Do Machine-Learning Models Beat Simple Benchmarks in Forecasting Next-Day Absolute Price Changes? Evidence from the Tehran Stock Exchange

Ali Soltani ghombavani[1], Amirmohammad Rezaei[2], MohamadAhadzadeh
[1]Department of Accounting, Alborz Campus, University of Tehran
[2]Department of Accounting, Ma.C., Islamic Azad University, Mashhad, Iran
[3]Department of Accounting, Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract
Forecasting the size of daily price movements matters for risk management, yet evidence on how well machine-learning models perform out of sample in emerging markets is scarce. This study asks whether machine-learning models can forecast the next-day absolute percentage change of the Tehran Stock Exchange (TSE) index, defined as the absolute open-to-close percentage change and used as a magnitude-based proxy for realized volatility. Using 1,512 daily observations and a strictly chronological 80/20 split, we compare five approaches (K-Means clustering, a rule-based classifier, a decision tree, gradient boosting regression (GBR), and support vector regression (SVR)) with five benchmarks, including a HAR-type regression and GARCH(1,1). Accuracy is measured by the out-of-sample R^2 relative to the historical mean (R^2OS). In the final test block, which is calmer than the training period, R^2OS is 10.7% for HAR, 7.8% for GARCH, 6.4% for GBR, and −0.7% for SVR; in a rolling-origin evaluation with 756 pooled forecasts it is 13.8%, 12.6%, 11.2%, and 6.8%. No machine-learning model using the eight lagged predictors improves on HAR, which additionally uses a 22-day average and is significantly better than GBR ($p = 0.029$) and SVR ($p = 0.006$) in the rolling evaluation. The decision tree is less accurate than always predicting the low-movement class (69.0% against 75.6%), and K-Means clusters carry no information on next-day movement. Same-day information, in contrast, explains over 90% of the variance mechanically, which cautions against high R^2 values obtained with contemporaneous inputs.

Uncorrected Proof

Keywords: Machine Learning, Absolute Percent Change, Volatility Proxy, Forecast Evaluation, Emerging Market

1. Introduction

The Tehran Stock Exchange (TSE) is a pivotal financial market that significantly impacts Iran's economy. Stock markets, including the TSE, experience considerable fluctuations driven by multifaceted and often unpredictable factors, making precise predictions of market movements a challenging yet highly sought-after task for investors and policymakers alike. In particular, predicting market volatility has gained considerable attention due to its importance in investment decisions, pricing of securities, risk management, and monetary policy formulation. Financial market volatility can also affect public confidence and overall economic stability (Tehrani et al., 2011).

Machine learning (ML) techniques have emerged as powerful tools for tackling financial forecasting challenges, with substantial success in diverse domains such as natural language processing, especially in tasks that involve sequence learning. These advancements have sparked interest in their application to financial markets, with early studies suggesting that machine learning methods could help uncover the underlying nonlinear structures in financial markets (Huck, 2009). Despite their widespread success in various sectors, including computer science and statistics, machine learning methods have also become an essential empirical tool in finance, gaining popularity as a reliable approach for modeling market behavior (Diaz et al., 2024).

Over recent decades, numerous studies have been conducted to examine volatility in financial markets, emphasizing its importance in understanding fluctuations in asset prices, risk management, and strategic decision-making in investment (Yan & Li, 2024; Mensah et al., 2025; Zhang et al., 2024; Letteri, 2024; Lopez et al., 2025; Diaz et al., 2024; Campisi et al., 2024). One of the key challenges in financial markets is forecasting volatility, which remains a critical area of research due to its significant impact on market behavior. Volatility indices have been proven to provide essential insights into current and future stock returns, with some studies indicating that high volatility levels might signal oversold markets, suggesting potential buy signals in the short to medium term (Giot, 2005).

Accurate predictions of market volatility are crucial for investors, as they enable more informed decision-making while mitigating exposure to potential risks and enhancing trading strategies. The predictive power of machine learning has recently improved considerably, helping investors better anticipate market fluctuations and achieve more reliable investment decisions (Adebayo et al., 2025; Sadhukhan et al., 2022).

The TSE is a suitable setting for this question because several institutional features shape the distribution of daily price movements. Daily price-limit bands cap how far a price can move within a session and can spread large information shocks over several days; thin trading and liquidity constraints in parts of the market, trading halts, and recurring political and geopolitical shocks, including sanctions-related news, produce abrupt shifts in activity. As a result, the magnitude of daily moves is persistent, truncated, and heterogeneous across regimes, which are patterns that flexible nonlinear methods may capture and that rigid parametric specifications may not.

Volatility in the TSE has traditionally been modeled with GARCH-family and related econometric specifications (Bollerslev, 1986; Tehrani et al., 2011). These models impose a parametric structure on conditional variance. Truncation at price limits, regime shifts, and interactions between the intraday range, overnight gaps, and past absolute movements are difficult to represent in that structure. Whether machine-learning models can capture such features and improve on simple persistence-based benchmarks is an empirical question that, to our knowledge, has received limited attention for the magnitude of TSE movements (Section 2.2). This paper addresses it by (i) forecasting the next-day absolute percentage change of the TSE index with five models, (ii) evaluating them strictly out of sample under a chronological design, and (iii) comparing them with simple benchmarks using formal forecast-comparison tests.

Uncorrected Proof

Terminology. Throughout the paper, absolute percentage change (AbsMov) denotes the absolute value of the daily open-to-close percentage change of the index. It measures the magnitude of price movement and is used as a proxy for realized volatility, which is not directly observable (Hördahl & Packer, 2007). It is not the conditional variance targeted by GARCH-type models. The word volatility is used only when discussing the literature or the intraday range measure (PV) defined in Section 3.4.

The study makes three contributions. First, it provides a benchmarked, leakage-controlled evaluation of five machine-learning approaches for the next-day magnitude of movements in an emerging market with price limits. Second, it reports the result that the data support, namely that none of the models improves on a simple HAR-type regression out of sample, together with a transparent account of why. Third, it shows how strongly same-day information inflates apparent accuracy, which is relevant for interpreting high reported R^2 values. The study addresses three research questions:

RQ1: Do machine-learning models forecast next-day absolute percentage changes of the TSE index more accurately out of sample than simple benchmarks?

RQ2: Does GBR forecast more accurately than SVR?

RQ3: How well do rules, clusters, and a decision tree separate low- from high-movement days out of sample?

2. Literature Review

This section synthesizes research on the model families used here, summarizes the evidence for the TSE, positions the study against prior work, and states the expectations that guide the analysis.

2.1. Machine Learning for Forecasting Absolute Movements and Volatility

Studies fall into three groups that correspond to the paradigms combined in this paper. The first uses supervised regression to forecast the level of volatility. Tree ensembles such as random forests and gradient boosting outperform classical EWMA and GARCH forecasts out of sample for the S&P 500 and reveal which inputs matter (Lolic, 2025). Stacked combinations of gradient boosting, random forests, support vector machines, and neural networks improve S&P 500 volatility forecasts (Ramos-Pérez et al., 2019), and gains over conventional techniques are attributed to the ability to capture nonlinear relations (Christensen et al., 2023; Kumar et al., 2022). Recurrent networks have also been applied (Pattanayak et al., 2024), including to range-based volatility estimators (Petneházi & Gáll, 2019).

The second group frames the task as classification of movement size or direction. Rodriguez et al. (2024) predict the daily absolute percent change of the S&P 500 with decision trees, kernel methods, rule-based classifiers, and K-Means clustering, reporting accuracies of about 80% for decision trees and kernels and 88% for rule-based classifiers. Campisi et al. (2024) compare machine-learning methods for predicting the direction of the US market from volatility indices, and Srivastava et al. (2022) design machine-learning classifiers for stock-market prediction. Because accuracy depends on how classes are defined and on class balance, such figures are not comparable across studies, and this paper reports precision, recall, and F1 alongside accuracy.

The third group asks which information improves forecasts beyond prices. News sentiment is used to predict prices (Gite et al., 2021) and improves volatility prediction in deep-learning models (Yao et al., 2022); other work uses Google domestic trends in hybrid models of index volatility (Seo et al., 2019) or intraday commonality (Zhang et al., 2024). Our data contain only price information, so this stream is not exploited here (see Section 5.3). Finally, validation that respects time order, such as forward testing (Letteri et al., 2022), is essential because random splits leak future information; this study uses a chronological split and time-series cross-validation.

2.2. Evidence from the TSE and Comparable Markets

Evidence for the TSE on volatility is dominated by econometric models. Tehrani et al. (2011) compare conditional and non-conditional models of TEDPIX volatility at daily frequency using RMSE. Among the two groups, the non-conditional MA250 and the conditional CGARCH performed best, non-conditional models performed better than conditional ones in combined models, and a Diebold–Mariano test found no significant difference between MA250 and CGARCH. Two lessons follow: simple benchmarks can be hard to beat, so machine-learning forecasts should be compared with them formally (as done here), and evidence on machine-learning forecasts of the magnitude of TSE movements, benchmarked against volatility models, appears limited.

