International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 533-549
https://doi.org/10.70917/ijcisim-2026-4759
Article

# WHICH MODEL FAMILIES PAY? ECONOMETRIC, GRADIENT BOOSTING AND RECURRENT NEURAL NETWORK VOLATILITY FORECASTS IN ACTIVE TRADING STRATEGIES ON THE RUSSIAN STOCK MARKET

**Nikita I. Lysenok**

Faculty of Economic Sciences, HSE University, Moscow, Russian Federation
https://orcid.org/0000-0003-4660-1410
nilysenok@hse.ru

**Abstract:** . This paper asks which families of volatility forecasting models create economic value in active trading, and through which integration channel that value is transmitted. Seven models drawn from four families — econometric (GJR-GARCH, HAR-J), gradient boosting (XGBoost, LightGBM), recurrent neural networks (LSTM, GRU) and a hybrid combining HAR-J with boosting — are compared on the realized volatility of 17 liquid Moscow Exchange stocks computed from 10-minute returns over 2014–2026, using a feature space of 234 variables and QLIKE as the loss function. Two findings follow. First, the machine learning advantage is neither uniform across families nor stable across horizons: gradient boosting improves on the HAR-J benchmark by 5–9% at the one-day horizon but loses to it at the five-day horizon under rolling re-estimation, whereas recurrent networks are 22–31% worse than HAR-J at the one-day horizon and dominate it at no horizon; only the hybrid ranks at or near the top throughout. Second, the translation of accuracy into trading performance is governed by the integration channel rather than by the size of the accuracy gain: a 28% reduction in QLIKE raises the Sharpe ratio of a six-strategy portfolio by +0.79 and +0.85 under regime filtering and volatility targeting ($p < 0.001$), but by an insignificant +0.04 when the forecast only scales the parameters of an individual trade. Model selection therefore pays where the forecast governs the entry decision and the position size, and not otherwise. Statistical accuracy is a valid proxy for economic value in this setting, but only conditionally on the channel through which the forecast reaches the trading decision.

**Keywords:** realized volatility forecasting; gradient boosting; recurrent neural networks; model comparison; algorithmic trading; position sizing; economic value; emerging markets

## 1. Introduction

Volatility is the most predictable second-order property of financial returns, and a large methodological literature has been devoted to forecasting it more accurately. That literature is now dominated by a single question — whether machine learning beats econometric benchmarks — and it is usually answered with a loss function. The answer is reported as a percentage reduction in mean squared error or in QLIKE, and the implicit assumption is that a model which forecasts better is a model worth having.

This paper questions that assumption in two respects, and it does so on a market where the assumption is unusually costly to get wrong. The first question is whether "machine learning" is a meaningful unit of comparison at

all. The label covers algorithm families with very different inductive biases: gradient-boosted decision trees, which partition a tabular feature space, and recurrent neural networks, which impose a sequential representation on it. If those families behave differently on the same data, then a paper reporting that "machine learning improves on HAR" has averaged over a distinction that matters more than the one it reports. The second question is whether a reduction in forecast loss survives the passage into a trading decision. A volatility forecast can reach a portfolio through several distinct mechanisms — it can scale the stop-loss distance of an individual trade, it can gate entry into a position, or it can determine the size of that position — and there is no reason in principle for these mechanisms to be equally sensitive to forecast quality.

The empirical setting is the Russian equity market over 2022–2026, which imposes an unusually demanding test on both questions. The Moscow Exchange index lost 28% over the period while the policy rate of the Bank of Russia reached 21% and stayed above 16% from late 2023, so that the opportunity cost of holding an equity position was extraordinarily high. A trading system in this environment enters the market only when the signal is strong enough to beat a money-market alternative yielding a fifth of capital per year; strategies of the kind studied here are in position between 7% and 16% of the time. Any forecast-driven improvement must therefore be earned on a small fraction of the calendar, which makes the setting a hard case rather than a favourable one.

Seven models from four families are compared on the same data, the same feature space and the same loss: GJR-GARCH and HAR-J (econometric), XGBoost and LightGBM (gradient boosting), LSTM and GRU (recurrent networks), and a hybrid combining HAR-J with the best boosting model. Their forecasts are then passed into a book of six systematic strategies on 17 liquid Moscow Exchange stocks through three distinct integration channels, and the resulting multi-strategy portfolios are evaluated on a walk-forward control period.

Two results follow. The first is that the machine learning advantage is neither uniform across families nor stable across horizons. Gradient boosting improves on HAR-J at the one-day horizon by 5–9% in QLIKE, but under rolling re-estimation both boosting models lose to HAR-J at the five-day horizon, and XGBoost is the worst of the four re-estimated models at the twenty-two-day horizon. Recurrent networks are 22–31% worse than HAR-J at the one-day horizon and improve on it at no horizon on the held-out test year. Only the hybrid ranks at or near the top at every horizon, which is an argument for combination rather than for replacement.

The second result concerns the transmission of accuracy into performance. Moving from the econometric benchmark to the hybrid model improves QLIKE at the one-day horizon by 28% under rolling re-estimation. Passed through the two channels in which the forecast governs whether to trade and how large the position should be, that improvement raises the Sharpe ratio of the multi-strategy portfolio by +0.79 and +0.85 ($p < 0.001$). Passed through the channel in which the forecast only scales the stop-loss and take-profit distances of a trade whose entry decision is unchanged, the same improvement yields +0.04, indistinguishable from zero. The value of a better model is therefore conditional on the architecture of the system that consumes it, and a practitioner choosing where to spend model-development effort should first establish that such a channel exists in the system at all.

This paper is related to two prior studies by the author, and the boundary is stated explicitly. Lysenok (2025) constructed the hybrid forecasting model and established its accuracy advantage; Lysenok (2026) evaluated the three integration channels with the forecasting model held fixed and reported the aggregate gains of the multi-strategy portfolio over passive alternatives. Both are used here as established background and are cited as such. The present paper reverses the question: instead of fixing the model and varying the channel, it compares model families against one another and asks which of them, through which channel, produces a measurable economic result. The family-level comparison of forecasting architectures and the economic comparison of forecasting models against one another are reported here for the first time.

The remainder of the paper is organised as follows. Section 2 positions the contribution within the literatures on volatility forecasting, on model classes for tabular data and on the economic value of forecasts. Section 3 describes the data, the models, the evaluation protocol, the strategies and the integration channels. Section 4 reports accuracy at the model and family levels and the translation of accuracy into portfolio performance and economic value. Section 5

discusses the mechanisms behind both results, Section 6 states limitations and directions for further work, and Section 7 concludes.

