMOHAMMADI, PhD (corresponding author)

mirzamohammadi@iust.ac.ir

Iran University of Science and Technology, Tehran, Iran

Babak AMIRI, PhD

babakamiri@iust.ac.ir

Iran University of Science and Technology, Tehran, Iran

A System-Dynamics-Inspired Transformer for Realized-Volatility Forecasting in Foreign-Exchange Markets

Abstract. Forecasting foreign-exchange (FX) return volatility matters for risk management, option pricing, and position sizing, yet deep-learning gains often vanish under realistic, cost-aware evaluation. We propose the SD-Transformer, coupling a Transformer encoder to a differentiable, system-dynamics-inspired stock-flow cell (SDCell) that predicts a residual on a per-fold Heterogeneous Autoregressive (HAR) anchor. Evaluation is leakage-safe, multi-seed, and cost-aware, using an expanding walk-forward design with formal predictive tests. Across eleven FX pairs (2011-2025), the model attains the lowest pooled QLIKE among random-walk, EWMA, GARCH, HAR, SHAR, and LightGBM benchmarks, is Diebold-Mariano-significant against HAR in every fold at the one-day horizon, and is the sole member of a 90% Model Confidence Set. These daily gains are statistically robust but numerically small; the economically large effect emerges at four-hour frequency, where the model outperforms HAR by 16 to 21 R-squared points. Ablations show the stock-flow core significantly outperforms matched GRU, PatchTST, Neural-ODE, and core-free Transformer alternatives at one day. We also document a label-accuracy illusion, whereby smoothed labels inflate apparent directional accuracy without real trading skill, and apply a uniform, validation-fitted QLIKE calibration across all models.

Keywords: deep learning, Transformer neural networks, realized volatility, foreign exchange markets, system dynamics, Diebold-Mariano test, model confidence set.

JEL Classification: C45, C53, C58, F31, G17.

Received: 29 July 2026	Revised: 9 September 2026	Accepted: 11 September 2026
------------------------	---------------------------	-----------------------------

1. Introduction

Volatility in exchange-rate returns is one of the few features of currency markets that can actually be predicted. Daily FX returns are close to serially unpredictable, consistent with fifty years of market-efficiency research, yet their magnitude exhibits

strong, persistent volatility clustering (Figure 1). Forecasting volatility therefore makes sense even where direction cannot be forecast: volatility forecasts drive value-at-risk limits, option pricing, and position sizing, and even modest improvements carry economic value.

Machine-learning research on FX volatility faces two recurring problems. First, many reported results rely on evaluation designs that leak information or ignore transaction costs, producing directional-accuracy rates of 90-99% that disappear once tested honestly (Hu, Zhao, & Khushi, 2021; Friday, Pati, & Mishra, 2025). Second, deep networks are typically used as black boxes disconnected from economic reasoning, so even genuine profits are hard to explain or trust.

This study addresses both problems with an architecture that embeds a recurrent core informed by an economic prior, tested under a protocol designed to interrogate its own assumptions. The central idea is that financial regimes (risk appetite, carry accumulation, valuation gaps) behave like stocks accumulating flows over time, in the sense of the system-dynamics (SD) framework of Forrester (1961) and Sterman (2000). We embed in a Transformer encoder a differentiable stock-flow cell whose gains and decays are modulated by the encoder context, combining structured, interpretable memory with attention.

The main contributions of the paper are as follows.

(1) **The SD-Transformer architecture**, which couples a Transformer encoder with a differentiable stock-flow recurrent cell (SDCell) whose gains and decays are modulated by the encoder context, and whose output heads correct a per-fold HAR anchor.

(2) **Significance-tested volatility results**. On eleven FX pairs, the model attains the lowest pooled QLIKE against random-walk, EWMA, GARCH, HAR, SHAR, and LightGBM, is Diebold-Mariano significant against HAR in every fold at one day (five of six at five days), and is the sole member of a 90% Model Confidence Set. These daily margins are statistically robust but small (a relative QLIKE gain below 0.5% over a near-ceiling HAR); the large accuracy effect is at four-hour frequency, where the R^2 margin over HAR is 16 to 21 points.

(3) **Ablations that locate the source of the gain**. Removing the accumulation prior reduces accuracy; on the daily panel the SDCell is significantly more accurate at one day than a matched GRU, a PatchTST trunk (Nie et al., 2023), a Neural-ODE core, and a core-free Transformer (all $p \leq 0.03$), converging with them at five days.

(4) **A reproducible evaluation methodology**, including a validation-fitted QLIKE level calibration that removes Jensen bias, and a documented label-accuracy illusion in directional forecasting.

Section 2 reviews related work; Section 3 covers the data, architecture, and evaluation protocol; Section 4 reports the volatility results; Section 5 the ablations and directional studies; Section 6 interpretability; and Section 7 concludes.

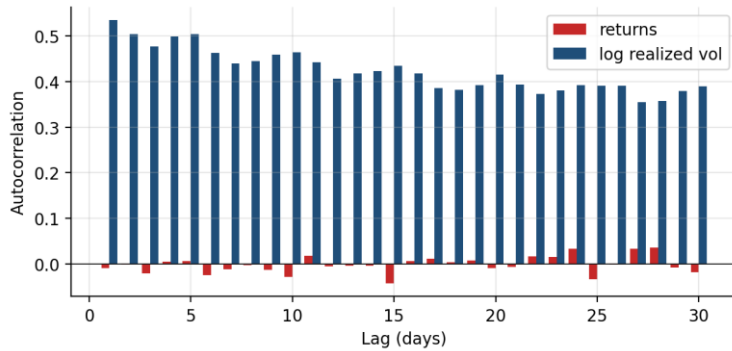


Figure 1. Autocorrelation of Daily Returns vs. Realized Volatility (EURUSD)

Source: Authors' own calculations.

2. Literature Review

Three strands of work bear on this paper: classical and deep volatility forecasting, Transformer architectures with structured memory, and the evidence on FX predictability and honest evaluation. The Heterogeneous Autoregressive (HAR) model (Corsi, 2009) remains the key benchmark, capturing long-memory behaviour with few parameters, alongside GARCH-family models and the semivariance extension SHAR. Patton (2011) showed that only some loss functions, including QLIKE and MSE, yield consistent rankings when the true volatility is proxied imperfectly, a result this paper builds on directly. Recurrent and hybrid deep architectures dominate the applied literature: autoencoder-LSTM combinations (Jung & Choi, 2021), LSTM-GARCH hybrids (Michańkow, Sakowski, & Ślepaczuk, 2024), and neural networks for forex volatility (Liao, Chen, & Ni, 2021). Further examples include multivariate deep-model comparisons (Ge, Lalbakhsh, Isai, Lenskiy, & Suominen, 2023), decomposition ensembles (Kausar et al., 2024), and deep forecasting during COVID-19 (Naas & Zouaoui, 2024). Hybrid ARMA-GARCH-MLP designs have been applied to volatile commodity prices (Sari-Hassoun, Boudjourfa, & Mekidiche, 2025) and neural solvers to derivative pricing (Park, Moon, & Kim, 2025). Related work spans fuzzy-inference trading systems (Hernandez-Aguila et al., 2021) and machine-learning classifiers for the euro-dollar rate (El Mahjouby et al., 2024). These studies show deep models can match or exceed classical baselines, but are rarely evaluated under a walk-forward, multi-seed, cost-aware protocol with formal predictive-ability tests.