Machine learning has been applied to the TSE, but to price levels and trends rather than to movement magnitude. Nabipour et al. (2020) forecast the values of four TSE sector groups (diversified financials, petroleum, non-metallic minerals, and basic metals) one to 30 days ahead from ten technical indicators, using tree-based models and neural networks (decision tree, bagging, random forest, AdaBoost, gradient boosting, XGBoost, ANN, RNN, and LSTM), and report that LSTM is the most accurate. Their targets are sector-index values, not absolute movements, and, to our reading, their evaluation does not include volatility benchmarks such as HAR or GARCH.

Uncorrected Proof

2.3. Research Gap and Positioning

Table 1 positions the present study. Existing machine-learning studies of absolute movements use developed-market indices; TSE evidence either relies on econometric models without machine-learning comparators (Tehrani et al., 2011) or applies machine learning to price levels and trends of sector indices (Nabipour et al., 2020). The gap addressed here is the absence of a leakage-controlled, chronologically evaluated analysis of machine-learning forecasts of next-day absolute movements in an emerging market with price limits, benchmarked against HAR- and GARCH-type models.

Table 1. Positioning of the Present Study Relative to Prior Research

Study	Market / data	Dependent variable	Method	Main limitation	How the present study differs
Andersen et al. (2001)	DJIA stocks; high-frequency intraday data	Realized daily volatilities and correlations	Realized-variance measures; distributional analysis	Describes distribution and persistence; no machine-learning forecasting; developed market	Machine-learning forecasts of next-day absolute change in an emerging market with price limits
Tehrani et al. (2011)	TEDPIX, daily data	Volatility of TEDPIX	Conditional (e.g., CGARCH) and non-conditional (e.g., MA250) models; RMSE and DM test	No machine-learning models	Adds nonlinear machine-learning models, rules, and regime analysis
Ramos-Pérez et al. (2019)	S&P 500	Volatility	Stacked GB, RF, SVM and ANN	Developed-market index only	TSE; adds regime and rule-based analysis
Rodriguez et al. (2024)	S&P 500	Daily absolute percent change	Decision trees, rule-based classifiers, K-Means, kernel methods	Developed-market index; accuracy depends on class definition	TSE; adds gradient boosting and a chronological leakage-controlled evaluation
Lolic (2025)	S&P 500	Volatility	Random forest and gradient boosting vs. EWMA and GARCH	Developed-market index only	TSE; absolute-movement target with classification and regime analysis
Nabipour et al. (2020)	TSE; four sector groups; ten years of daily data	Future values of sector indices, one to 30 days ahead	Decision tree, bagging, random forest, AdaBoost, gradient boosting, XGBoost, ANN, RNN, LSTM; ten technical indicators	Targets index levels rather than movement magnitude; volatility benchmarks not reported (to our reading)	Total index; absolute next-day movement; HAR and GARCH benchmarks, DM tests, leakage-controlled design

Source: Authors' compilation from the cited studies.

Uncorrected Proof

2.4. Expectations Guiding the Analysis

The reviewed evidence leads to three expectations that the empirical analysis examines. First, tree ensembles and kernel methods can represent nonlinear dependence in absolute price movements (Christensen et al., 2023; Lolic, 2025; Rodriguez et al., 2024), so out-of-sample forecasts of next-day magnitude should have explanatory power beyond simple benchmarks (RQ1). Second, gradient boosting handles heterogeneous predictors, threshold effects, and interactions without distance-based scaling, whereas SVR depends on scaling and kernel choice, and tree ensembles have outperformed classical methods in comparable settings (Lolic, 2025); GBR should therefore forecast more accurately than SVR (RQ2). Third, absolute movements are strongly persistent (Andersen et al., 2001), so rules and clusters built on lagged absolute movements should separate calm from active periods (RQ3). The results in Section 4 test these expectations out of sample and do not support the first two.

2.5. Analytical Framework

Figure 1 summarizes the framework. Daily open, high, low, and close data are transformed into lagged features (percentage changes, intraday range, and moving averages). The same features feed five models: K-Means (regime identification), a rule-based classifier (interpretable benchmark), a decision tree (movement class), and GBR and SVR (movement magnitude). Models are fitted on the first 80% of observations and evaluated once on the last 20%, and their forecasts are compared with five benchmarks. The models run in parallel on the same data and split, and no model uses another model's output. We call the framework hybrid in the sense that it combines unsupervised, rule-based, and supervised classification and regression approaches; it is a comparison of complementary approaches, not a stacked or sequential architecture.

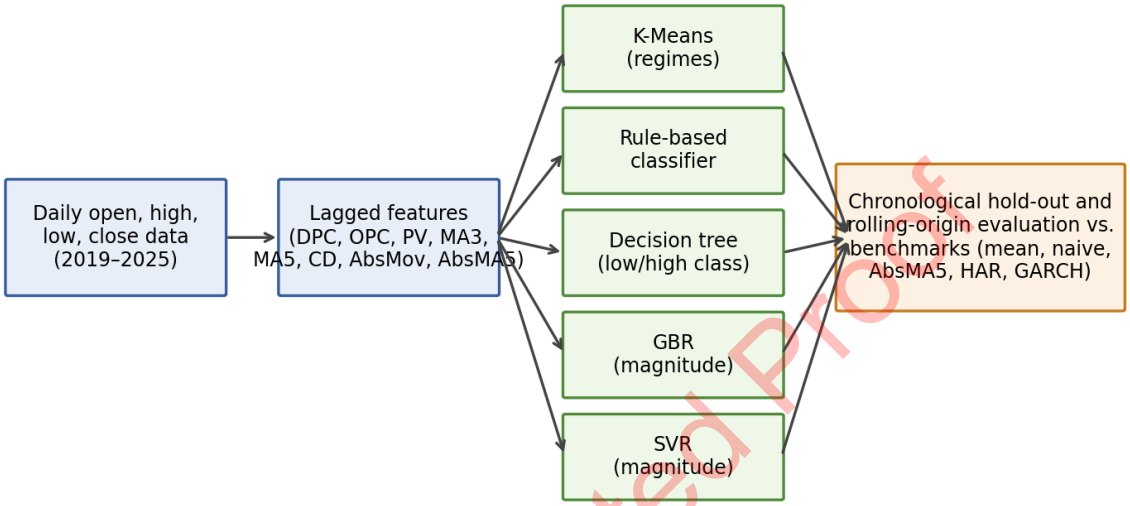


Figure 1. Proposed Prediction and Evaluation Framework
Source: Authors' calculations

Uncorrected Proof

3. Methodology

3.1. Research Design and Terminology

The study forecasts the next-day absolute percentage change of the TSE index over 25 March 2019 – 5 September 2025 (1,534 trading days, of which 1,512 remain after lags and rolling windows). Five models from three paradigms are estimated on the same lagged feature set and evaluated on a common chronological hold-out sample: (i) unsupervised regime detection (K-Means); (ii) classification of the next-day movement into a low and a high class (rule-based classifier and decision tree); and (iii) regression of the movement magnitude (GBR and SVR). The models are complementary rather than sequential, and the rule-based classifier is an interpretable benchmark and not a component of the regression models. K-Means is an unsupervised clustering algorithm and is used for regime description, not as a classifier. Five simple benchmarks and a rolling-origin evaluation are added (Section 3.7). The term index refers to the TSE index series downloaded from Rahavard365; the abbreviation TSE refers to the exchange.

3.2. Target Variable and Information Timing

Let O_t, H_t, L_t, and C_t denote the opening, highest, lowest, and closing price of trading day t. The daily percent change is $DPC_t = 100 \times (C_t - O_t) / O_t$. The forecasting target is the absolute value of the next day's daily percent change:

$$\text{AbsMov}_{t+1} = |\text{DPC}_{t+1}| = |100 \times (C_{t+1} - O_{t+1}) / O_{t+1}| \qquad (1)$$

A forecast is made after the close of day t and uses only information dated t or earlier (the feature vector X_t in Table 2). The target in row t+1 is aligned with the features of row t (a one-day lag). The forecast horizon is one trading day. Because the target is an open-to-close measure, the overnight change of day t+1 is not part of the target and is not used as a predictor of it; the overnight change of day t is known at the open of day t and is therefore admissible. Variables that contain contemporaneous information (DPC, PV, MA3, MA5, AbsMov) enter only through their values at day t or earlier. Table 2 lists each variable, its formula, the time at which it is available, and its role.