## 2. Related work

### *2.1. Econometric and machine learning models of realized volatility*

The parametric tradition begins with the ARCH and GARCH families (Engle, 1982; Bollerslev, 1986), extended by Glosten, Jagannathan and Runkle (1993) to accommodate the asymmetric response of volatility to the sign of returns. A separate line follows the formalisation of realized volatility from intraday returns (Andersen, Bollerslev, Diebold and Labys, 2001; 2003) and culminates in the heterogeneous autoregressive model of Corsi (2009), which represents long memory through an additive cascade of daily, weekly and monthly components and remains the standard benchmark. Andersen, Bollerslev and Diebold (2007) add a jump component, separating the continuous part of quadratic variation with the bipower variation of Barndorff-Nielsen and Shephard (2004). The comparative evidence is unfavourable to elaborate parametric specifications: Hansen and Lunde (2005) find no consistent improvement over GARCH(1,1) among 330 specifications, while HAR-type models on intraday data routinely dominate the GARCH family.

Machine learning entered the field through the same door as the rest of empirical asset pricing (Gu, Kelly and Xiu, 2020). Christensen, Siggaard and Veliyev (2023) report that tree ensembles improve volatility forecasts when a rich feature set is available; Bucci (2020) documents gains from neural architectures on realized volatility. The evidence is not one-directional. Branco, Rubesam and Zevallos (2024), forecasting the realized volatility of ten global indices, find no evidence that nonlinear machine learning statistically beats linear models. This ambiguity is the point of departure for the present paper: if the aggregate verdict on "machine learning" is unstable across studies, one explanation is that the aggregate is not the right unit of analysis.

### *2.2. Model classes and the structure of the data*

That explanation is supported by a body of work outside finance. Comparative studies of tabular learning consistently find that gradient-boosted decision trees remain competitive with, and frequently superior to, deep architectures on heterogeneous tabular data (Shwartz-Ziv and Armon, 2022; Grinsztajn, Oyallon and Varoquaux, 2022), and surveys of deep learning on tabular problems reach compatible conclusions (Borisov et al., 2022). The proposed explanations are relevant to the present setting: tree ensembles are robust to uninformative features and to non-smooth target functions, and they do not require the sample sizes that high-capacity sequential models need in order to estimate their parameters. A volatility panel constructed from a few hundred engineered predictors on a market with seventeen liquid instruments is exactly the regime in which this literature predicts trees to win and sequence models to struggle. The forecasting results reported in Section 4 are consistent with that prediction, and the present paper contributes the observation that the same ordering survives into an economic evaluation.

Combination provides a further reference point. Bates and Granger (1969) established that a weighted combination of forecasts can dominate each of its constituents, and the subsequent literature has repeatedly found simple combinations difficult to beat (Genre et al., 2013). The hybrid model examined here is an instance of that principle applied across families rather than within one.

### *2.3. From forecast accuracy to economic value*

Whether a better forecast is worth having is a separate question from whether it is more accurate, and it has its own methodology. Fleming, Kirby and Ostdiek (2001; 2003) evaluate volatility timing through a mean-variance utility function and estimate that an investor would pay 50–200 basis points per year for improved covariance estimates, with realized volatility from intraday data outperforming daily-data alternatives as an input. Moreira and Muir (2017) show that scaling exposure by the inverse of recent variance earns alpha for major equity factors. Two qualifications follow. Harvey et al. (2018) conclude that the principal effect of volatility targeting across asset classes is the stabilisation of realized risk rather than the enhancement of returns; Cederburg et al. (2020), evaluating 103 equity strategies, find

that volatility-managed versions rarely outperform their unmanaged counterparts out of sample. Kaminski and Lo (2014) show that stop-loss rules add value under momentum dynamics but predict the magnitude to be modest.

What this literature has largely studied is a single channel — exposure scaling applied to a passive or factor portfolio. A practitioner operating a book of heterogeneous systematic strategies faces a richer menu, and the question of which mechanism carries the value is empirical. Lysenok (2026) addresses that question for the Russian market with the forecasting model held fixed. The present paper holds the channel structure fixed and varies the model, which is the complementary experiment and the one that determines whether investment in forecasting technology is justified.

### *2.4. Volatility research on the Russian market*

Early work applied the GARCH family to daily index data and established the significance of oil prices, the exchange rate and the policy rate as drivers (Fedorova and Nazarova, 2010), together with pronounced asymmetry of conditional variance in crisis periods. Asaturov and Teplova (2014) documented volatility spillovers from global markets to the Moscow Exchange using multivariate specifications, which motivates the inclusion of global assets in the feature space used here. The decisive shift came with Aganin (2017), who compared 88 GARCH specifications, 10 HAR-RV modifications and 4 ARFIMA models on ten Russian assets using 5-minute data and found HAR-RV models entering the Model Confidence Set; Aganin (2020) linked index volatility to oil and sanctions. Peresetsky and co-authors examined the Russian market in an international context and extracted a global stochastic trend from non-synchronous volatility observations (Pogorelova and Peresetsky, 2020). Recent Russian-language work has begun to apply neural architectures, though largely to index and ETF data (Glebova and Kovaleva, 2024; Patlasov, 2025). Gurov (2023) documents the microstructure constraints — wide spreads outside the most liquid names and measurable price impact even within the index — that govern the choice of sampling frequency for realized volatility on this market.

Two gaps remain. No study on Russian data compares econometric, tree-based and sequential architectures within a single design, and none evaluates competing forecasting models by their economic consequences rather than by a statistical loss. This paper addresses both.

## 3. Data and methodology

### *3.1. Sample and construction of the target variable*

The empirical base consists of 17 liquid stocks of the Moscow Exchange whose combined weight in the IMOEX index exceeds 80%. Inclusion required trading to have begun no later than June 2014, coverage by 10-minute data above 99%, and active trading at the end of the sample. The resulting set spans six sectors and is listed in Table 1.

*Table 1. Sample: 17 Moscow Exchange stocks*

| Ticker | Company | Sector |
|---|---|---|
| SBER | Sberbank | Financials |
| GAZP | Gazprom | Oil and gas |
| LKOH | LUKOIL | Oil and gas |
| GMKN | Norilsk Nickel | Metals and mining |
| ROSN | Rosneft | Oil and gas |
| NVTK | NOVATEK | Oil and gas |
| MGNT | Magnit | Consumer |
| MTSS | MTS | Telecommunications |

| | | |
|---|---|---|
| ALRS | ALROSA | Metals and mining |
| CHMF | Severstal | Metals and mining |
| NLMK | NLMK | Metals and mining |
| PLZL | Polyus | Metals and mining |
| TATN | Tatneft | Oil and gas |
| VTBR | VTB Bank | Financials |
| MOEX | Moscow Exchange | Financials |
| IRAO | Inter RAO | Utilities |
| SNGS | Surgutneftegas | Oil and gas |

*Source: compiled by the author.*

OHLCV quotations are obtained from the MOEX ISS API at 10-minute and daily frequencies. The choice of a 10-minute sampling interval is a compromise between the informativeness of the realized volatility estimate and the suppression of microstructure noise (Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2004); the 5-minute standard established for developed markets is not appropriate for an order book with the spread structure documented by Gurov (2023). Realized volatility for trading day t is computed as the sum of squared 10-minute logarithmic returns within the main session and aggregated to horizons h = 1, 5 and 22 days. Bars with zero volume, anomalous returns, violations of the OHLC ordering or time gaps exceeding ten minutes within trading hours are excluded from the computation, and days with less than 80% session coverage are dropped.