Attention-based architectures form the second strand. Since the Transformer (Vaswani et al., 2017), attention-based models have entered finance: time-embedded Transformers for FX spot movements (Fischer, Sterling, & Lessmann, 2024), trend-aware variants for FX returns (Liu, Ślepaczuk, & Tang, 2025), and cascading language-model (Pornwattanavichai, Maneeroj, & Boonsiri, 2022) and Vision-Transformer (Fisichella & Garolla, 2021) systems. Closest to our own conditional result, Han, Huang, Huang, and Zhou (2026) report Transformer skill in daily market returns concentrated in high-sentiment, low-VIX states. Long-horizon architectures

such as Informer (Zhou et al., 2021), Autoformer (Wu, Xu, Wang, & Long, 2021), and PatchTST (Nie, Nguyen, Sinthong, & Kalagnanam, 2023) target attention efficiency and patching, and the Temporal Fusion Transformer (Lim, Arik, Loeff, & Pfister, 2021) adds variable-selection gating. None embeds an economic prior about how financial state accumulates, which is the gap the SDCell fills. The system-dynamics approach (Forrester, 1961; Sterman, 2000) represents systems as stocks integrating flows under balancing feedback, and gated recurrent networks (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) use a similar first-order update; we make the analogy explicit so the cell's latent units behave like accumulating stocks.

The third strand concerns FX predictability itself and the standards used to judge it. Innovations in global FX volatility price the cross-section of carry returns (Menkhoff, Sarno, Schmeling, & Schrimpf, 2012), volatility risk premia predict exchange rates (Della Corte, Ramadorai, & Sarno, 2016), and spot and forward volatility have been studied jointly (Della Corte, Sarno, & Tsiakas, 2011), all motivating volatility as the primary, more-predictable target. Geopolitical risk moves real exchange rates directly (Doğanlar, Mike, Kızılkaya, & Akyıldız, 2025), which motivates the geopolitical-news family we reserve for ablation, and time-varying analyses tie global uncertainty to market volatility (Hurduzeu, Lupu, Lupu, & Filip, 2024). Surveys document the prevalence of optimistic evaluation (Hu, Zhao, & Khushi, 2021), and multimodal architectures reporting above 95% accuracy typically permit label leakage (Friday, Pati, & Mishra, 2025). Our protocol follows the predictive-ability tradition: Diebold-Mariano (Diebold & Mariano, 1995), Giacomini-White (Giacomini & White, 2006), the Model Confidence Set (Hansen, Lunde, & Nason, 2011), and Pesaran-Timmermann (Pesaran & Timmermann, 1992), with backtest-overfitting safeguards (Bailey & Lopez de Prado, 2014).

3. Data and Methodology

3.1 Data and Volatility Measurement

We analyse eleven liquid FX pairs (AUDUSD, EURAUD, EURCAD, EURCHF, EURGBP, EURJPY, EURUSD, GBPUSD, USDCAD, USDCHF, USDJPY) from December 2011 through July 2025, about 3,530 trading days per pair. The higher-frequency analysis uses six USD majors at four-hour intervals from 2010 to 2025 (about 146,000 bars); this narrower universe is deliberate, restricting the intraday study to pairs whose hourly Dukascopy quotes are dense enough for reliable four-hour realized variance. The sample ends in late July 2025, so the final test fold is a partial year and its results should be read accordingly. Descriptive statistics (Table A1) show mean daily returns close to zero with consistently positive excess kurtosis, especially for pegged or managed pairs such as EURCHF and USDCHF: fat tails with no clear drift, typical of FX markets.

Intraday FX prices are one-hour bid/ask OHLCV bars from the Dukascopy Bank historical data feed for the eleven pairs over 2010-2025 (the extra 2010 year

builds lagged components only, so the first modelled observation is in 2011); the one-hour mid quote is the average of the bid and ask close. Daily realized variance is the sum of squared one-hour mid log-returns within each UTC calendar day, so timestamps are kept in UTC and no daylight-saving adjustment is required. With daily close $P(t)$, daily log-returns and realized volatility from intraday returns $r(t,i)$ are

$$r_t = \ln P_t - \ln P_{t-1} \quad (1)$$

$$RV_t = \left(\sum_i r_{t,i}^2 \right)^{1/2} \quad (2)$$

The Stage-1 target is log realized volatility, one-half the log of the daily realized variance, at horizons of one and five days (one, three, and six bars at 4H). Hours with missing quotes are dropped, as are days with too few valid hourly bars. Macro-financial features come from public series identified by their standard tickers: interest-rate, Treasury, and dollar-index series from FRED, and the VIX, VVIX, MOVE, OVX, and GVZ implied-volatility indices from the CBOE, each lagged by one publication day; a geopolitical-news family is derived from the GDELT Global Knowledge Graph. Each daily input is a 65-dimensional vector spanning five families (Table 1), standardized using training-fold statistics only. Leakage is controlled by publication lags on every macro source, strictly forward-looking targets, fitting standardization and the HAR anchor on training data only, and a five-day embargo between training and test blocks.

Table 1. Feature Families (65 features total)

Family	Count	Examples
Shared risk / intermarket	21	VIX, VVIX, MOVE, credit stress
News and sentiment	18	sentiment stocks, news pressure
Price / volatility (own)	13	multi-horizon returns, MACD
System-dynamics states	7	carry stock, PPP gap, rate diff.
Calendar	6	day-of-week, NFP, quarter-turn

Source: Authors' own calculations.

3.2 The SD-Transformer architecture

Figure 2 shows the architecture and Figure 3 the cell mechanism. An input window of L past feature vectors is projected to a d -dimensional space, augmented with learned positional and per-asset embeddings, and passed through a pre-norm Transformer encoder with multi-head self-attention, giving a contextual representation $h(k)$ at each step.

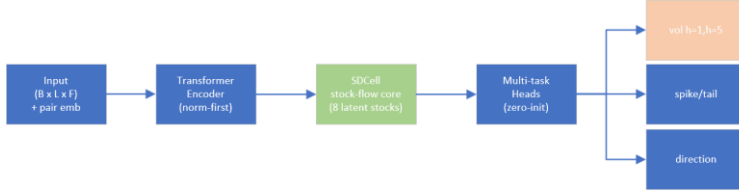


Figure 2. SD-Transformer Architecture: Encoder, Stock-flow Core (SDCell), and HAR-anchored Heads

Source: Authors' own creation.

The core (SDCell) maintains n latent stocks s updated in discrete time:

$$s_k = \text{clamp}(s_{k-1} + g_k \odot f_k - d_k \odot s_{k-1}) \quad (3)$$

where the inflow, gain, and decay are

$$f_k = \tanh(W_f x_k), \quad g_k = \sigma(W_g h_k), \quad d_k = 1/2 \sigma(W_d h_k) \quad (4)$$

with the inflow bounded by $\tanh$, the gain and decay modulated by the encoder context $h(k)$, and the state clamped for numerical stability. Equation (3) is a Transformer-gated, Euler-discretized stock-flow integrator: each latent unit accumulates a bounded flow subject to first-order balancing feedback, the defining computational elements of a system-dynamics model. The decay bias is initialized so the cell begins near a stable integrator.