Table 2. Variables, Formulas and Information Timing

Variable	Formula	Information available at	Role in forecasting day t+1
Daily percent change DPC_t	$100\ (C_t - O_t) / O_t$	Close of day t	Predictor
Overnight percent change OPC_t	$100\ (O_t - C_{t-1}) / C_{t-1}$	Open of day t	Predictor
Percentage volatility PV_t	$100\ (H_t - L_t) / L_t$	Close of day t	Predictor
Moving averages MA3_t, MA5_t	Mean of DPC over days t−2…t and t−4…t	Close of day t	Predictor; rule-based input (MA3)
Direction CD_t	+1 if $C_t > O_t$, −1 otherwise	Close of day t	Predictor
Absolute change AbsMov_t	$\lvert\text{DPC}_t\rvert$	Close of day t	Predictor (lagged target)
AbsMA5_t	Mean of AbsMov over days t−4…t	Close of day t	Predictor; rule-based input
AbsMA22_t	Mean of AbsMov over days t−21…t	Close of day t	HAR benchmark only
Target AbsMov_{t+1}	$\lvert\text{DPC}_{t+1}\rvert$	Close of day t+1	Target (never a predictor)

Source: Authors' construction. t denotes a trading day; rolling windows count consecutive trading days. Price levels and volume are not predictors (Section 3.3).

Uncorrected Proof

3.3. Data and Preprocessing

Daily open, high, low, and close prices of the TSE total index (TEDPIX) for 25 March 2019 – 5 September 2025 (1,534 trading days) were obtained from Rahavard365 The original data file also contained a column labelled volume. Verification showed that it is not trading volume: it equals the closing price on 464 days and its correlation with the closing price is 0.998. Volume was therefore excluded from all analyses, and all models were re-estimated without it. The four price levels are also excluded from the predictors, because the index rose from about 0.18 to 3.25 million points and test-period levels lie outside the training range, which destabilizes tree-based and kernel models: in a specification that included the price levels (and the faulty volume column), the test R^2 of GBR was −6.11. On 200 days the open or close lies marginally outside the high–low range (by at most 0.008% of the close); we attribute this to rounding and leave the data unchanged. Non-trading days have no observations, so lags and rolling windows are computed over consecutive trading days.

Price-limit regime. The daily price band for listed stocks changed during the sample period. It was ±5% for years before 2021; from mid-February 2021 prices could fall by at most 2% and rise by at most 6% in a session (Financial Tribune, 2021); the symmetric ±5% band was restored in May 2021 (EghtesadOnline, 2021; Tehran Stock Exchange, 2021); and for First Market stocks it was widened to ±6% on 18 April 2022 (Tehran Stock Exchange, 2022a) and to ±7% in August 2022 (Tehran Stock Exchange, 2022b), while Second Market stocks remained at ±5%. Band changes after August 2022, if any, are not documented here. An index move aggregates many constituents, so the band does not bound the index directly, but it truncates

constituent moves and shapes the distribution of AbsMov; the extremes of about ±4% in Table 5 are therefore consistent with truncation at the stock level and not with an index-level limit. The main analysis does not model the regime explicitly, which is noted as a limitation in Section 5.3.

3.4. Feature Engineering

Features are defined mathematically below and consolidated, with their timing, in Table 2. With O, H, L, C as in Section 3.2:

$$DPC_t = 100 \times (C_t - O_t) / O_t \qquad (2)$$

$$OPC_t = 100 \times (O_t - C_{t-1}) / C_{t-1} \qquad (3)$$

$$PV_t = 100 \times (H_t - L_t) / L_t \qquad (4)$$

$$MA_{k,t} = (1/k) \Sigma_{j=0}^{k-1} DPC_{t-j}, \; k \in \{3, 5\} \qquad (5)$$

$$AbsMA5_t = (1/5) \Sigma_{j=0}^{4} AbsMov_{t-j}, \; AbsMov_t = |DPC_t| \qquad (6)$$

The direction variable CD_t equals +1 when the close exceeds the open and −1 otherwise.

3.5. Sample Split, Class Definition and Cross-Validation

After dropping the first 21 days (needed for the 22-day average used by the HAR benchmark) and the last day (which has no next-day target), 1,512 feature–target pairs remain. The sample is split chronologically: the first 80% (1,209 pairs; feature dates 2019-04-29 to 2024-05-11) form the training sample and the last 20% (303 pairs; feature dates 2024-05-12 to 2025-09-02, with targets through 2025-09-05) form the test sample. No shuffling is applied. Hyperparameters are tuned with five-fold expanding-window TimeSeriesSplit cross-validation inside the training sample only. For the classification tasks, the next-day class is high-movement (Class 1) if $AbsMov_{t+1}$ exceeds $\theta = 1$ percentage point and low-movement (Class 0) otherwise; θ is a fixed round value above the sample mean of AbsMov (0.83%) and is not tuned. Class counts are reported in Table 6. Precision, recall, and F1-score refer to the positive (high-movement) class.

3.6. Models and Implementation

Rule-based classifier. Three deterministic rules based on lagged moving averages (Rule 1: MA3 ≤ 0.3 on days t and t−1; Rule 2: AbsMA5 ≤ 0.516 on days t and t−1; Rule 3: AbsMA5 ≥ 1 on days t and t−1). Rules 1 and 2 signal low movement and Rule 3 signals high movement. The three thresholds are fixed values carried over from the initial analysis and are not re-estimated because they were fixed in the initial analysis, they may have been informed by the full sample, so the rule results are treated as descriptive and not as strictly out-of-sample evidence; Table 8 reports the share of training days satisfying each condition (65.5%, 28.5%, and 32.7%) and the sensitivity of the test accuracy to scaling all thresholds by 0.6 to 1.4.

Decision tree. A decision tree classifier partitions the feature space through recursive binary splits; its depth is tuned using TimeSeriesSplit cross-validation within the training sample (Table 3).

K-Means clustering. K-Means partitions the standardized features (training-sample means and standard deviations) into k clusters by minimizing within-cluster variance. The number of clusters is k = 3 because it maximizes the silhouette score in the training sample among k = 2 to 6 (k = 2: 0.320, k = 3: 0.354, k = 4: 0.319, k = 5: 0.296, k = 6: 0.302) (Rousseeuw, 1987). The silhouette values are low and differ little across k (0.354 at k = 3 against 0.320 at k = 2), which indicates weak cluster structure; the choice of k = 3 is therefore descriptive. Because K-Means is unsupervised, it has no accuracy in the predictive sense. We report (i) the silhouette score, (ii) the agreement between clusters and movement classes, obtained by

Uncorrected Proof

mapping each cluster to the majority class of the training sample and applying this fixed mapping to the test sample, which is a purity-type measure and not a forecasting accuracy, and (iii) the adjusted Rand index (Hubert & Arabie, 1985) and normalized mutual information.

Gradient boosting regression (GBR). An additive ensemble of shallow regression trees fitted sequentially to the residuals of the previous stage (Friedman, 2001); the learning rate is tuned by grid search with temporal folds (Table 3).

Support vector regression (SVR). A kernel regression with an epsilon-insensitive loss that captures nonlinear dependence; the kernel, C, and epsilon are tuned using temporal cross-validation (Table 3).

Table 3. Tuned Hyperparameters, Search Grids and Selected Values

Model	Search grid	Selection criterion	Selected values
Rule-based	Thresholds 0.3, 0.516, 1 (fixed)	–	0.3; 0.516; 1
Decision tree	Maximum depth {2, 3, 4, 5, 6, 8}; minimum leaf size {5, 10, 20}	F1-score in time-series CV	Depth 8; leaf size 10
K-Means	k = 2 to 6; 20 initializations; seed 42	Silhouette score (training)	k = 3
GBR	Learning rate {0.03, 0.05, 0.1}; trees {100, 300}; depth {2, 3}	MSE in time-series CV	Learning rate 0.03; 100 trees; depth 2
SVR	RBF kernel; C {0.1, 1, 10}; epsilon {0.01, 0.05, 0.1}; standardized inputs	MSE in time-series CV	C = 1; epsilon = 0.1

Source: Authors' implementation. Random seed 42 throughout. Several selected values lie at the edge of the grid (tree depth 8, GBR learning rate 0.03, 100 trees and depth 2, SVR epsilon 0.1); see Section 5.3.

3.7. Benchmarks and Forecast Comparison

To test whether the machine-learning models add information beyond simple approaches, five benchmarks forecast $AbsMov_{t+1}$: (B0) the training-sample mean; (B1) a naive forecast equal to $AbsMov_t$; (B2) the five-day average $AbsMA5_t$; (B3) a HAR-type linear regression of $AbsMov_{t+1}$ on $AbsMov_t$, $AbsMA5_t$, and the 22-day average $AbsMA22_t$, estimated by OLS on the training sample (Corsi, 2009), which uses AbsMA22 as an input that the machine-learning models do not receive, so that the comparison with HAR is not like-for-like (Section 5.3); and (B4) a GARCH(1,1) model for DPC with constant mean, estimated by Gaussian quasi-maximum likelihood on the training sample (Bollerslev, 1986), whose one-step-ahead conditional standard deviation is converted into an expected absolute change under normality (multiplication by the square root of $2/\pi$). Forecast accuracy is compared on the test sample using RMSE and the out-of-sample R^2 relative to the historical mean, $R^2OS = 1 - SSEmodel / SSEmean$, where the benchmark forecast is the training-sample mean (B0), which is available at forecast time. The conventional R^2 computed against the test-sample mean is reported only as a supplementary column in Table 12. Differences in squared-error loss are tested with the Diebold–Mariano test at horizon one with the small-sample correction of Harvey et al. (1997) (Diebold & Mariano, 1995).