## *3.2. Feature space*

The predictor set comprises 234 variables organised in four blocks, summarised in Table 2. It includes HAR components (daily, weekly and monthly realized volatility, bipower variation and the jump component), intraday measures, lagged returns at horizons from 1 to 22 days, cross-sectional characteristics locating each asset relative to the others, and macroeconomic and global market variables.

*Table 2. Structure of the feature space*

| Data block | Source | Assets | Features |
|---|---|---|---|
| MOEX indices | MOEX ISS API | 11 | 99 |
| Global assets | Yahoo Finance | 10 | 80 |
| Macroeconomic indicators | CBR, FRED, Trading Economics | 8 | 32 |
| Own asset features | Computed | — | 23 |
| **Total** | | | **234** |

*Source: compiled by the author.*

## *3.3. Model families*

Seven models are compared, grouped into four families by the structure of the hypothesis class rather than by chronology.

**Econometric models.** GJR-GARCH (Glosten, Jagannathan and Runkle, 1993) models conditional variance parametrically with an asymmetry term. HAR-J (Corsi, 2009; Andersen, Bollerslev and Diebold, 2007) regresses realized volatility on its daily, weekly and monthly components together with a jump vector, and serves as the principal

benchmark: it is the specification that the comparative literature on this market places in the Model Confidence Set (Aganin, 2017).

**Gradient boosting.** XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) construct additive ensembles of decision trees on the full 234-variable feature space. Both are given identical inputs and an identical validation protocol, so that any difference between them is attributable to the implementation of the boosting procedure rather than to information.

**Recurrent neural networks.** LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) impose a sequential representation on the same features, processing them as an ordered window rather than as independent rows. They are the natural comparison class for the boosting models because they consume the same information under a different inductive bias.

**Hybrid model.** The hybrid combines the HAR-J forecast with that of the best boosting model at each horizon, with weights selected on a validation window and updated annually while the constituent models are not re-estimated. The weight on HAR-J rises with the horizon, from approximately one fifth at h = 1 to nearly one half at h = 22. Its construction follows Lysenok (2025) and is not re-derived here.

## *3.4. Evaluation protocol*

The chronological partition is strict. Models are trained on 2014–2016, validated on 2017–2018 and tested on a held-out year, 2019, which no model sees at any stage of specification or hyperparameter selection. Rolling validation with an expanding window is then conducted over 2017–2026, yielding 39,488 observations per ticker. All strategy evaluation is walk-forward over the control period 2022–2026, with model training and parameter calibration strictly preceding each evaluation year. The separation between the model evaluation window and the strategy evaluation window prevents information leakage between the two stages of the design.

The primary loss function is QLIKE, which preserves the true ranking of forecasts under an imperfect volatility proxy (Patton, 2011) and penalises underestimation of volatility more heavily than overestimation — the asymmetry that matters for risk management. Statistical significance of accuracy differences is assessed by the Diebold–Mariano test (Diebold and Mariano, 1995).

## *3.5. Trading strategies and integration channels*

Six directional strategies are grouped by the type of market inefficiency they exploit, as set out in Table 3. The counter-trend strategies rest on the mean-reversion evidence of Poterba and Summers (1988); the trend strategies on time-series momentum (Moskowitz, Ooi and Pedersen, 2012); the range strategies exploit intraday equilibrium levels.

*Table 3. Six systematic strategies*

| Strategy | Block | Signal | Risk management |
|---|---|---|---|
| S1 Mean Reversion | Counter-trend | Z-score vs moving average | ATR stop / take-profit |
| S2 Bollinger Bands | Counter-trend | Bollinger Bands | ATR stop / take-profit |
| S3 Donchian Channel | Trend | Channel breakout | Trailing stop |
| S4 Supertrend | Trend | Supertrend (ATR-based) | Trailing stop |
| S5 Pivot Points | Range | Pivot levels S1/R1 | Pivot / ATR stops |
| S6 VWAP | Range | VWAP ± deviation band | VWAP band / ATR stops |

*Source: compiled by the author.*

Each strategy is run under four regimes that differ in how, and whether, the volatility forecast enters the trading decision. Approach A is the base system: signal and risk parameters are optimised on training data and fixed, and no

forecast is used. Approach B is the trade-parameter channel: the forecast scales the stop-loss and take-profit distances proportionally, while the decision to enter remains unchanged (Kaminski and Lo, 2014). Approach C is the regime-filtering channel: the percentile rank of the forecast determines an admission mask, with counter-trend strategies permitted in low-volatility regimes and trend strategies in high-volatility regimes (Fleming, Kirby and Ostdiek, 2001). Approach D is the volatility-targeting channel: position size is scaled inversely to the forecast, with a binary gate at extreme levels (Moreira and Muir, 2017).

The distinction between B on one side and C and D on the other is the analytical core of the design. In B the forecast modifies how a trade is managed once it has been entered; in C and D it determines whether the trade occurs at all and how much capital it commits. If forecast accuracy has an economic value, the two groups of channels should not respond to it identically.

### *3.6. Portfolio aggregation and inference*

Results are aggregated at three levels. At the instrument level each of the 24 combinations of six strategies and four approaches is tested on 17 instruments by walk-forward validation with 10 optimisation phases for approaches B, C and D. At the strategy level the 17 instrument sleeves are combined into a portfolio under four weighting schemes — equal weights, inverse volatility, risk parity and maximum Sharpe — re-estimated annually. At the third level the six strategy portfolios are combined into a multi-strategy portfolio in the ALL6 × EW configuration. Free capital is placed in money-market instruments at the key rate of the Bank of Russia, which is material given that the rate reached 21% during the control period. Transaction costs of 0.05–0.06% per side are applied throughout, corresponding to institutional execution.

Inference on Sharpe ratio differences follows Ledoit and Wolf (2008) using a block bootstrap with 10,000 replications and stationary resampling (Politis and Romano, 1994). Economic value is measured by the certainty equivalent of a mean–variance investor (Fleming, Kirby and Ostdiek, 2001), $CE = r - (\gamma/2)\sigma^2$, evaluated at $\gamma = 3$, 5 and 10, with $\Delta CE$ defined as the difference between the certainty equivalent of a forecast-based approach and that of the base approach A.

## 4. Results

### *4.1. Accuracy at the model level*

Table 4 reports QLIKE for all seven models on the held-out test year at three horizons. Lower values indicate more accurate forecasts.