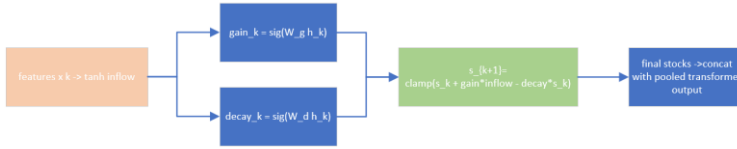


Figure 3. The SDCell Stock-flow Mechanism: Bounded Inflow, Context-modulated Gain, Balancing Decay

Source: Authors' own creation.

We are explicit about what the cell is and is not. It implements the three core computational elements of system dynamics: a stock as the integral of a net flow, an input-excited flow, and first-order balancing feedback. It does not implement the named, dimensioned loop structure of a classical Forrester model, so we describe it as SD-inspired. A coupled multi-loop variant with a learnable stock-to-stock coupling matrix (Section 5) is statistically indistinguishable from the base cell: the accumulation element, not loop complexity, carries the predictive value.

The final representation (pooled encoder output concatenated with the terminal stock vector) feeds a small MLP with zero-initialized output heads. Crucially, the heads predict a residual on a per-fold HAR anchor

$$a_t = \beta_0 + \beta_1 \ell_t^{(d)} + \beta_2 \ell_t^{(w)} + \beta_3 \ell_t^{(m)} \quad (5)$$

where $a(t)$ is the anchor forecast of the target $y(t+h)$ (log realized volatility h days ahead, $h \in \{1, 5\}$), and the regressors are backward-looking averages of log realized variance over the trailing k days, with $k = 1$ (daily), 5 (weekly), and 21

(monthly). Coefficients are estimated by pooled OLS on training data only, so the network is initialized at HAR performance and learns only the correction beyond it, which stabilizes training and keeps the comparison to HAR conservative.

Our headline volatility model (gsdt_comb) adds one lightweight layer on top of the base SDCell network (gsdt): for each horizon the final forecast is a convex combination of the network output and the HAR anchor, with a logistic weight driven by a single causal volatility-regime feature. Its two parameters are fit on the validation window only, so the layer sees no test data, and the gate leans on the network in turbulent regimes while reverting toward HAR in calm ones. We report the pure network (gsdt) and the combined model (gsdt_comb) separately so the SDCell contribution stays distinguishable from the combination layer. Ablation variants are tformer (encoder only, no core), sdt (SDCell with a plain head), gsdt (SDCell with the regime-gated head), sdt_gru (matched-capacity GRU core), and sdt_ode (continuous-time Neural-ODE core); all share the same data, anchor, losses, and capacity (about 112,000-150,000 parameters), so only the architecture changes. The training configuration is listed in the Appendix.

3.3 Evaluation Protocol

We apply an expanding-window walk-forward evaluation with six folds, using daily test years 2020-2025 and 4H test years 2021-2025 (Figure 4). Each model is trained on all previous data, validated on the last 10% of the training span, and tested out-of-sample with a five-day embargo; neural results are averaged over three seeds {23, 55, 123}.

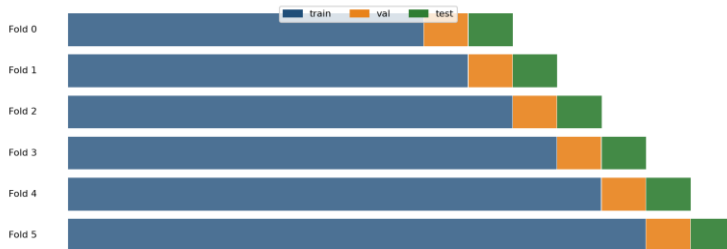


Figure 4. Expanding Walk-forward Evaluation: Six Folds with a Five-day Embargo

Source: Authors' own creation.

The primary loss is the consistent scoring rule of Patton (2011),

$$\text{QLIKE}(\hat{v}_t, RV_t) = \ln \hat{v}_t + \frac{RV_t^2}{\hat{v}_t} \quad (6)$$

where the variance forecast $\hat{v}(t) = \exp(2\hat{y}(t))$ is obtained from the model's log-volatility output $\hat{y}(t)$. Because an MSE-optimal log-volatility forecast systematically under-predicts variance (a Jensen effect), raw QLIKE can reward accidentally biased forecasters. We therefore fit, for every forecaster, horizon, and fold, one scalar log-volatility shift on the validation window only, minimizing the calibrated QLIKE; the

same procedure is applied uniformly to every model, so no model's calibration is imposed on another. We stress the uniformity of this calibration rather than its novelty: its absence, common in deep-volatility studies that calibrate only the network, can distort comparisons with tree and linear baselines.

Models are compared with the Diebold-Mariano (DM) test using HAC variance (Diebold & Mariano, 1995); regime concentration uses the conditional Giacomini-White (GW) test (Giacomini & White, 2006) with a causal, trailing-median volatility-regime indicator; and the block-bootstrap Model Confidence Set (MCS; Hansen, Lunde, & Nason, 2011) checks which models are statistically tied with the best. Directional skill uses the Pesaran-Timmermann (PT) test (Pesaran & Timmermann, 1992), which adjusts for the market base up-rate, and event forecasts are reported via AUC and calibrated Brier score. Trading results are reported both gross and net of a one-basis-point cost per unit turnover, so each rule cost sensitivity is visible rather than assumed. Volatility baselines are random walk, EWMA, GARCH(1,1), HAR, SHAR, and LightGBM; directional baselines are always-long, 21-day momentum, logistic regression, and a LightGBM classifier, all computed locally on the same inputs.

The learned and cross-sectionally estimated models are pooled: a single SD-Transformer, distinguished across pairs only by a per-pair embedding, is trained jointly on the eleven pairs, and HAR, SHAR, and LightGBM are likewise fit by pooled regression within each fold, so all models see identical information. GARCH(1,1), EWMA, and the random walk remain per-series by construction. DM and GW statistics use a Newey-West HAC long-run variance with a Bartlett kernel and a lag equal to the forecast horizon in bars, with loss differentials pooled across pairs and dates and significance confirmed by the MCS (stationary block bootstrap, 1,500 resamples, block length $\max(h+1, 3)$ bars).

4. Volatility Results

4.1 Daily Accuracy, Significance, and Regime Concentration

Aggregate QLIKE by model, pooled over all pair-days rather than averaged over the six per-fold means, is reported in Table 2 (lower is better). The combined SD-Transformer achieves the lowest QLIKE at both horizons; the pure model also beats LightGBM at $h = 1$. We are explicit about effect size, not only significance: the improvements are statistically robust but numerically small. The combined model improves pooled QLIKE over HAR by only 0.23% (-9.729 vs. -9.707), over LightGBM by 0.23%, and over EWMA by 0.40% at $h = 1$, for standardized effect sizes of just -0.09 , -0.10 , and -0.08 . The large DM statistics that follow (Table 3; -10.7 pooled against HAR) therefore reflect a very large effective sample acting on a consistent small edge, not a large per-observation effect. HAR is close to the achievable daily ceiling; the economically meaningful magnitude lives at four-hour frequency (Section 4.2). One ranking detail: at $h = 5$ the random walk overtakes EWMA and GARCH, because realized volatility is highly persistent weekly, so the

most recent level is already competitive while EWMA and GARCH revert too aggressively.