Because a single test block can be unrepresentative, we also run a rolling-origin evaluation. The first 756 feature–target pairs (50%) form the initial training window; the remaining 756 pairs are split into five consecutive blocks of 151 or 152 observations each, since 756/5 is not an integer, and each block is forecast by models fitted on all earlier observations; HAR and GARCH are re-estimated at the start of each block on all earlier observations. Hyperparameters are fixed at the values selected in the main analysis. Pooled forecasts are compared by RMSE and by the out-of-sample R^2 relative to the forecast of the mean of all earlier observations (the same R^2OS as in the test block, with the expanding historical mean as the benchmark), with Diebold–Mariano tests on the pooled forecasts. As a further check, the decision tree is re-estimated for alternative class thresholds θ.

Uncorrected Proof

3.8. Evaluation and Data Leakage Control

Table 4 summarizes the evaluation protocol. All test results are computed on the final 20% of observations, which are never used for fitting or hyperparameter selection. Preprocessing and model tuning were designed so that no test-period information enters model development. (i) All predictors are lagged transformations of past observations, and rolling averages use only current and earlier days. (ii) Hyperparameters were tuned with expanding-window TimeSeriesSplit cross-validation inside the training sample only, and the test sample was used once, for the final evaluation. (iii) Standardization statistics, the K-Means centroids, the cluster-to-class mapping, and the HAR coefficients were estimated on the training sample only and applied unchanged to the test sample. (iv) The class threshold θ is a fixed round value, and the rule thresholds were fixed before the present evaluation (their origin is described in Section 3.6; their sensitivity is reported in Table 8). These controls refer to the main chronological split; the rolling-origin evaluation is subject to the limitation stated in Section 5.3.

Table 4. Evaluation Framework and Data Leakage Prevention Protocol

Component	Specification	Implementation Details
Regression Evaluation Metrics	Model performance for the continuous target AbsMov(t+1)	R^2, MSE and RMSE computed on the hold-out test set
Classification Evaluation Metrics	Low- versus high-movement class of day t+1 Volatility direction/regime prediction	Accuracy, Precision, Recall, F1-score computed on hold-out test set
Cross-Validation Strategy	Stability and generalization assessment	TimeSeriesSplit cross-validation applied on training set only
Validation Purpose	Model stability check	Cross-validation results reported separately from test results
Train-Test Protocol	Temporal generalization framework	Chronological split (first 80% train, last 20% test)
Predictor Construction Rule	Information timing Non-stationarity handling	All predictors are computed from data up to the close of day t and aligned with the target of day t+1 (one-day lag)
Data Leakage Prevention	No-lookahead bias elimination	No future data used in features, training, or validation
Hyperparameter Optimization	Model tuning strategy	GridSearchCV performed only within training folds using TimeSeriesSplit
Methodological Integrity Condition	Temporal ordering condition Forecasting validity guarantee	Feature timestamps precede target timestamps for every observation

Source: Authors' construction.

Uncorrected Proof

3.9. Model Development Workflow

Figure 2 shows the workflow: data collection and cleaning, lagged feature construction, a chronological split, tuning inside the training sample, and a single test evaluation against the benchmarks.

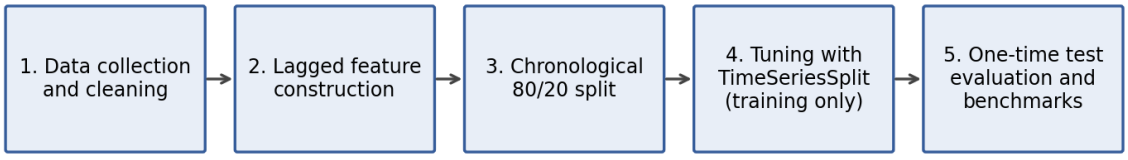


Figure 2. Overview of the Model Development and Evaluation Process
Source: Authors' calculations

3.10. Reproducibility

All analyses were run in Python (pandas 3.0, NumPy 2.4, SciPy 1.17, scikit-learn 1.8; Pedregosa et al., 2011) with random seed 42. The complete analysis script and the cleaned data file used for all reported results (without the invalid volume column) are provided as supplementary material to this submission.

4. Results

This section reports descriptive evidence and the out-of-sample performance of the five models and the five benchmarks. Variables are defined in Section 3.2 (Table 2) and abbreviations in Table 17. Performance metrics are computed on the hold-out sample unless stated otherwise.

4.1. Data Overview and Descriptive Statistics

The index rose from about 0.18 to about 3.25 million points over the sample (Figure 3). Because price levels trend strongly, the percentage-based movement measures in Table 5 are the relevant objects for forecasting. The daily percent change has a mean of 0.06% and a standard deviation of 1.11%, with extremes of -4.24% and 4.10%. The absolute change has a mean of 0.83% and a median of 0.61%; the gap between them and a 75th percentile of 1.18% indicate a right-skewed distribution in which a minority of large-move days dominates squared-error metrics. Extremes near ±4% are consistent with the effect of daily price limits. About 50.8% of days close above their open. Mean absolute movement is lower in the test sample (0.67%) than in the training sample (0.86%), which matters for the interpretation of the out-of-sample R^2 below.

Uncorrected Proof

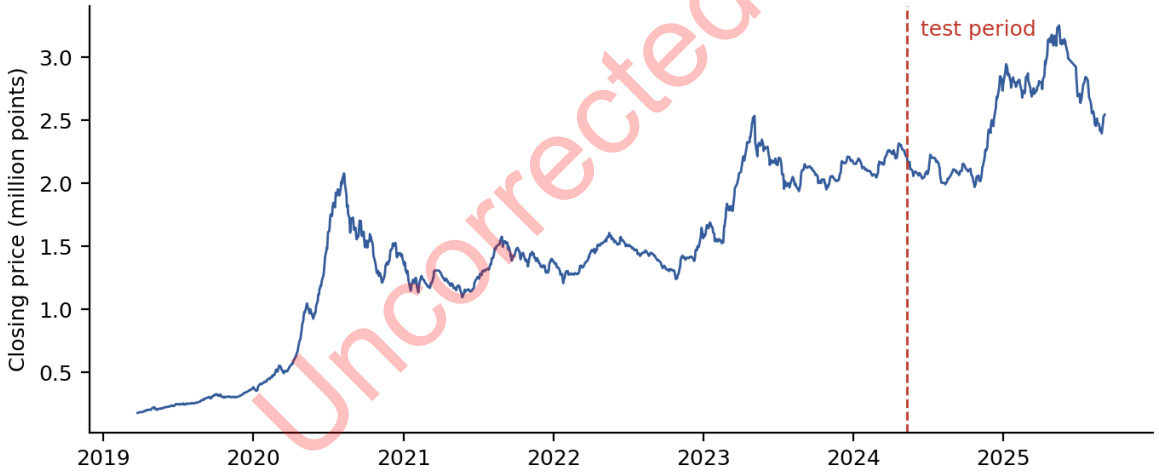


Figure 3. Closing price of the index and the start of the test period
Source: Authors' calculations

Table 5. Descriptive Statistics of the Index Prices and Engineered Variables

Variable	N	Mean	SD	Min	25%	Median	75%	Max
Opening price	1,534	1,575,066	745,398	178,788	1,238,872	1,506,597	2,110,406	3,252,839
Highest price	1,534	1,582,350	747,763	179,194	1,245,734	1,514,757	2,114,228	3,259,432
Lowest price	1,534	1,568,228	742,460	178,743	1,233,598	1,502,439	2,103,094	3,244,670
Closing price	1,534	1,575,210	744,722	179,194	1,238,390	1,509,287	2,110,463	3,252,219

Daily percent change (DPC)	1,534	0.060	1.108	-4.241	-0.553	0.015	0.666	4.100
Overnight percent change (OPC)	1,533	0.121	0.516	-3.294	-0.033	0.042	0.259	3.078
Percentage volatility (PV)	1,534	0.958	0.732	0.037	0.403	0.757	1.331	4.428
MA3	1,532	0.060	0.769	-2.824	-0.404	0.015	0.514	2.710
MA5	1,530	0.058	0.654	-2.351	-0.344	0.023	0.461	2.468
Direction (CD)	1,534	0.017	1.000	-1.000	-1.000	1.000	1.000	1.000
Absolute change (AbsMov)	1,534	0.825	0.741	0.000	0.259	0.610	1.183	4.241
AbsMA5	1,530	0.826	0.453	0.057	0.482	0.749	1.034	2.602

Source: Authors' calculations from the 1,534-day data file. DPC, OPC, PV, MA3, MA5 and AbsMA5 are percentages. Direction is coded +1 or −1.