*Table 4. QLIKE of seven forecasting models, held-out test year 2019*

| Model | Family | h = 1 | h = 5 | h = 22 |
|---|---|---|---|---|
| **Hybrid** | Hybrid | **0.276** | **0.373** | **0.441** |
| XGBoost | Gradient boosting | 0.277 | 0.396 | 0.470 |
| LightGBM | Gradient boosting | 0.290 | 0.382 | 0.455 |
| **HAR-J** | Econometric (benchmark) | **0.305** | **0.424** | **0.467** |
| LSTM | Recurrent | 0.372 | 0.435 | 0.485 |
| GRU | Recurrent | 0.399 | 0.411 | 0.503 |
| GJR-GARCH | Econometric | 0.501 | 0.567 | 0.601 |

*Source: compiled by the author. Models ordered by QLIKE at h = 1; row shading denotes model family. The HAR-J row is the benchmark against which the other models are assessed.*

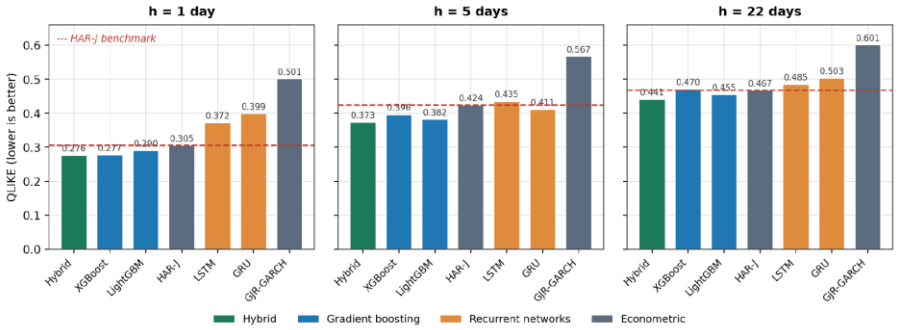


*Figure 1. QLIKE of seven forecasting models at three horizons, held-out test year 2019. The dashed line marks the HAR-J benchmark.*

Three observations follow. The hybrid model ranks first at every horizon. The gradient boosting models occupy the next two positions at the one-day horizon, improving on HAR-J by 9.2% (XGBoost) and 4.9% (LightGBM). The recurrent networks are placed between HAR-J and GJR-GARCH, and their deficit relative to HAR-J is substantial at the shortest horizon: 22.0% for LSTM and 30.8% for GRU. GJR-GARCH is the least accurate model at every horizon, trailing HAR-J by 64.3% at h = 1, which reproduces on Russian data the pattern established by Hansen and Lunde (2005) and by Aganin (2017).

The ordering is not stable across horizons, and the instability is informative. At h = 5 the ranking of the two boosting models reverses, and GRU (0.411) overtakes HAR-J (0.424) while LSTM does not. At h = 22 XGBoost (0.470) falls behind HAR-J (0.467), so that one of the two boosting models loses its advantage entirely at the monthly horizon. The advantage of tree ensembles is therefore concentrated where the information in the recent intraday history is richest, and it dissipates as the target is aggregated over longer windows and the mean-reverting component of volatility dominates.

## *4.2. Aggregation by model family*

Table 5 aggregates the same forecasts to the family level, which is the comparison of interest for a practitioner choosing an architecture rather than a specification.

*Table 5. QLIKE by model family, held-out test year 2019*

| Family | h = 1 | h = 5 | h = 22 | Mean |
|---|---|---|---|---|
| **Hybrid** | **0.276** | **0.373** | **0.441** | **0.363** |
| Gradient boosting | 0.284 | 0.389 | 0.463 | 0.379 |
| Recurrent networks | 0.386 | 0.423 | 0.494 | 0.434 |
| Econometric | 0.403 | 0.496 | 0.534 | 0.478 |

*Source: compiled by the author. Family values are unweighted means of the constituent models in Table 4.*

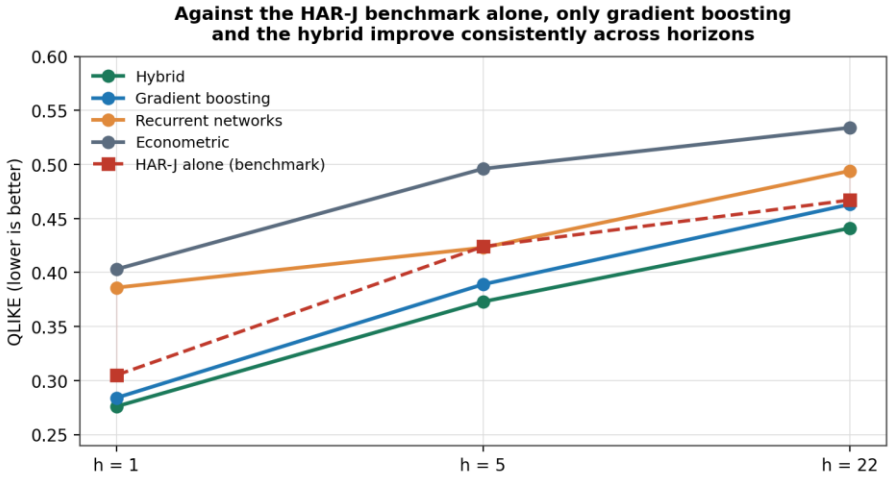


*Figure 2. Family means against the HAR-J benchmark. The econometric family mean is depressed by GJR-GARCH; measured against HAR-J alone, the recurrent family holds no advantage at any horizon.*

At the family level the ordering is unambiguous and identical at all three horizons: hybrid, then gradient boosting, then recurrent networks, then econometric models. Read without qualification, this supports the conventional claim that machine learning improves on econometric benchmarks. The qualification is essential, however, and it is visible only at the model level. The econometric family mean is depressed by GJR-GARCH, a model that the literature has known to be dominated for two decades. Measured against HAR-J alone — the benchmark that any serious study of this problem would adopt — the recurrent networks do not improve on econometrics at all: LSTM and GRU are worse than HAR-J at h = 1 and at h = 22, and only GRU is marginally better at h = 5. The family-level result "machine learning wins" is thus an artefact of including a weak parametric baseline in the comparison, and the defensible statement is narrower: gradient boosting improves on the strong econometric benchmark at short horizons, while recurrent networks do not improve on it at any horizon in this sample.

This is the sense in which "machine learning" is not a useful unit of analysis. Two families that consume identical features under identical protocols differ by 26% in QLIKE at the one-day horizon — a gap larger than that between either of them and the econometric benchmark.

## *4.3. Accuracy under rolling re-estimation*

The held-out test year is a single-period evaluation. Table 6 reports the rolling validation over 2017–2026 with annual re-estimation, which subjects each model to nine distinct market environments including the structural break of 2022. Recurrent networks were excluded from this stage; the implications are discussed in Section 6.