Table 2. Volatility Models vs. Baselines: Aggregate QLIKE (lower is better), daily

Model	QLIKE (h=1)	QLIKE (h=5)
gsdt_comb (ours)	-9.729	-9.695
gsdt (pure, ours)	-9.713	-9.692
HAR	-9.707	-9.687
LightGBM	-9.707	-9.684
SHAR	-9.705	-9.687
EWMA	-9.690	-9.650
GARCH(1,1)	-9.674	-9.640
Random walk	-9.575	-9.659

Source: Authors' own calculations.

That edge survives formal testing. Fold-by-fold DM statistics against EWMA, HAR, and LightGBM are reported in Table 3 (negative = ours better; $|\text{DM}|$ above 1.96 is significant at 5%), and at one day the model beats all three baselines in every fold. At five days it wins in five of six: the exception is F1 (2021), where HAR is significantly better (DM = +4.94), an unusually calm year in which the parsimonious HAR structure was already near-optimal weekly and the regime-sensitive correction added noise. The full per-fold loss matrix confirms the combined model has the lowest loss in five of six folds, the exception being F2 (2022), where GARCH is marginally lower (-9.154 vs. -9.153). Nor is the result driven by extreme pairs: removing the pegged EURCHF and USDCHF (excess kurtosis 437.9 and 132.8) leaves the pooled daily DM against HAR at -9.2 (from -10.7) and pooled QLIKE at -9.661, still strongly significant; the per-pair breakdown (Table A2) shows the model more accurate than HAR in all eleven pairs.

Table 3. Diebold-Mariano Statistics, Combined Model vs. Baselines (negative = ours better)

h	vs.	F0	F1	F2	F3	F4	F5	Wins
1	EWMA	-7.44	-1.86	-1.47	-3.51	-3.06	-5.22	6/6
1	HAR	-3.23	-2.83	-5.56	-7.57	-4.93	-4.56	6/6
1	LGBM	-5.16	-9.11	-2.00	-8.29	-1.70	-3.22	6/6
5	EWMA	-5.07	+1.38	-2.52	-4.57	-3.90	-6.53	5/6
5	HAR	-1.29	+4.94	-5.91	-2.74	-1.68	-2.73	5/6
5	LGBM	+1.37	-7.67	-2.84	-3.09	-2.55	-0.78	5/6

Source: Authors' own calculations.

The advantage is also concentrated where it is most valuable. Table 4 reports the conditional GW test against HAR with within-regime DM statistics ($h = 1$), for the same forecaster as Table 3. The GW test rejects equal conditional predictive ability in all six folds ($p \leq 0.018$), the within-regime DM is negative in both regimes in every fold, and in five of six the turbulent-regime DM is more negative, so the advantage concentrates where accurate forecasts matter most. The 90% MCS includes only the combined model at $h = 1$ and the two SD variants at $h = 5$. Figure

5 shows the model tracking out-of-sample realized volatility well, and the event heads add early warnings: Table 5 reports spike-event AUC between 0.59 and 0.67 in every fold.

**Table 4. Conditional Predictive Ability vs. HAR (h=1):
GW Test and within-regime DM**

Fold	GW stat	p-value	DM (turb.)	DM (calm)
F0	10.61	0.005	-2.79	-1.70
F1	8.08	0.018	-2.38	-1.60
F2	31.04	0.000	-5.40	-1.37
F3	61.17	0.000	-2.85	-7.28
F4	26.58	0.000	-3.70	-3.61
F5	21.11	0.000	-4.27	-1.82

Source: Authors' own calculations.

Table 5. Volatility-event Prediction (h=5): AUC and Calibrated Brier Score per Fold

Fold	AUC (spike)	AUC (tail)	Brier (spike)
F0	0.612	0.541	0.257
F1	0.671	0.577	0.242
F2	0.599	0.524	0.175
F3	0.643	0.655	0.184
F4	0.670	0.615	0.247
F5	0.592	0.654	0.157

Source: Authors' own calculations.

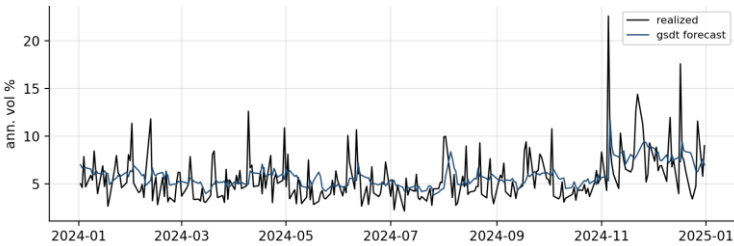


Figure 5. Predicted vs. Realized Volatility, EURUSD, out-of-sample (fold 4)

Source: Authors' own calculations.

4.2 Higher Frequency (4H)

Daily realized volatility is close to the HAR limit, but at four-hour intervals the SD-Transformer shows a clear advantage. Table 6 gives R^2 against the unconditional mean and Table 7 the corresponding DM tests and MCS membership. The SD models outperform HAR by 16-21 points at $h = 1$ and $h = 3$, and beat a locally reproduced LightGBM at $h = 1$ (DM = -5.6, LightGBM excluded from the 90% MCS), while being roughly equal at $h = 3$ and tied at $h = 6$. The need for a recurrent core is clearest at one bar: the core-free transformer is significantly worse than sdt (DM = +3.33, $p < 0.001$) and excluded from the MCS. The margin narrows at $h = 3$ (+1.88, $p = 0.06$), and the architectures converge at $h = 6$, where the advantage over

HAR is only about 3 R^2 points (45.7 vs. 42.6). The large margins are thus specific to the shorter horizons, where HAR daily-weekly-monthly structure suits 4H sampling least.

Table 6. Four-hour Volatility: R^2 vs. Mean (%), Three Seeds, Five Folds (2021-2025)

Model	h=1	h=3	h=6
sdt (ours)	44.0	48.6	45.7
gsdt (ours)	43.8	48.3	45.5
sdt_gru	44.0	48.9	45.6
tformer (no core)	43.3	48.1	45.4
LightGBM (fair)	40.7	46.9	44.8
HAR	23.1	32.4	42.6
EWMA	-8.1	17.2	34.2

Source: Authors' own calculations.

Table 7. Four-hour Volatility: DM Tests of GSDT and 90% MCS Membership

Comparison	h=1	h=3	h=6
gsdt vs. HAR	-25.2***	-18.4***	-6.5***
gsdt vs. LightGBM	-5.6***	-1.67 (p=.10)	-0.66 (n.s.)
tformer vs. sdt	+3.33***	+1.88 (p=.06)	+1.42 (n.s.)
90% MCS	{sdt}	{sdt,gsdt,sdt_gru,LGBM,tf}	{gsdt,sdt,LGBM,tf}

Source: Authors' own calculations.