Table 6. Chronological Sample Split and Class Distribution (θ = 1 percentage point)

Sample	Feature dates	Pairs	Low-movement (Class 0), n (%)	High-movement (Class 1), n (%)
Training	2019-04-29 to 2024-05-11	1,209	822 (68.0%)	387 (32.0%)
Test	2024-05-12 to 2025-09-02	303	229 (75.6%)	74 (24.4%)
Full sample	2019-04-29 to 2025-09-02	1,512	1,051 (69.5%)	461 (30.5%)

Source: Authors' calculations. Class membership refers to the next-day absolute change.

4.2. Daily Changes, Overnight Gaps and Intraday Range

Figures 4 and 5 present the daily, overnight, and intraday-range series. The persistence of large absolute moves, a stylized fact of financial series (Andersen et al., 2001), is visible in the autocorrelations of AbsMov, which are 0.223 at lag 1 and between 0.198 and 0.229 at lags 2 to 5. The intraday range PV is almost mechanically tied to the absolute daily change: the open-to-close move cannot exceed the high–low range, and the two series have a correlation of 0.95. A regression of AbsMov on the same-day PV therefore explains 90.7% of the variance in the training sample and 92.8% in the test sample. Such same-day information is unavailable when forecasting, and the forecasting design uses the one-day lag of PV. The check also shows how much a design that mixes same-day information into the predictors inflates apparent accuracy.

Uncorrected Proof

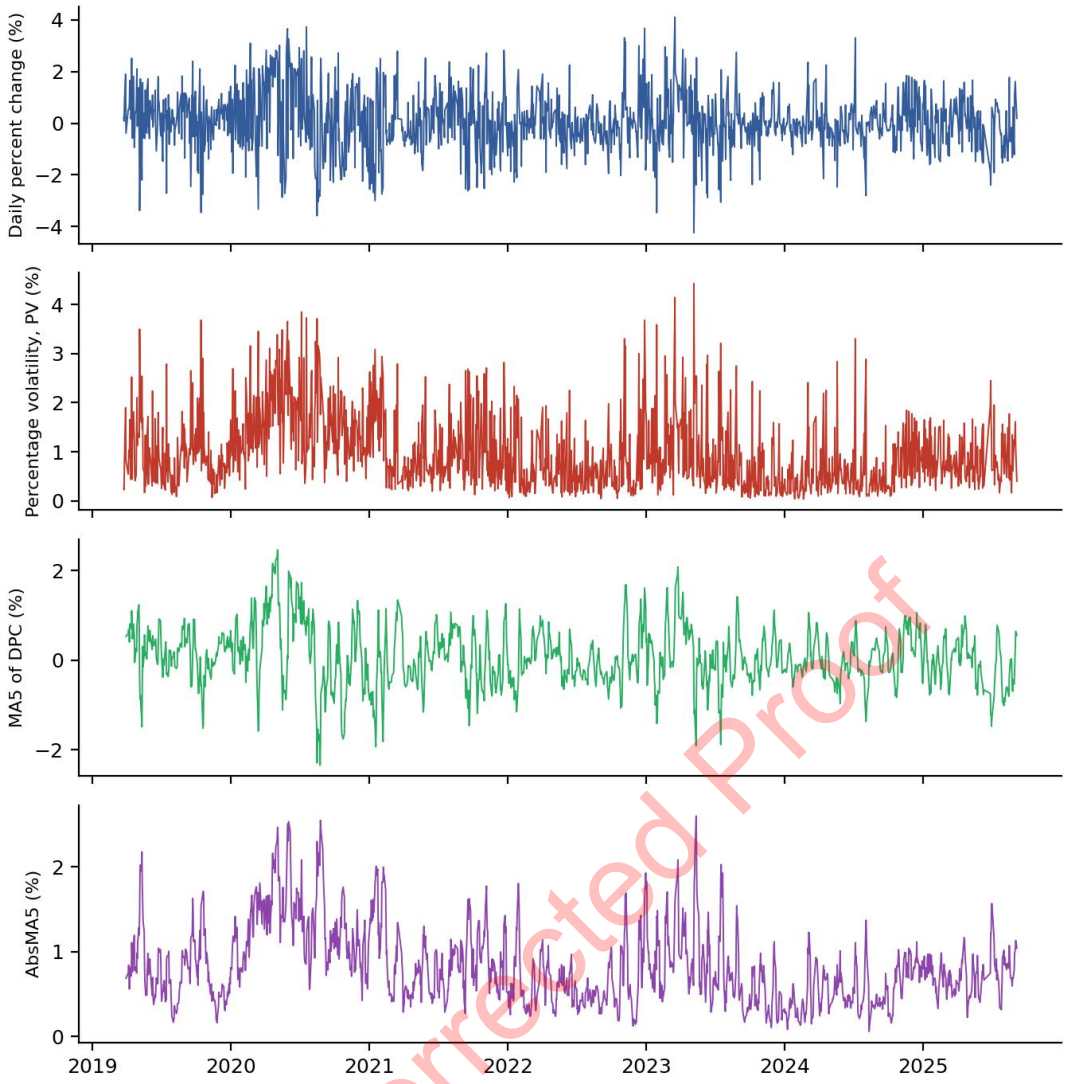


Figure 4. Daily percent change, percentage volatility, MA5 and AbsMA5
Source: Authors' calculations

Uncorrected Proof

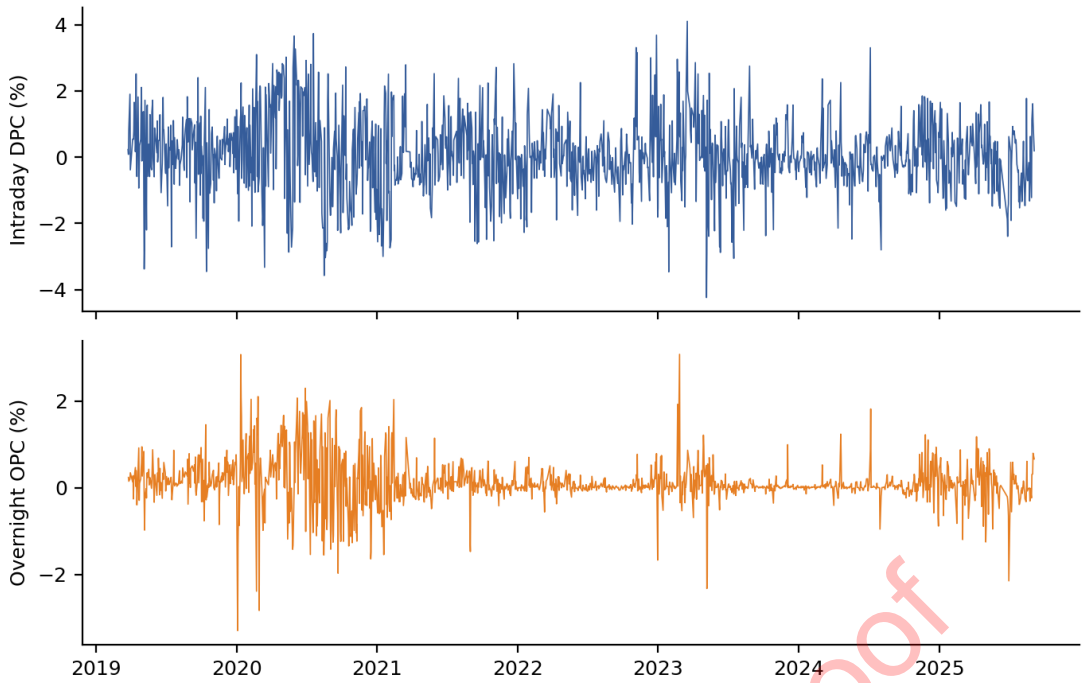


Figure 5. Intraday (DPC) and overnight (OPC) percentage changes
Source: Authors' calculations

4.3. Rule-Based Classifier

Table 7 reports the three rules on the training, test, and full samples, together with the base rate of the predicted class. Rules are informative only if their accuracy exceeds that base rate. Rule 1 (MA3 ≤ 0.3 on days t and t−1) is correct in 75.0% of its test cases against a base rate of 75.6%, so it adds nothing out of sample. Rule 2 (AbsMA5 ≤ 0.516) is correct in 84.8% of its 79 test cases against a base rate of 75.6%, and in 86.4% of its training cases against 68.0%: a low recent average of absolute movements does signal a calm next day. Rule 3 (AbsMA5 ≥ 1) is correct in 52.6% of its training cases against a base rate of 32.0%, but it fires only 10 times in the test sample (30.0% correct against 24.4%), so no conclusion can be drawn for the test period, which was calmer than the training period. The rules are deterministic benchmarks and are not inputs to the regression models. Table 8 reports the position of each threshold in the training distribution and the sensitivity of the test accuracy.

Uncorrected Proof

Table 7. Rule-Based Classifier: Coverage, Accuracy and Base Rate by Sample

Rule	Sample	Days flagged	Coverage (%)	Accuracy (%)	Base rate of predicted class (%)
Rule 1	Train	683	56.5	71.4	68.0
	Test	176	58.1	75.0	75.6
	Full	859	56.8	72.2	69.5
Rule 2	Train	273	22.6	86.4	68.0
	Test	79	26.1	84.8	75.6
	Full	352	23.3	86.1	69.5
Rule 3	Train	331	27.4	52.6	32.0
	Test	10	3.3	30.0	24.4

Full	341	22.6	51.9	30.5

Source: Authors' calculations. Rules 1 and 2 predict the low-movement class and Rule 3 the high-movement class.