*Table 6. QLIKE under rolling validation, 2017–2026*

| Model | h = 1 | h = 5 | h = 22 | DM p-value vs HAR-J |
|---|---|---|---|---|
| **Hybrid** | **0.376** | **0.762** | 0.888 | **< 0.001 (all h)** |
| LightGBM | 0.378 | 0.821 | **0.870** | — |
| XGBoost | 0.400 | 0.956 | 1.127 | — |
| **HAR-J** | 0.524 | **0.796** | 0.961 | — |

*Source: rolling validation figures for the four re-estimated models follow Lysenok (2026); Diebold–Mariano t-statistics are 13.55, 3.41 and 7.03 at h = 1, 5 and 22 respectively.*

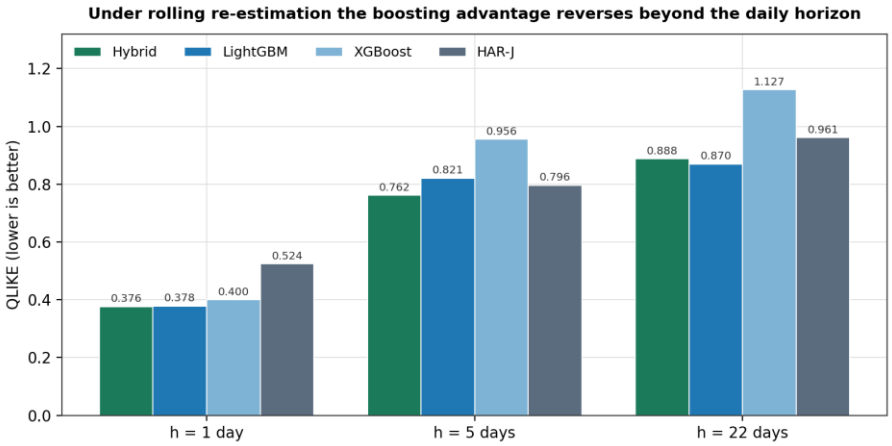


*Figure 3. QLIKE under rolling validation with annual re-estimation, 2017–2026.*

Re-estimation reverses part of the single-period picture. At h = 1 the hybrid and LightGBM are effectively tied (0.376 and 0.378) and both improve on HAR-J by approximately 28%, the largest accuracy gap in the study. At h = 5, however, HAR-J (0.796) outperforms both pure boosting models, and at h = 22 XGBoost (1.127) is the least accurate of the four, trailing HAR-J by 17%. The pattern is consistent with the interpretation offered above: the boosting models extract short-horizon information that the additive HAR cascade cannot represent, but they generalise less reliably when the target is aggregated and the effective sample of independent observations shrinks. The hybrid is the only specification that is at or near the top at every horizon, and its advantage over HAR-J is significant at all three by the Diebold–Mariano test ($p < 0.001$). Combination, rather than replacement, is what survives re-estimation.

## *4.4. From accuracy to strategy performance*

The remainder of the analysis asks what these accuracy differences are worth. Two models anchor the economic comparison: the hybrid, which is the most accurate specification under rolling re-estimation, and HAR-J, the econometric benchmark. At the one-day horizon that separates them by 28.2% of QLIKE (0.376 against 0.524), which is the largest accuracy differential available in this design and therefore the most favourable case for detecting an economic effect. Each model is passed through the three integration channels, and the resulting multi-strategy portfolios are compared on the control period 2022–2026. For channel C the strategy parameters are held identical across the two models and only the input forecast varies, which makes that comparison the cleanest of the three; for channels B and D parameters are optimised separately for each model on validation data before out-of-sample comparison.

*Table 7. Sharpe ratio of the multi-strategy portfolio: hybrid model against econometric benchmark, 2022–2026*

| Integration channel | Hybrid | HAR-J | Difference | p-value |
|---|---|---|---|---|
| A (no forecast) | 1.32 | 1.32 | — | not applicable |
| B (trade parameters) | 1.47 | 1.43 | +0.04 | 0.52 |
| **C (regime filtering)** | **2.51** | 1.70 | **+0.79** | **0.0002** |
| **D (volatility targeting)** | **2.54** | 1.69 | **+0.85** | **< 0.0001** |

*Source: compiled by the author. Significance assessed by block bootstrap with 10,000 replications (Ledoit and Wolf, 2008).*

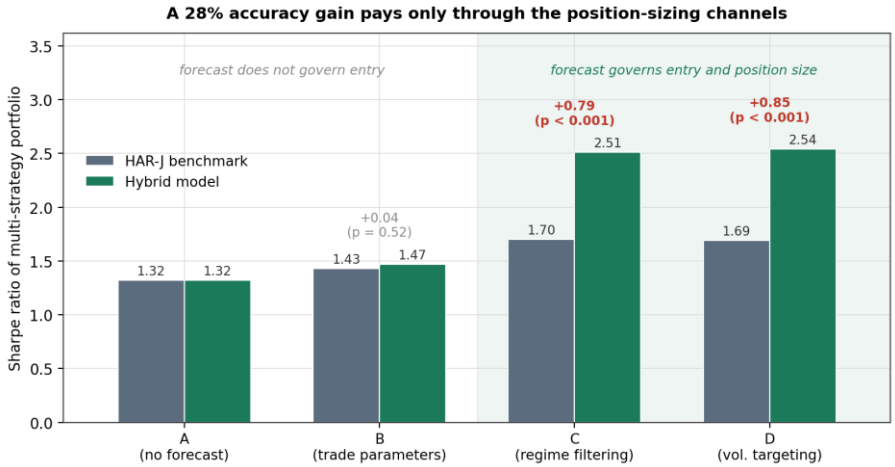


*Figure 4. Translation of forecast accuracy into portfolio performance by integration channel, 2022–2026. The same 28% improvement in QLIKE is worth +0.79 and +0.85 of Sharpe ratio where the forecast governs entry and position size, and nothing measurable where it does not.*

The result is sharply channel-dependent. In the two channels where the forecast governs the entry decision and the position size, replacing the econometric benchmark with the more accurate model raises the Sharpe ratio by +0.79 and +0.85, both significant at conventional levels. In the trade-parameter channel the same substitution produces +0.04, which cannot be distinguished from zero ($p = 0.52$). The accuracy improvement is identical in all three cases; what differs is the mechanism through which it reaches the portfolio.

The economic interpretation of the null result in channel B is not that the forecast is uninformative there, but that the channel is insensitive to the information. Widening or narrowing a stop-loss distance modifies the management of a position whose existence has already been determined by the technical signal; the parameter optimisation in approach B can compensate for a less accurate forecast by recalibrating the response function, and it does. Channels C and D admit no such compensation, because the forecast determines whether capital is committed at all.