4.3 Economic Value of the Volatility Edge

Statistical loss differences translate into two operational tests, reported in Table 8. The clearest gain is in volatility targeting: in a 10% annualized exercise, portfolios scaled by the model's one-day forecast track the target with the lowest realized tracking error (1.54 percentage points, vs. 1.61 for EWMA, 1.74 for LightGBM, 1.77 for HAR; the perfect-foresight bound is 1.09). We emphasize this result because it follows directly from forecast accuracy without a distributional assumption. A one-day 95% Gaussian VaR backtest is less favourable: only EWMA achieves correct coverage (Kupiec $p = 0.64$, Christoffersen $p = 0.84$), while the SD-Transformer, HAR, and LightGBM all reject ($p < 0.001$), the familiar difficulty of matching tail quantiles with a Gaussian VaR built on a conditional log-variance forecast. We therefore claim lower tracking error but explicitly not better VaR coverage than a well-tuned EWMA. On cost-benefit: at daily frequency, with EWMA already near the frontier and holding correct coverage, the case for a 115k-parameter Transformer on tracking error alone is weak. Its justification rests elsewhere: (i) at four-hour frequency the margin over HAR is large; (ii) the SDCell is trained jointly for the spike/tail early-warning heads and an interpretable stock state, which one EWMA recursion cannot provide; and (iii) the daily contribution is as much methodological as architectural.

Table 8. Volatility-targeting Tracking Error and one-day 95% VaR Backtest

Model	Track. err. (pp)	VaR-95 exc. (%)	Kupiec LRuc (p)	Christ. LRcc (p)
SD-Transformer (ours)	1.54	5.9	23.9 (<.001)	27.4 (<.001)
EWMA	1.61	5.1	0.22 (.64)	0.35 (.84)
HAR	1.77	6.0	29.6 (<.001)	34.7 (<.001)
LightGBM	1.74	5.9	23.2 (<.001)	24.0 (<.001)
Perfect foresight	1.09	--	--	--

Gaussian one-day 95% VaR, pooled 2020-2025, $n = 15,884$; perfect-foresight row is a tracking-error bound only.

Source: Authors' own calculations.

5. Ablation and Honest-Evaluation Studies

5.1 Is the Stock-flow Core Doing the Work?

Removing the core (tformer) reduces volatility accuracy: the pure SDCell network (sdt, without the regime-gated combination layer) outperforms HAR in five of six daily folds, while the core-free model does so in only three. Replacing the SD core with a matched-capacity GRU is the most direct test, pooled over three seeds at 4H. The SDCell is significantly better at one bar (DM -2.13 , $p = 0.03$) and six bars (DM -2.48 , $p = 0.01$), and indistinguishable at three bars, where the GRU has a marginally higher R^2 and stays inside the 90% MCS. We therefore make the narrow claim that the stock-flow core is never significantly worse than a generic recurrent core and is better at the short and long horizons, not that it is uniformly superior. One apparent tension with Table 6 deserves comment: sdt and sdt_gru report the same rounded R^2 at $h = 1$ (44.0), yet the QLIKE-based DM test separates them, because QLIKE penalizes under-prediction of variance and so weights turbulent bars more heavily. A difference concentrated there can be significant under QLIKE while moving pooled R^2 by less than a rounding step. A coupled multi-loop (Forrester-style) variant with a stock-coupling matrix performs the same as the standard cell (paired-t $p = 0.45$ at $h = 1$), showing that the accumulation prior, not loop complexity, is what matters.

We establish this comparison directly on the daily panel and broaden it to a modern deep baseline. Table 9 reports the pooled three-seed DM test of the SDCell against a matched GRU core, a PatchTST trunk (Nie et al., 2023), a Neural-ODE core, and the core-free transformer. The Neural-ODE is a continuous-time, RK4-integrated cousin of the stock-flow cell, of which the SDCell is the explicit-Euler special case; all four alternatives are trained on the identical daily panel, HAR anchor, folds, seeds, and ~ 112 -116k-parameter budget. At one day the SDCell is significantly better than every alternative: the GRU (DM -9.7), PatchTST (-11.8), the Neural-ODE (-8.0), the core-free transformer (-11.5), HAR (-13.1), and LightGBM (-2.2), all $p \leq 0.03$. At five days the deep architectures converge: the SDCell ties the GRU, PatchTST, and transformer, and still beats HAR and the Neural-ODE. That the Neural-ODE, the alternative closest to the SDCell, is its strongest competitor yet is still beaten confirms that the specific stock-flow

parameterization, not continuous-time recurrence in general, carries the edge; PatchTST, with no state-accumulation prior, matches the SDCell at neither horizon. These daily gaps are small in absolute terms (every alternative is within 0.006 of sdt), so the ablation establishes a consistent, significance-tested ranking rather than a large per-observation effect. Separately, the GDELT news family reserved for ablation adds a marginal, insignificant amount to daily accuracy (DM -1.55 , $p = 0.12$ at $h = 1$) and is retained for completeness but not relied upon.

Table 9. Daily-panel Architecture Ablation: DM of SDCell vs. each Alternative (negative = SDCell better)

Comparison (sdt vs.)	h=1	h=5
GRU core	-9.7***	-0.2 (n.s.)
PatchTST (Nie et al., 2023)	-11.8***	-0.7 (n.s.)
Neural-ODE core	-8.0***	-3.3**
Transformer (no core)	-11.5***	-0.6 (n.s.)
HAR	-13.1***	-3.7***
LightGBM	-2.2*	-1.5 (n.s.)

Note: Pooled daily QLIKE ($h=1$): sdt -9.713 , GRU -9.709 , Neural-ODE -9.709 , transformer -9.708 , PatchTST/HAR/LightGBM -9.707 . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Source: Authors' own calculations.

5.2 Direction is a different problem

For the secondary task of predicting return direction, the SD core offers no advantage. Table 10 shows that sdt (0.551) and tformer (0.554) are indistinguishable in the stable folds under honest evaluation, with an LSTM slightly behind (0.527); pooled over all days, accuracy falls to about 0.51. When evaluation is safe from leakage, increasing model complexity does not improve directional accuracy, consistent with market efficiency.

Table 10. Architecture Comparison on Direction (DA, stable folds 1-4, $h=21$)

Architecture	Directional accuracy
Sdt	0.551
Tformer	0.554
Lstm	0.527

Source: Authors' own calculations.

Direction can only be predicted in calm conditions. Across volatility terciles (Table 11), the calmest third shows real skill (PT above 5), the middle third weak skill, and the most turbulent third is indistinguishable from chance. A finer quintile-level split shows the same pattern, with the calmest quintile of pair-days carrying essentially the whole edge.

Table 11. Directional Accuracy by Causal Volatility Regime

Regime	Horizon	DA	PT
Calm	5	0.539	+5.83
Calm	21	0.535	+5.41

Regime	Horizon	DA	PT
Middle	5	0.521	+3.03
Middle	21	0.516	+2.28
Turbulent	5	0.492	-0.79
Turbulent	21	0.490	-0.43

Source: Authors' own calculations.