Table 8. Rule Thresholds: Position in the Training Distribution and Test Accuracy under Scaled Thresholds

Rule	Training days satisfying rule (%)	×0.6	×0.8	×1.0	×1.2	×1.4
Rule 1: MA3 ≤ 0.3	65.5	74.1 (162)	74.4 (172)	75.0 (176)	74.5 (188)	75.0 (200)
Rule 2: AbsMA5 ≤ 0.516	28.5	93.8 (16)	92.7 (41)	84.8 (79)	81.8 (110)	79.2 (144)
Rule 3: AbsMA5 ≥ 1	32.7	29.6 (162)	23.6 (72)	30.0 (10)	40.0 (5)	0.0 (1)

Source: Authors' calculations. Cells show test accuracy in percent and, in parentheses, the number of test days flagged. The base rates are 75.6% (Rules 1 and 2) and 24.4% (Rule 3).

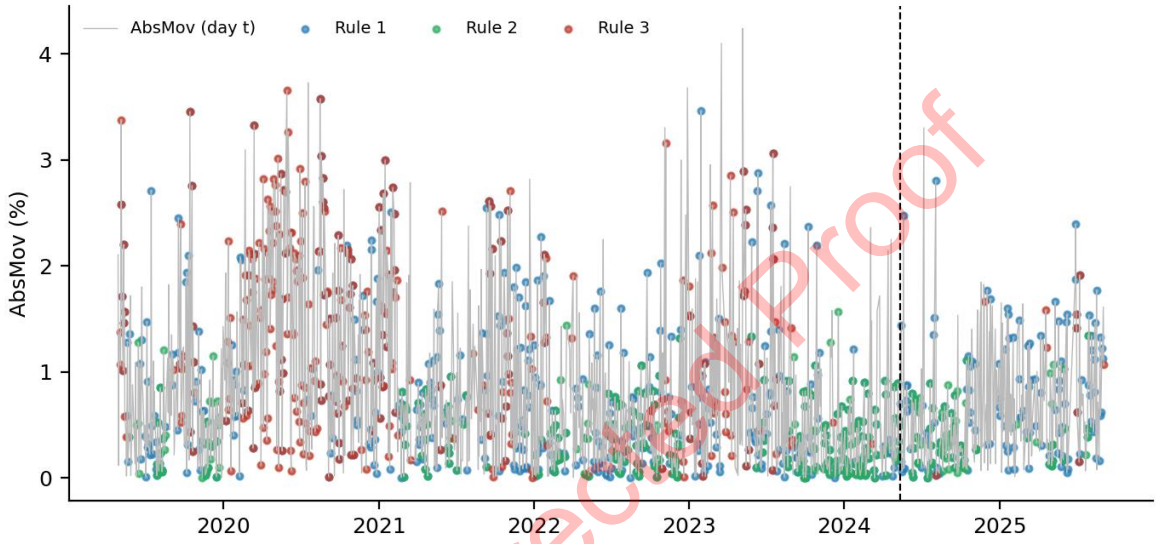


Figure 6. Days flagged by each rule (dashed line: start of the test period)
Source: Authors' calculations

Uncorrected Proof

4.4. Feature Distributions Across Clusters

Figure 7 compares the distributions of the daily percent change, MA3, MA5, and PV across the three K-Means clusters (construction in Section 3.6, evaluation in Section 4.7). The clusters mainly separate the character of the current day: Cluster 1 groups days with small ranges and near-zero changes, Cluster 2 groups days with positive changes, and Cluster 3 groups days with negative changes and larger ranges.

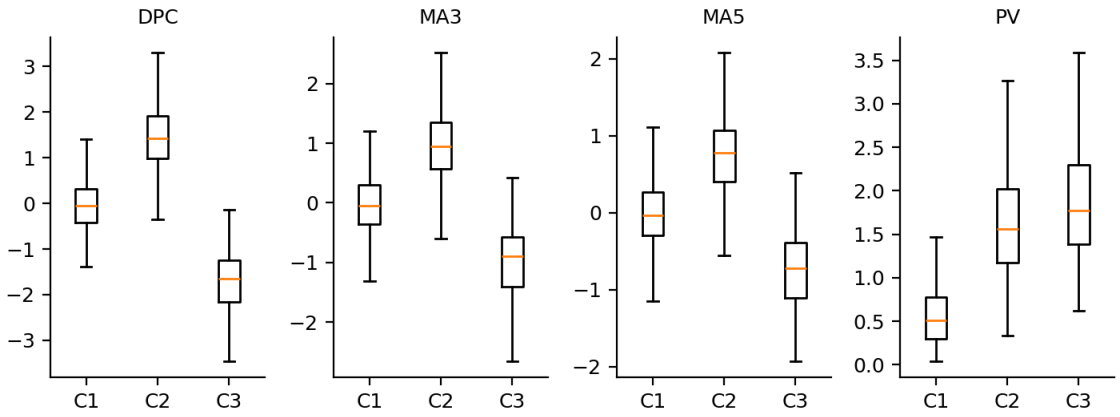


Figure 7. Distribution of selected features across the K-Means clusters
Source: Authors' calculations

4.5. Relationships Among Features

Figure 8 shows the correlations among the features and with next-day AbsMov. The strongest correlates of the target are AbsMA5 (0.35), PV (0.28), and AbsMov (0.22); the signed variables DPC, MA3, MA5, and the direction indicator are essentially uncorrelated with it ($|r| \leq 0.04$). Absolute movement and PV are strongly related, as expected from the mechanical link described in Section 4.2. The information in the features is thus concentrated in recent absolute movements, which is what the HAR benchmark uses.

Uncorrected Proof

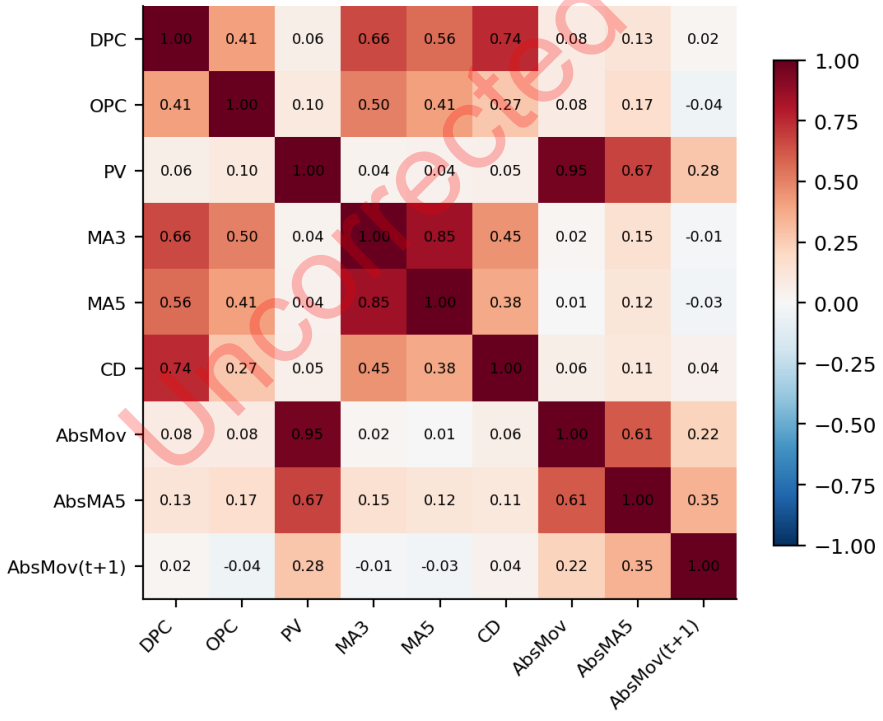


Figure 8. Correlation matrix of the features and the next-day target
Source: Authors' calculations

4.6. Decision Tree Classification

Table 9 reports the decision tree on the training and test samples. In the test sample, accuracy is 69.0%, precision 29.2%, recall 18.9%, and the F1-score 23.0%. Accuracy is below the 75.6% that a classifier always predicting the low-movement class would obtain, and recall shows that only 14 of the 74 high-movement days were detected. The training F1-score of 59.0% against the test F1-score of 23.0% shows that the tree overfits. Table 10 gives the confusion matrix.

Table 9. Decision Tree Performance (%)

Sample	Accuracy	Precision	Recall	F1-score
Training	78.16	73.93	49.10	59.01
Test	68.98	29.17	18.92	22.95
Test: always predict low movement	75.58	–	0.00	–

Source: Authors' calculations. Precision, recall and F1-score refer to the high-movement class.