Table 8 places these portfolios in context against the passive alternative.

*Table 8. Characteristics of the multi-strategy portfolio and the IMOEX index (ALL6 × EW, 2022–2026)*

| Portfolio | Sharpe | Return, % | Vol., % | MDD, % | Exposure, % |
|---|---|---|---|---|---|
| A (no forecast) | 1.32 | 19.6 | 4.4 | −2.8 | 15.7 |
| B (trade parameters) | 1.47 | 19.5 | 3.8 | −1.6 | 13.0 |
| C (regime filtering) | 2.51 | 21.2 | 3.0 | −0.9 | 7.2 |
| D (volatility targeting) | 2.54 | 20.9 | 2.9 | −1.1 | 12.2 |
| **IMOEX buy-and-hold** | **−0.69** | **−7.1** | **30.8** | **−70.3** | **100** |

*Source: figures for the multi-strategy portfolios and the index follow Lysenok (2026).*

Two features of Table 8 bear on the interpretation. First, the improvement delivered by channels C and D is achieved through a simultaneous reduction in volatility and drawdown with no loss of return, which is the risk-management signature identified by Harvey et al. (2018) rather than a return-enhancement effect. Second, channel C attains the higher Sharpe ratio while halving average exposure, from 15.7% to 7.2% of the time in position. Trading

less than half as often without losing return implies that the filter removes predominantly unprofitable entries, which is the behaviour a genuine forecast-driven regime filter should exhibit and which a mechanical reduction of exposure would not produce.

## *4.5. Economic value for a mean-variance investor*

A difference in Sharpe ratios does not by itself establish that an investor would pay for the forecasting system. Table 9 reports the increment in the certainty equivalent when moving from the base approach to each forecast-based channel, at three levels of risk aversion.

*Table 9. Economic value of the integration channels (ΔCE, % per annum)*

| γ | B: ΔCE | p | C: ΔCE | p | D: ΔCE | p |
|---|---|---|---|---|---|---|
| 3 | −0.09% | 0.59 | +1.77% | 0.03 | +1.46% | 0.07 |
| 5 | −0.05% | 0.55 | +1.86% | 0.02 | +1.56% | 0.05 |
| 10 | +0.04% | 0.45 | **+2.07%** | **0.01** | **+1.79%** | **0.02** |

*Source: figures follow Lysenok (2026). Bootstrap with 10,000 replications.*

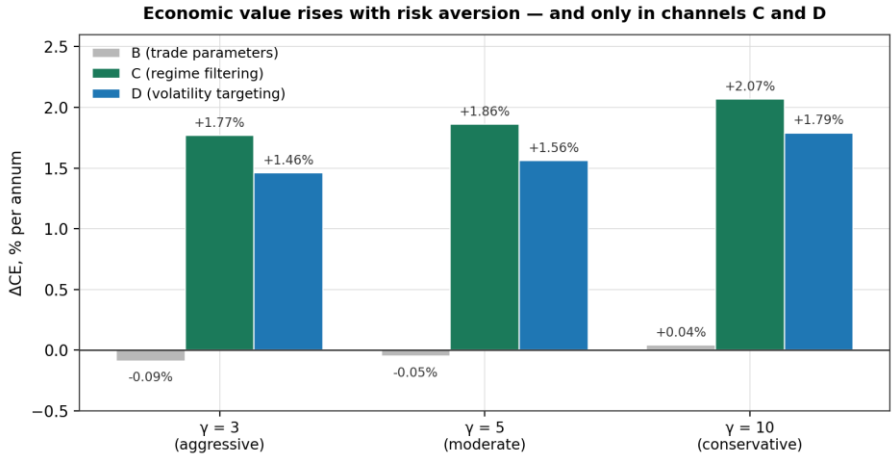


*Figure 5. Increment in certainty-equivalent return by integration channel and level of risk aversion.*

The pattern in Table 9 corroborates the channel result at the level of investor utility. Channels C and D deliver statistically significant economic value at every level of risk aversion, reaching +2.07% and +1.79% per annum at γ = 10, while channel B delivers nothing distinguishable from zero at any level. The monotonic increase of ΔCE in γ indicates that the gain is generated predominantly through the reduction of portfolio variance rather than through higher average return, which is consistent with the exposure and volatility figures in Table 8. The magnitudes sit at the upper end of the 50–200 basis point range estimated by Fleming, Kirby and Ostdiek (2001) for developed markets, which is unsurprising given the level of the risk-free rate in the control period.

## *4.6. Summary of findings*

The findings can be stated as three propositions. First, the accuracy advantage attributed in the literature to "machine learning" is carried in this sample by gradient boosting alone; recurrent networks consuming the same features do not improve on the strong econometric benchmark at any horizon, and the gap between the two machine learning families exceeds the gap between either and the benchmark. Second, the advantage of boosting is horizon-dependent and weakens under re-estimation, whereas a hybrid of the econometric and boosting forecasts is at or near the top at every horizon; combination dominates replacement. Third, a 28% improvement in forecast accuracy is worth +0.79 to +0.85 of Sharpe ratio and up to +2.07% of certainty-equivalent return per year when the forecast governs

entry and position size, and is worth nothing measurable when it only scales the parameters of a trade whose entry has already been decided.

## 5. Discussion

### *5.1. Why gradient boosting outperforms recurrent networks here*

The deficit of the recurrent architectures is large enough to require explanation, particularly since they consume exactly the same predictors as the boosting models. Three properties of the problem favour tree ensembles, and all three are recognised in the literature on tabular learning (Shwartz-Ziv and Armon, 2022; Grinsztajn, Oyallon and Varoquaux, 2022; Borisov et al., 2022).

The first is sample size relative to parametric capacity. The panel comprises seventeen instruments over a training window of three years, and the sequential models must estimate substantially more parameters than the boosting models from that data. The second is the structure of the feature space. Of the 234 predictors, a considerable share is uninformative or nearly collinear; tree ensembles select features implicitly at each split and are robust to the presence of irrelevant inputs, whereas a recurrent network must learn to suppress them through its weights. The third is the shape of the target function. Volatility responds to its own recent history in a manner well approximated by piecewise-constant partitions of the predictor space, which is precisely the hypothesis class that trees represent efficiently and that smooth sequential models approximate only with additional capacity.

There is also a specific consideration for this market. The training window contains a structural break — the market disruption of 2022 — whose magnitude relative to the length of the sample is unusually large. A high-capacity sequential model exposed to such a break with limited data has more scope to fit the break itself rather than the regularity, and less opportunity to recover through re-estimation than a model with fewer parameters. This suggests that the ordering reported here is a property of the data regime rather than of the architectures in general, and that it should be expected to weaken on markets where longer samples of comparable quality are available.