Timing matters as much as regime. During the four stable years (2021-2024) the model performs well (up to 0.588, PT = 9.81), while in the two transition years, 2020 (pandemic onset) and 2025 (regime change), accuracy drops below 0.45 and PT is strongly negative (Table 12). This reflects the nature of regime change, not a model flaw: recent data becomes misleading exactly when regimes break.

Table 12. Directional Accuracy per Test Year (h=21): Model vs. Key Baselines

Year	DA (model)	PT	DA (long)	DA (LGBM)
2020	0.402	-10.41	0.486	0.469
2021	0.576	+7.65	0.467	0.484
2022	0.525	+3.26	0.511	0.503
2023	0.514	+6.09	0.516	0.524
2024	0.588	+9.81	0.532	0.521
2025	0.439	-4.69	0.604	0.505

Source: Authors' own calculations.

Selective trading turns this conditional skill into an economic signal, and we report cost sensitivity explicitly rather than only the gross number. All confidence rules are validation-calibrated (seed agreement, calibrated probability, a regime gate, a conformal coverage threshold). Table 13 gives each rule gross and net of a one-basis-point cost per unit turnover, with a Newey-West t and a block-bootstrap 95% CI on the net daily return series (pooled across folds with at least 100 selected pair-days; thin folds are excluded rather than forced into a noisy average). Directional skill is genuine at every coverage level (PT up to +12.2), but risk-adjusted return is far more cost-sensitive than the gross numbers suggest. The plain conformal rule at a 20% target (15.2% realized coverage) is most robust: gross Sharpe 1.26 falls to a net 1.05 ($t = 2.03$, 95% CI [0.11, 2.03], excluding zero). Adding a regime gate to that same filter raises the gross Sharpe (1.45) but tightens coverage to 7.8%, raising turnover per unit of return enough that the net Sharpe falls to 0.46 ($t = 0.95$) with a CI including zero: a rule that looks better gross is not automatically better net. The remaining overlays fare worse, with "Regime + agreement" outright negative (-0.21) and "Meta-label" collapsing to essentially zero (0.05, $t = 0.14$) despite a strongly significant PT statistic, a reminder that directional significance and tradable return are distinct. We therefore lead with the plain conformal rule and report the rest, negative results included, for completeness.

Table 13. Selective Trading: Sharpe Gross vs. Net of a 1bp Turnover Cost (h=21, all pairs)

Selection rule	Cov. (%)	DA	PT	Gross Sh.	Net Sh. (t)	95% CI (net)
All days (no selection)	100.0	0.509	+2.41	0.44	0.32 (0.74)	[-0.49, 1.08]
Seed agreement	58.6	0.519	+4.22	0.35	0.25 (0.56)	[-0.53, 1.06]
Calibrated probability	34.5	0.579	+12.23	0.75	0.60 (1.38)	[-0.22, 1.47]
Regime + agreement	30.6	0.524	+3.62	0.66	-0.21 (-0.47)	[-0.99, 0.61]
Conformal, 20% target	15.2	0.628	+11.92	1.26	1.05 (2.03)	[0.11, 2.03]
+ regime gate	7.8	0.633	+8.80	1.45	0.46 (0.95)	[-0.32, 1.33]
Meta-label	34.8	0.564	+9.89	0.72	0.05 (0.14)	[-0.60, 0.87]

Note: Net Sharpe uses a 1bp-per-unit-turnover cost model; Sharpe, t and CI come from the same pooled daily net-return series. Tables 13-14 use the no-GDELT SD-Transformer (GDELT is reserved for the Section 5.1 ablation); Tables 10-12 use the full-feature variant.

Source: Authors' own calculations.

As a further check on the boundary-year weakness (2020, 2025), we add an auxiliary volatility-prediction head during training (sdt-mtl), used only in optimization, not at inference. We evaluate the same base model as Table 13 and this variant under a cross-sectional long-short book, gross and net of the same 1bp cost. The book is long the top five and short the bottom five of the eleven pairs by cross-sectional score, risk-parity weighted and volatility-targeted, with a validation-locked smoothing span and a pre-committed filter dropping the least-confident 20% of days. Table 14 reports the result. The auxiliary head raises the net Sharpe from 0.50 (t = 1.14, insignificant) to 1.06 (t = 2.34, significant) and reduces the boundary-year losses (2020 from -0.36 to -0.22, 2025 from -0.74 to -0.44). Every interior fold except 2024 improves as well, and 2024 itself stays strongly positive at 1.21, down from 1.73. The multi-task variant is therefore the more reliable of the two here.

Table 14. Cross-Sectional Long-Short Book, Gross vs. Net of 1bp Cost: Base Model (as in Table 13) vs. Multi-task Variant

Model	Cov. (%)	Gross Sh.	Net Sh. (t)	2020	2021	2022	2023	2024	2025
sdt (base)	83.3	0.55	0.50 (1.14)	-0.36	1.22	0.51	0.42	1.73	-0.74
sdt-mtl (+ aux. vol head)	72.6	1.11	1.06 (2.34)	-0.22	2.33	2.71	2.50	1.21	-0.44

Note: Columns 2020-2025 are net-of-cost Sharpe by fold. Base model is the same no-GDELT SD-Transformer run as Table 13, re-evaluated here under the cross-sectional book.

Source: Authors' own calculations.

Finally, we document a label-accuracy illusion. When a directional label is smoothed by moving-average filtering or trend-scanning, the label becomes strongly autocorrelated and therefore easy to predict: a classifier trained and scored on such a label reaches apparent accuracy as high as 0.69, none of which transfers to the market. Scored against the true next-period sign, the same predictions collapse to about 0.50 with zero net trading skill. Table 15 scores six representative labeling schemes with one common classifier, always against the true sign: most engineered labels are no better than the raw sign and several are worse, because their advertised gains exist only in their own label space. The single genuine improvement is drift adjustment (DA 0.527, PT 6.96 vs. 1.33 for the raw sign), which removes benchmark drift from the target instead of smoothing it. This illusion explains a class of 90%-

plus accuracy results in the literature and is why every metric in this study is computed against the economic ground truth.

Table 15. Labeling Bake-off (common LightGBM classifier, h=21, scored on true sign)

Label	DA	MCC	PT	Net Sharpe
Raw sign	0.502	0.011	+1.33	−0.10
Drift-adjusted	0.527	0.056	+6.96	+0.23
Trend-scanning	0.500	0.006	+0.72	−0.37
Triple-barrier (K=1)	0.498	0.006	+0.69	+0.09
1-sigma band	0.506	0.019	+2.43	−0.03
Drift-adj. + gray zone	0.525	0.049	+6.11	+0.26

Source: Authors' own calculations.