Table 10. Decision Tree Confusion Matrix on the Test Sample

Actual \ Predicted	Low movement	High movement	Total
Low movement	195	34	229
High movement	60	14	74
Total	255	48	303

Source: Authors' calculations.

Figure 9 shows the top three levels of the fitted tree; its first splits use AbsMA5, PV, MA3, and OPC, i.e., recent absolute movements and the range.

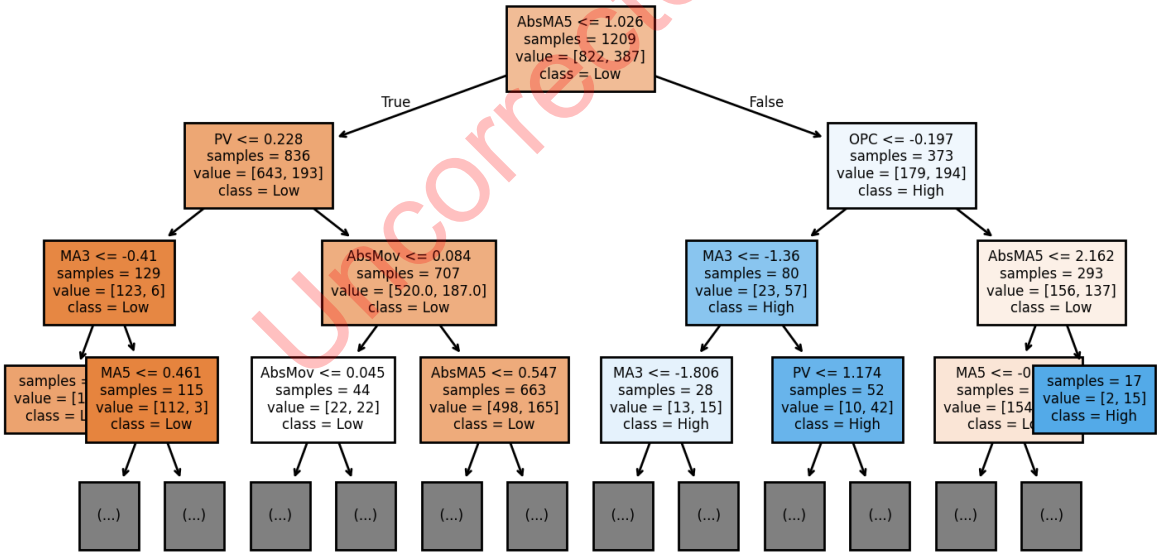


Figure 9. Decision tree (top three levels)
Source: Authors' calculations

Uncorrected Proof

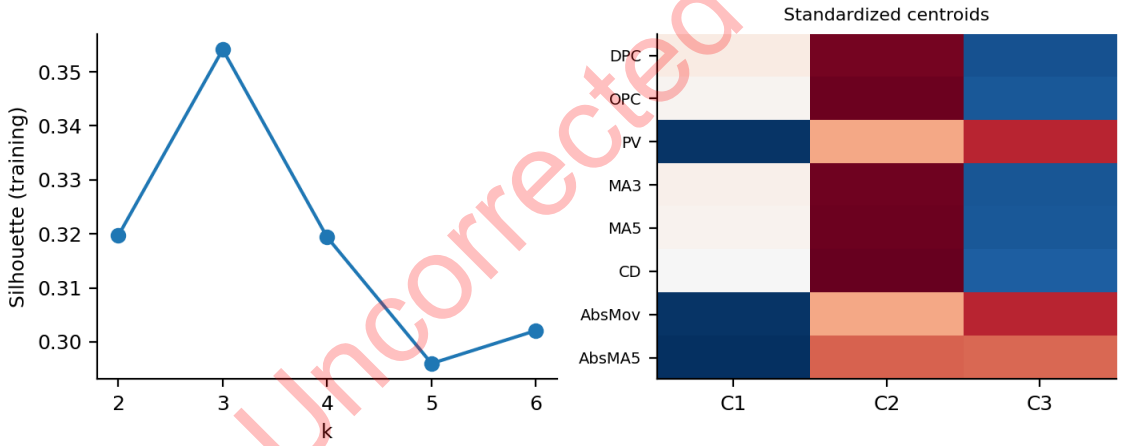
4.7. K-Means Market Regimes

Table 11 reports the silhouette scores for k = 2 to 6, which are highest at k = 3 (0.354) but low overall, indicating weak cluster structure, and the evaluation of the three-cluster solution. The cluster–class agreement on the test sample is 75.6%, which equals the share of the majority class (75.6%); the adjusted Rand index is −0.007 and the normalized mutual information is 0.001. In the training sample, mean next-day AbsMov differs across clusters (0.73, 1.03, and 1.22), but this structure does not persist out of sample (0.68, 0.60, and 0.70; Kruskal–Wallis p = 0.56). The regimes describe the current day and carry no information on the size of the next move.

Table 11. K-Means Clustering: Silhouette Scores and Out-of-Sample Evaluation

Item	Value
Silhouette (training), k = 2 / 3 / 4 / 5 / 6	0.320 / 0.354 / 0.319 / 0.296 / 0.302
Silhouette (test), k = 3	0.252
Cluster–class agreement (test)	75.6%
Majority-class share (test)	75.6%
Adjusted Rand index (test)	−0.007
Normalized mutual information (test)	0.001
Mean next-day AbsMov by cluster, training (C1 / C2 / C3)	0.73 / 1.03 / 1.22
Mean next-day AbsMov by cluster, test (C1 / C2 / C3)	0.68 / 0.60 / 0.70
Kruskal–Wallis test across clusters (test), p-value	0.56

Source: Authors' calculations.



Uncorrected Proof

Figure 10. K-Means: silhouette scores and standardized cluster centroids
Source: Authors' calculations

4.8. Regression Models: Gradient Boosting and SVR

Table 12 reports the regressions and benchmarks. In the training sample the R^2 is 0.206 for GBR and 0.230 for SVR. In the test block, the out-of-sample R^2 relative to the training-sample mean (R^2OS) is 6.4% for GBR and −0.7% for SVR (RMSE 0.554 and 0.574), against 10.7% for the HAR-type benchmark (RMSE 0.541), the best of all forecasts, and 7.8% for GARCH(1,1). SVR therefore does not improve on the historical-mean forecast in this block. The conventional R^2 computed against the test-sample mean, which is not available at forecast time, is −0.054, −0.134, −0.006, and −0.038 for GBR, SVR, HAR, and GARCH(1,1); on that basis even the training-sample mean has an R^2 of −0.126, which shows how much

calmer the test period is than the training period. GBR is not statistically different from HAR (Table 13), SVR is significantly worse, and GBR and SVR are not distinguishable from each other. HAR is also significantly better than GARCH (p = 0.013). The naive forecast (R^2OS of −72.8%) is far worse than the training mean, and the five-day average (−0.8%) is no better. Figures 11 and 12 show that both models track the general level of absolute movements but miss the large days.

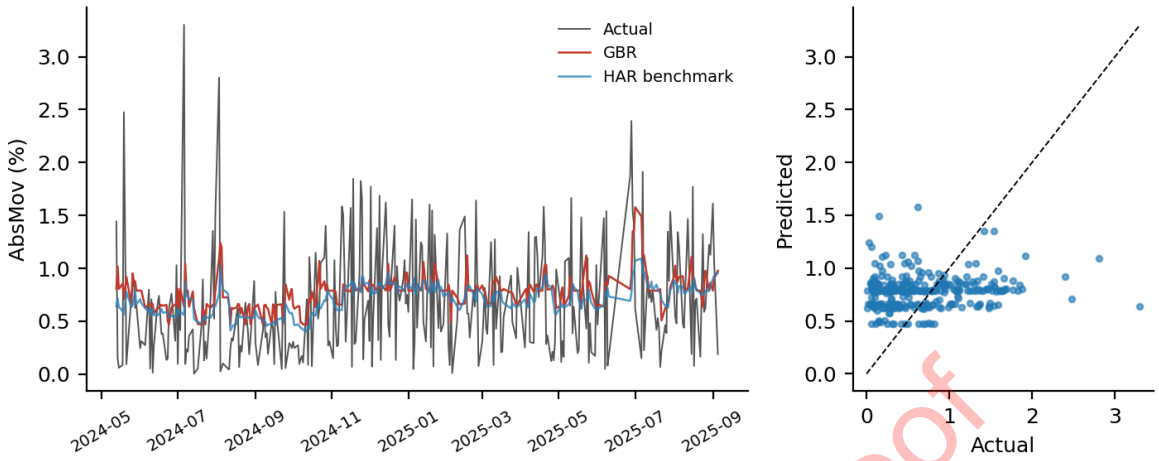
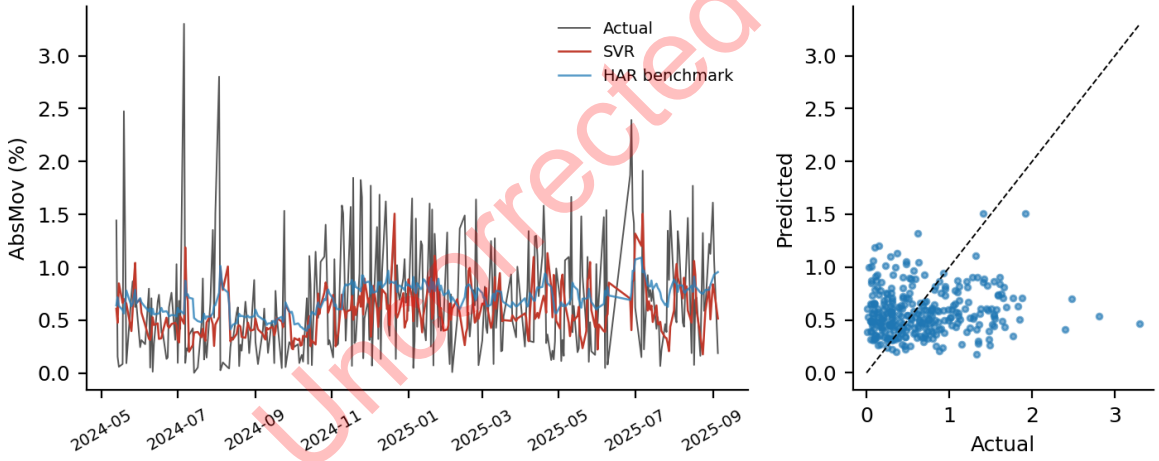