### *5.2. Horizon dependence and the case for combination*

The reversal of the boosting advantage between h = 1 and h = 22 is the second result that resists a simple reading. At the daily horizon the target is dominated by the recent intraday history, where the boosting models exploit interactions among the realized measures that the additive HAR cascade cannot represent. As the horizon lengthens the target increasingly reflects the mean-reverting component of volatility, which the HAR hierarchy captures by construction, and the marginal information available to a flexible learner declines while its variance does not. Under rolling re-estimation this produces the pattern in Table 6: HAR-J ahead of both boosting models at h = 5, and XGBoost last at h = 22.

The hybrid resolves this by shifting weight toward the econometric component as the horizon lengthens — from approximately one fifth at h = 1 to nearly one half at h = 22 — which is the empirical counterpart of the theoretical argument of Bates and Granger (1969). The practical implication for system design is that the choice of forecasting model should be indexed to the horizon at which the trading system consumes the forecast, and that a single model selected on aggregate accuracy may be the wrong choice at both ends of the horizon range.

### *5.3. The channel is the binding constraint*

The most consequential result for practice is the asymmetry between the integration channels, because it bounds the return on investment in forecasting technology. A development team can improve QLIKE by a quarter and observe no effect whatever on the portfolio if the forecast enters the system only through the parameters of individual trades. The same improvement, routed through a regime filter or a position-sizing rule, produces an effect large enough to justify the cost of the infrastructure that generates it.

This extends the findings of Moreira and Muir (2017) and Harvey et al. (2018), who evaluate exposure scaling applied to factor or asset-class portfolios without decomposing the mechanism of transmission, and it qualifies the

caution of Cederburg et al. (2020) by locating the condition under which volatility management does add value. It also aligns with the prediction of Kaminski and Lo (2014) that the gain from stop-loss rules should be modest: the observed +0.04 of Sharpe ratio in channel B is of the predicted order of magnitude.

For a practitioner the sequence of decisions therefore runs in the opposite direction to the usual research workflow. The first question is not which model forecasts best, but whether the trading system contains a channel through which forecast quality can propagate to the allocation of capital. Where it does not, the appropriate investment is in the architecture of the system rather than in the model; where it does, the evidence here indicates that moving from a strong econometric benchmark to a hybrid specification is worth approximately two percentage points of certainty-equivalent return per year for a conservative investor.

### *5.4. The interest rate environment*

The control period is characterised by a policy rate that reached 21% and remained above 16% from late 2023, and this shapes the interpretation of the results in two ways. It raises the hurdle for any active equity position, so that the strategies studied here are in the market between 7% and 16% of the time and the released capital earns the money-market rate. This makes the setting a conservative test of forecast value: the forecast must justify entry against an alternative that is itself attractive. At the same time it magnifies the value of a filter that correctly declines to trade, since capital withheld is not idle. The result of Table 9 — that channel C attains the highest certainty equivalent while holding the lowest average exposure — is a joint product of forecast quality and the rate environment, and the decomposition of these two contributions is not attempted here.

## 6. Limitations and directions for further research

Four limitations qualify the conclusions, and each defines a direction for subsequent work.

**Scope of the economic comparison.** The accuracy comparison covers seven models, but the economic comparison covers two: the most accurate specification and the econometric benchmark. It became apparent at the analysis stage that extending the economic evaluation to the full model × channel matrix is a separate computational exercise rather than an incremental one, because channels B and D require the strategy parameters to be re-optimised under each forecasting model, whereas channel C admits a direct substitution of the forecast with parameters held fixed. The two models compared here are the extremes of the accuracy range and therefore provide the strongest available test of transmission, but the shape of the relationship between accuracy and economic value — in particular, whether it is monotone across the intermediate models — is not established by this design. Establishing it is the natural next step, and channel C offers the cheapest route to it.

**Recurrent networks under re-estimation.** LSTM and GRU are evaluated on the held-out test year but were excluded from the rolling validation, so their performance under annual recalibration on an expanding sample is unknown. Given that the proposed explanation for their deficit is the ratio of sample size to parametric capacity, an expanding-window protocol is precisely the condition under which that deficit should narrow. The comparison reported here is therefore conservative with respect to the recurrent family, and a re-estimated evaluation could revise the family ordering at longer samples.

**Sample and market coverage.** The study covers seventeen liquid instruments on a single market, restricted to continuously traded large capitalisations. This is a selection choice: it excludes instruments that ceased trading and it excludes the less liquid segment where the microstructure noise documented by Gurov (2023) is more severe. The concentration of the Russian market makes the sample economically representative of the index, but the results should not be extrapolated to the second and third tiers without re-estimation, nor to markets with a different rate environment.

**Architectures and information sources.** The comparison covers three architectural families and omits others of current interest, notably CatBoost among boosting implementations and attention-based sequence models. The feature space, while large, is constructed exclusively from price, volume and macroeconomic series; order-book depth, news sentiment and other alternative sources are absent. Both extensions bear directly on the explanation advanced in

Section 5.1, since an architecture designed for sparse high-dimensional inputs may perform differently in this data regime than the recurrent models evaluated here.

## 7. Conclusion

This paper compared seven realized volatility forecasting models from four architectural families on 10-minute data for seventeen liquid Moscow Exchange stocks, and then evaluated what their accuracy differences are worth when the forecasts are routed into a book of six systematic trading strategies through three distinct integration channels.

The forecasting results do not support the undifferentiated proposition that machine learning improves on econometric benchmarks. Gradient boosting improves on the HAR-J benchmark by 5–9% in QLIKE at the one-day horizon, but the advantage weakens with the horizon and, under rolling re-estimation, reverses: HAR-J outperforms both boosting models at h = 5 and XGBoost is the least accurate of the re-estimated models at h = 22. Recurrent networks, consuming an identical feature space, are 22–31% worse than HAR-J at the one-day horizon and improve on it at no horizon on the held-out test year. The gap between the two machine learning families is larger than the gap between either of them and the econometric benchmark, which is the sense in which the family label is the wrong unit of comparison. Only the hybrid of econometric and boosting forecasts ranks at or near the top at every horizon and improves on HAR-J significantly under re-estimation at all three ($p < 0.001$).

The economic results establish that the value of accuracy is conditional on the architecture of the system consuming the forecast. A 28% reduction in QLIKE raises the Sharpe ratio of the multi-strategy portfolio by +0.79 under regime filtering and +0.85 under volatility targeting, both significant at $p < 0.001$, and the corresponding certainty-equivalent gains reach +2.07% and +1.79% per annum for a conservative investor. The same improvement, routed through the scaling of trade parameters, yields +0.04 of Sharpe ratio and no measurable economic value at any level of risk aversion. Statistical accuracy is thus a valid proxy for economic value in this setting, but only where the forecast governs the decision to commit capital and the amount committed.