6. Interpretability

Permutation importance for the directional task, grouped by family (Table 16), shows that risk-state, price/volatility, and system-dynamics carry/value families each contribute about 30% of total importance, while news and calendar features contribute very little; the pattern is consistent across pairs. Information-coefficient analysis by regime adds two findings: carry and valuation signals provide most of their directional information in calm markets, matching risk-on carry behaviour, and news signals mainly predict high volatility rather than price direction. This fits the two-stage design: magnitude information feeds Stage 1, and the small directional advantage appears only in calm states. The eight latent stocks of the SDCell can be examined directly and change slowly, acting like regime state variables, in the spirit of the wider push for interpretable machine-learning models in asset pricing (Ferrara & Ciano, 2024).

Table 16. Feature-Family share of Total Importance (directional task)

Family	Share (%)
Risk / intermarket	32.3
Price / volatility	30.2
SD carry / value	29.9
News / sentiment	6.9
Calendar	0.6

Source: Authors' own calculations.

7. Discussion and Conclusions

We introduced the SD-Transformer, a Transformer encoder coupled to a differentiable stock-flow cell, for FX realized-volatility forecasting under a leakage-safe, multi-seed, cost-aware protocol. Two facts must be read together. At the daily horizon the model is statistically the best forecaster: lowest pooled QLIKE, DM-significant against HAR in every fold at $h = 1$, GW-confirmed turbulent-regime concentration, and sole member of a 90% MCS. The margin, however, is numerically small (a relative QLIKE gain below 0.5% over a near-ceiling HAR,

robust to dropping the pegged pairs), so we claim robustness, not a large daily effect. The large accuracy effect is at four-hour frequency, where the model beats HAR by 16 to 21 R^2 points. Ablations locate the source: the recurrent core is essential, and on the daily panel the SDCell is significantly more accurate at $h = 1$ than a matched GRU, PatchTST, a Neural-ODE core, and a core-free transformer (all $p \leq 0.03$), converging at five days. The contribution is thus as much methodological as architectural.

Two methodological contributions accompany the model. First, the validation-fitted QLIKE level calibration removes a Jensen-type bias that could give false wins to biased forecasters. Second, the label-accuracy illusion shows how smoothed labels can create the appearance of directional accuracy without real trading skill. Together with the honest directional results, namely architecture equalization at the efficiency ceiling, significant skill in calm regimes, and a selective net-of-cost Sharpe of 1.05 at a 20%-target conformal coverage ($t = 2.03$, 95% CI excluding zero), these place the claims in the credible range of FX research rather than the overly optimistic one. We also report, rather than omit, the overlays that did not survive cost accounting.

Threats to validity were addressed by design: leakage via embargoes, lags, and training-only fitting; seed luck via three-seed ensembles; data snooping via fold-wise DM, the MCS, and pre-set success criteria; and cost sensitivity by reporting every trading rule gross and net. Limitations remain. The stock-flow cell is mathematically a gated recurrence, so the system-dynamics reading is a guiding idea rather than a literal simulator. Directional accuracy appears capped by the available information: no architecture or feature set we tried moved it materially above the levels in Tables 10-12, and progress may require new information types (order flow, options, positioning) rather than new models. What did move is the economic outcome: the multi-task auxiliary volatility head of Table 14 roughly doubles the net-of-cost Sharpe of the cross-sectional book (0.50 to 1.06) and halves the boundary-year losses without improving raw hit rates, so the gain comes from better-conditioned confidence. All neural results rest on three seeds; dispersion is small (about $\pm 0.3 R^2$ points at 4H), but borderline significance levels should be read with this in mind. Benchmarking against the Temporal Fusion Transformer (Lim et al., 2021), which targets multivariate point forecasting rather than HAR-anchored volatility, is a natural next step, as is a microstructure analysis beyond our linear cost model.

The practical takeaway is that a compact, interpretable architecture with an economically motivated recurrent prior gives significant, regime-focused improvements in FX volatility forecasting, under standards often missing in applied research. To support reproducibility, the authors will release a public repository with the model and evaluation code, the feature list, the training configuration, the fold boundaries, the seeds, and data-source identifiers.

Acknowledgements: *This work was conducted at the Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.*

References

- [1] Bailey, D.H., Lopez de Prado, M. (2014), *The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality*. *Journal of Portfolio Management*, 40(5), 94-107.
- [2] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014), *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014, Association for Computational Linguistics, Stroudsburg, PA, USA, 1724-1734, <https://doi.org/10.3115/v1/D14-1179>.
- [3] Corsi, F. (2009), *A simple approximate long-memory model of realized volatility*. *Journal of Financial Econometrics*, 7(2), 174-196.
- [4] Della Corte, P., Ramadorai, T., Sarno, L. (2016), *Volatility risk premia and exchange rate predictability*. *Journal of Financial Economics*, 120(1), 21-40.
- [5] Della Corte, P., Sarno, L., Tsiakas, I. (2011), *Spot and forward volatility in foreign exchange*. *Journal of Financial Economics*, 100(3), 496-513.
- [6] Diebold, F.X., Mariano, R.S. (1995), *Comparing predictive accuracy*. *Journal of Business & Economic Statistics*, 13(3), 253-263.
- [7] Doğanlar, M., Mike, F., Kızılkaya, O., Akyıldız, İ.E. (2025), *How Geopolitical Risks Influence Real Exchange Rate Dynamics in Türkiye? Evidence from Fourier Cointegration Test with Non-Normal Errors*. *Economic Computation and Economic Cybernetics Studies and Research*, 59(3), 238-253.
- [8] El Mahjoub, M., Bennani, M.T., Lamrini, M., El Far, M., Bossoufi, B., Alghamdi, T. A.H. (2024), *Machine learning algorithms for forecasting and categorizing euro-to-dollar exchange rates*. *IEEE Access*, 12.
- [9] Ferrara, M., Ciano, T. (2024), *Improving the Interpretability of Asset Pricing Models by Explainable AI: A Machine Learning-based Approach*. *Economic Computation and Economic Cybernetics Studies and Research*, 58(4), 5-19.
- [10] Fischer, T., Sterling, M., Lessmann, S. (2024), *FX-spot predictions with state-of-the-art transformer and time embeddings*. *Expert Systems with Applications*, 249, 123538.
- [11] Fisichella, M., Garolla, F. (2021), *Can deep learning improve technical analysis of forex data to predict future price movements?*. *IEEE Access*, 9, 153083-153101.
- [12] Forrester, J.W. (1961), *Industrial Dynamics*. MIT Press, Cambridge, MA, USA.
- [13] Friday, I.K., Pati, S.P., Mishra, D. (2025), *A multi-modal approach using a hybrid Vision Transformer and Temporal Fusion Transformer model for stock price movement classification*. *IEEE Access*, 13.
- [14] Ge, W., Lalbakhsh, P., Isai, L., Lenskiy, A., Suominen, H. (2023), *Comparing deep learning models for the task of volatility prediction using multivariate data*, arXiv preprint, arXiv:2306.12446, <https://arxiv.org/abs/2306.12446>.
- [15] Giacomini, R., White, H. (2006), *Tests of conditional predictive ability*. *Econometrica*, 74(6), 1545-1578.