For system design the implication is an ordering of priorities. Establishing that a position-sizing or regime-filtering channel exists precedes the choice of forecasting model, because in its absence no improvement in forecast quality will reach the portfolio. Where such a channel exists, the evidence indicates that combination across model families is preferable to replacement of one family by another, and that the choice of model should be indexed to the forecast horizon the system actually consumes.

**REFERENCES**


1. Aganin A.D. (2017). Comparison of GARCH and HAR-RV models for forecasting realized volatility in the Russian market. Applied Econometrics, 48, 63–84.
2. Aganin A.D. (2020). Volatility of the Russian stock index: oil and sanctions. Voprosy Ekonomiki, 2, 86–100.
3. Andersen T.G., Bollerslev T. (1998). Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. International Economic Review, 39(4), 885–905.
4. Andersen T.G., Bollerslev T., Diebold F.X. (2007). Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility. The Review of Economics and Statistics, 89(4), 701–720.
5. Andersen T.G., Bollerslev T., Diebold F.X., Labys P. (2001). The Distribution of Realized Exchange Rate Volatility. Journal of the American Statistical Association, 96(453), 42–55.
6. Andersen T.G., Bollerslev T., Diebold F.X., Labys P. (2003). Modeling and Forecasting Realized Volatility. Econometrica, 71(2), 579–625.
7. Asaturov K.G., Teplova T.V. (2014). Construction of hedging coefficients for highly liquid stocks of the Russian market based on GARCH class models. Economics and Mathematical Methods, 50(1), 37–54.
8. Barndorff-Nielsen O.E., Shephard N. (2004). Power and Bipower Variation with Stochastic Volatility and Jumps. Journal of Financial Econometrics, 2(1), 1–37.
9. Bates J.M., Granger C.W.J. (1969). The Combination of Forecasts. Journal of the Operational Research Society, 20(4), 451–468.
10. Bollerslev T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

11. Borisov V., Leemann T., Seßler K., Haug J., Pawelczyk M., Kasneci G. (2022). Deep Neural Networks and Tabular Data: A Survey. IEEE Transactions on Neural Networks and Learning Systems.
12. Branco R.R., Rubesam A., Zevallos M. (2024). Forecasting Realized Volatility: Does Anything Beat Linear Models? Journal of Empirical Finance, 78, 101524.
13. Bucci A. (2020). Realized Volatility Forecasting with Neural Networks. Journal of Financial Econometrics, 18(3), 502–531.
14. Cederburg S., O'Doherty M.S., Wang F., Yan X.S. (2020). On the Performance of Volatility-Managed Portfolios. Journal of Financial Economics, 138(1), 95–117.
15. Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
16. 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. Proceedings of EMNLP 2014, 1724–1734.
17. Christensen K., Siggaard M., Veliyev B. (2023). A Machine Learning Approach to Volatility Forecasting. Journal of Financial Econometrics, 21(5), 1680–1727.
18. Corsi F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. Journal of Financial Econometrics, 7(2), 174–196.
19. Diebold F.X., Mariano R.S. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.
20. Engle R.F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50(4), 987–1007.
21. Fedorova E.A., Nazarova Yu.N. (2010). Identification of factors influencing stock market volatility using a cointegration approach. Economic Analysis: Theory and Practice, (3), 17–24.
22. Fleming J., Kirby C., Ostdiek B. (2001). The Economic Value of Volatility Timing. The Journal of Finance, 56(1), 329–352.
23. Fleming J., Kirby C., Ostdiek B. (2003). The Economic Value of Volatility Timing Using "Realized" Volatility. Journal of Financial Economics, 67(3), 473–509.
24. Genre V., Kenny G., Meyler A., Timmermann A. (2013). Combining Expert Forecasts: Can Anything Beat the Simple Average? International Journal of Forecasting, 29(1), 108–121.
25. Glebova A.G., Kovaleva A.A. (2024). Forecasting the volatility of the Russian stock market in the context of international economic sanctions. Finance: Theory and Practice, 28(1), 20–29.
26. Glosten L.R., Jagannathan R., Runkle D.E. (1993). On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. The Journal of Finance, 48(5), 1779–1801.
27. Grinsztajn L., Oyallon E., Varoquaux G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35.
28. Gu S., Kelly B., Xiu D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223–2273.
29. Gurov S.V. (2023). Illiquidity effects on the Russian stock market. Economic Journal of the Higher School of Economics, 27(1).
30. Hansen P.R., Lunde A. (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? Journal of Applied Econometrics, 20(7), 873–889.
31. Harvey C.R., Hoyle E., Korgaonkar R., Rattray S., Sargaison M., Van Hemert O. (2018). The Impact of Volatility Targeting. The Journal of Portfolio Management, 45(1), 14–33.
32. Hochreiter S., Schmidhuber J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
33. Kaminski K.M., Lo A.W. (2014). When Do Stop-Loss Rules Stop Losses? Journal of Financial Markets, 18, 234–254.
34. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30, 3149–3157.
35. Ledoit O., Wolf M. (2008). Robust Performance Hypothesis Testing with the Sharpe Ratio. Journal of Empirical Finance, 15(5), 850–859.
36. Lysenok N.I. (2025). Application of machine learning for volatility forecasting and trading strategy enhancement on the Russian stock market. Fundamental and Applied Mathematics, 25(4), 90–107.
37. Lysenok N.I. (2026). Effectiveness of volatility forecasts in active trading strategies of institutional investors on the Russian equity market. Fundamental and Applied Mathematics, 26(3), 33–42.
38. Moreira A., Muir T. (2017). Volatility-Managed Portfolios. The Journal of Finance, 72(4), 1611–1644.
39. Moskowitz T.J., Ooi Y.H., Pedersen L.H. (2012). Time Series Momentum. Journal of Financial Economics, 104(2), 228–250.
40. Patlasov D.A. (2025). Hybrid approaches to forecasting realized ETF volatility: deep learning and the recovery theorem. Economic Journal of the Higher School of Economics, 29(1), 103–131.
41. Patton A.J. (2011). Volatility Forecast Comparison Using Imperfect Volatility Proxies. Journal of Econometrics, 160(1), 246–256.
42. Pogorelova P.V., Peresetsky A.A. (2020). Isolation of a global stochastic trend from non-synchronous observations of financial index volatility. Applied Econometrics, 57, 53–71.

43. Politis D.N., Romano J.P. (1994). The Stationary Bootstrap. Journal of the American Statistical Association, 89(428), 1303–1313.
44. Poterba J.M., Summers L.H. (1988). Mean Reversion in Stock Prices: Evidence and Implications. Journal of Financial Economics, 22(1), 27–59.
45. Sharpe W.F. (1994). The Sharpe Ratio. The Journal of Portfolio Management, 21(1), 49–58.
46. Shwartz-Ziv R., Armon A. (2022). Tabular Data: Deep Learning is Not All You Need. Information Fusion, 81, 84–90.