- [16] Han, Y., Huang, D., Huang, R., Zhou, G. (2026), *Daily market return prediction with Transformer*. SSRN Working Paper No. 6835039, <https://doi.org/10.2139/ssrn.6835039>.
- [17] Hansen, P.R., Lunde, A., Nason, J.M. (2011), *The model confidence set*. *Econometrica*, 79(2), 453-497.
- [18] Hernandez-Aguila, A., Garcia-Valdez, M., Merelo-Guervos, J.-J., Castañón-Puga, M., Castillo Lopez, O. (2021), *Using fuzzy inference systems for the creation of forex market predictive models*. *IEEE Access*, 9, 69391-69404.
- [19] Hochreiter, S., Schmidhuber, J. (1997), *Long short-term memory*. *Neural Computation*, 9(8), 1735-1780.
- [20] Hu, Z., Zhao, Y., Khushi, M. (2021), *A survey of forex and stock price prediction using deep learning*. *Applied System Innovation*, 4(1), 9.
- [21] Hurduzeu, G., Lupu, R., Lupu, I., Filip, R.I. (2024), *Evolving Economic Relationships: A TVP-VAR Analysis of Trade, World Uncertainty, and Stock Market Volatility in Europe*. *Economic Computation and Economic Cybernetics Studies and Research*, 58(4), 20-41.
- [22] Jung, G., Choi, S.-Y. (2021), *Forecasting foreign exchange volatility using deep learning autoencoder-LSTM techniques*. *Complexity*, 2021, 6647534.
- [23] Kausar, R., Iqbal, F., Raziq, A., Sheikh, N., Rehman, A. (2024), *Enhanced foreign exchange volatility forecasting using CEEMDAN with Optuna-optimized ensemble deep learning model*. *Sains Malaysiana*, 53(9), 3229-3239, <https://doi.org/10.17576/jsm-2024-5309-25>.
- [24] Liao, S., Chen, J., Ni, H. (2021), *Forex trading volatility prediction using neural network models*, arXiv preprint, arXiv:2112.01166, <https://arxiv.org/abs/2112.01166>.
- [25] Lim, B., Arık, S.Ö., Loeff, N., Pfister, T. (2021), *Temporal Fusion Transformers for interpretable multi-horizon time series forecasting*. *International Journal of Forecasting*, 37(4), 1748-1764.
- [26] Liu, D., Ślepaczuk, R., Tang, Z. (2025), *EXFormer: A multi-scale trend-aware Transformer with dynamic variable selection for foreign exchange returns prediction*, arXiv preprint, arXiv:2512.12727, <https://arxiv.org/abs/2512.12727>.
- [27] Menkhoff, L., Sarno, L., Schmeling, M., Schrimpf, A. (2012), *Carry trades and global foreign exchange volatility*. *Journal of Finance*, 67(2), 681-718.
- [28] Michańkow, Z., Sakowski, P., Ślepaczuk, R. (2024), *LSTM-GARCH hybrid model for the prediction of volatility*. *Computational Economics*, 63, 1511-1542.
- [29] Naas, M.-N., Zouaoui, H. (2024), *Forecasting foreign exchange rate volatility using deep learning: Case of US dollar/Algerian dinar during the COVID-19 pandemic*. *Research Papers in Economics and Finance*, 8(1), 91-114.
- [30] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. (2023), *A time series is worth 64 words: Long-term forecasting with Transformers*. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 1-5 May 2023, <https://arxiv.org/abs/2211.14730>.
- [31] Park, S., Moon, K.-S., Kim, H. (2025), *Asian Option Pricing Using the Physics-Informed Neural Networks Method*. *Economic Computation and Economic Cybernetics Studies and Research*, 59(1), 5-20.

- [32] Patton, A.J. (2011), *Volatility forecast comparison using imperfect volatility proxies*. *Journal of Econometrics*, 160(1), 246-256.
- [33] Pesaran, M.H., Timmermann, A. (1992), *A simple nonparametric test of predictive performance*. *Journal of Business & Economic Statistics*, 10(4), 461-465.
- [34] Pornwattanavichai, A., Maneeroj, S., Boonsiri, S. (2022), *BERTFOREX: Cascading model for forex market forecasting using fundamental and technical indicator data based on BERT*. *IEEE Access*, 10, 23425-23438.
- [35] Sari-Hassoun, S., Boudjourfa, N., Mekidiche, M. (2025), *Forecasting the Weekly Spot Oil Price Using a Hybrid Model ARMA-GARCH-MLP and Prophet Forecasting Method*. *Economic Computation and Economic Cybernetics Studies and Research*, 59(1), 156-173.
- [36] Sterman, J.D. (2000), *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston, MA, USA.
- [37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017), *Attention is all you need*. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, Curran Associates, Red Hook, NY, USA, 5998-6008, <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [38] Wu, H., Xu, J., Wang, J., Long, M. (2021), *Autoformer: Decomposition Transformers with auto-correlation for long-term series forecasting*. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. Curran Associates, Red Hook, NY, USA, 22419-22430, <https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html>.
- [39] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W. (2021), *Informer: Beyond efficient Transformer for long sequence time-series forecasting*. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, AAAI Press, Palo Alto, CA, USA, 35(12), 11106-11115, <https://doi.org/10.1609/aaai.v35i12.17325>.

Appendix

Table A1. Descriptive Statistics of Daily Log-Returns per Pair (2011-2025)

Pair	Ann. vol (%)	Skew	Ex. kurt.	Pair	Ann. vol (%)	Skew	Ex. kurt.
AUDUSD	10.3	-0.18	2.5	EURUSD	8.1	0.05	1.9
EURAUD	8.9	0.39	3.2	GBPUSD	8.8	-1.03	16.6
EURCAD	7.9	0.14	2.2	USDCAD	7.3	0.03	1.4
EURCHF	7.8	-10.03	437.9	USDCHF	9.8	-3.91	132.8
EURGBP	7.4	0.73	9.5	USDJPY	9.1	-0.27	4.7
EURJPY	9.7	-0.28	5.4				

Source: Authors' own calculations.

The model and training configuration are as follows. Encoder $d = 64$, four heads, three pre-norm layers, FFN width 128, dropout 0.1; window 64 days (32 bars at 4H); $n = 8$ latent stocks. AdamW, weight decay $1e-4$, cosine schedule, grad-clip 1.0, learning rate $1.5e-4$ daily or $1e-3$ at 4H, batch 512, early stopping with patience 4-8. Each result is the ensemble mean over seeds {23, 55, 123}.

Table A2. Per-Pair Pooled QLIKE at h=1: Combined Model vs. HAR (lower is better)

Pair	gsdt_comb	HAR	Δ (ours-HAR)	Pair	gsdt_comb	HAR	Δ (ours-HAR)
AUDUSD	-9.030	-9.017	-0.012	EURUSD	-9.763	-9.736	-0.027
EURAUD	-9.596	-9.584	-0.012	GBPUSD	-9.497	-9.472	-0.025
EURCAD	-9.948	-9.937	-0.011	USDCAD	-9.964	-9.947	-0.016
EURCHF	-10.395	-10.345	-0.050	USDCHF	-9.679	-9.662	-0.017
EURGBP	-10.197	-10.188	-0.009	USDJPY	-9.474	-9.429	-0.044
EURJPY	-9.478	-9.460	-0.018				

Source: Authors' own calculations.