Figure 11. GBR forecasts versus actual next-day AbsMov (test sample)
Source: Authors' calculations



Uncorrected Proof

Figure 12. SVR forecasts versus actual next-day AbsMov (test sample)
Source: Authors' calculations

4.9. Comparative Performance and Forecast Comparison

Table 12 puts all regression forecasts side by side, and Table 13 tests the differences. HAR is significantly better than the training mean (p = 0.004), which confirms that recent absolute movements carry information. Neither machine-learning model improves on HAR. In the final test block, HAR, GARCH, and GBR improve on the historical-mean forecast (R^2OS of 10.7%, 7.8%, and 6.4%), whereas SVR does not (−0.7%). The decision tree is less accurate than the majority-class benchmark (Table 9), and the K-Means regimes have no out-of-sample association with the target (Table 11). Section 4.11 shows that the conclusion about

machine learning holds in a rolling-origin evaluation, where all models beat the expanding historical mean (R^2OS of 6.8% to 13.8%).

Table 12. Regression Models and Benchmarks: Forecast Accuracy for Next-Day AbsMov

Model	Sample	MSE	RMSE	R^2 vs test-sample mean (supplementary)	R^2OS vs training mean (%)
GBR	Training	0.4819	0.694	0.206	–
SVR	Training	0.4676	0.684	0.230	–
GBR	Test	0.3065	0.554	−0.054	6.4
SVR	Test	0.3297	0.574	−0.134	−0.7
B0 Training-sample mean	Test	0.3274	0.572	−0.126	0.0
B1 Naive ($AbsMov_t$)	Test	0.5658	0.752	−0.946	−72.8
B2 $AbsMA5_t$	Test	0.3301	0.575	−0.135	−0.8
B3 HAR-type OLS	Test	0.2924	0.541	−0.006	10.7
B4 GARCH(1,1)	Test	0.3020	0.550	−0.038	7.8

Source: Authors' calculations. $R^2OS = 1 - MSEmodel / MSE(B0)$, with the training-sample mean (B0) as the benchmark forecast, which is available at forecast time; the R^2 against the test-sample mean is not available at forecast time and is shown for reference.

Table 13. Diebold–Mariano Tests of Equal Forecast Accuracy (Squared-Error Loss, Horizon 1, Test Sample)

Comparison	DM statistic	p-value	Lower squared error
GBR vs. SVR	−1.281	0.201	GBR
GBR vs. HAR-type OLS	1.562	0.119	HAR-type OLS
SVR vs. HAR-type OLS	2.645	0.009	HAR-type OLS
HAR-type OLS vs. Training-sample mean	−2.869	0.004	HAR-type OLS
B4 GARCH(1,1) vs. HAR-type OLS	2.495	0.013	HAR-type OLS
GBR vs. B4 GARCH(1,1)	0.501	0.617	B4 GARCH(1,1)

Source: Authors' calculations. Harvey et al. (1997) small-sample correction applied. A negative statistic means that the first model has the lower loss.

Uncorrected Proof

4.10. Feature Importance

Table 14 reports permutation importance on the test sample, measured as the mean decrease in R^2 when a predictor is shuffled (Breiman, 2001). For GBR, the most important predictors are PV (0.035) and AbsMov (0.018); the signed variables contribute little or nothing, and DPC, MA3, and MA5 have slightly negative values. For SVR, PV (0.173), MA5 (0.111), and AbsMov (0.085) rank highest, followed by OPC (0.041), while shuffling DPC and CD does not reduce accuracy (−0.036 and −0.030). The importance of MA5 for SVR is not visible in the linear correlations of Section 4.5 ($|r| \leq 0.04$). Because permutation importance is sensitive to correlated inputs (MA5 is correlated with MA3 and DPC) and to the kernel's scaling, these SVR values should be read with caution. The importance values are small in absolute terms for GBR (at most 0.035), consistent with its weak forecasting performance.

Table 14. Permutation Importance of Lagged Predictors on the Test Sample

Predictor (value at day t)	GBR: mean decrease in R^2 (SD)	SVR: mean decrease in R^2 (SD)
DPC	−0.001 (0.001)	−0.036 (0.017)

OPC	0.008 (0.008)	0.041 (0.040)
PV	0.035 (0.012)	0.173 (0.038)
MA3	−0.003 (0.006)	0.007 (0.018)
MA5	−0.007 (0.010)	0.111 (0.047)
CD	0.000 (0.000)	−0.030 (0.017)
AbsMov	0.018 (0.005)	0.085 (0.037)
AbsMA5	0.010 (0.022)	0.030 (0.022)

Source: Authors' calculations. 30 repetitions.

4.11. Robustness: Rolling-Origin Evaluation and Alternative Class Thresholds

Table 15 reports the rolling-origin evaluation (756 pooled out-of-sample forecasts, five expanding-window blocks). Unlike the single test block, which was calmer than the training period, the rolling evaluation covers several volatility regimes, and all models beat the forecast of the historical mean: the out-of-sample R^2 relative to the expanding historical mean (R^2OS) is 13.8% for HAR, 12.6% for GARCH, 11.2% for GBR, and 6.8% for SVR. HAR has the lowest RMSE, and the Diebold–Mariano tests show that GBR ($p = 0.029$) and SVR ($p = 0.006$) forecast significantly worse than HAR, while HAR is significantly better than the historical mean ($p < 0.001$). The decision tree reaches an accuracy of 73.4% with a recall of 18.8% and an F1-score of 24.2%, below the 77.5% accuracy of the majority-class forecast. The conclusion is therefore the same as in the single split: recent absolute movements carry a persistent signal, but the machine-learning models do not add to it.

Table 15. Rolling-Origin Evaluation: Pooled Out-of-Sample Forecast Accuracy

Model	RMSE	Out-of-sample R^2 vs. historical mean (%)	DM p-value vs. HAR
Historical mean (B0)	0.700	0.0	–
HAR-type OLS (B3)	0.650	13.8	–
GARCH(1,1) (B4)	0.655	12.6	–
GBR	0.660	11.2	0.029
SVR	0.676	6.8	0.006

Source: Authors' calculations. 756 forecasts; first training window of 756 pairs; hyperparameters fixed at the values of Table 3. The DM test for GARCH against HAR was not computed in this evaluation.

Table 16 re-estimates the decision tree for alternative class thresholds. At no threshold does the tree beat the accuracy of always predicting the majority class; at $\theta = 0.5$, where the classes are balanced, its accuracy is 48.8% against a baseline of 51.8%.

Table 16. Decision Tree Test Performance under Alternative Class Thresholds (%)

θ (percentage points)	High-movement share (test)	Accuracy	Majority-class accuracy	Recall	F1-score
0.5	51.8	48.8	51.8	56.7	53.5
0.75	36.3	59.4	63.7	19.1	25.5
1.0 (baseline)	24.4	69.0	75.6	18.9	23.0
1.25	16.2	81.2	83.8	10.2	14.9

Source: Authors' calculations. The tree is re-tuned on the training sample for each threshold using the grid of Table 3.

Uncorrected Proof

5. Discussion and Conclusion

5.1. Interpretation of Findings

The main finding is that the machine-learning models, given the eight lagged predictors of Table 2, add no forecasting value beyond a simple HAR-type regression that also uses a 22-day average (RQ1). In the final test block, R^2OS relative to the training-sample mean is 6.4% for GBR, −0.7% for SVR, 7.8% for GARCH(1,1), and 10.7% for HAR. In the rolling-origin evaluation, all models beat the expanding historical mean (R^2OS of 11.2% for GBR and 6.8% for SVR) but HAR does better (13.8%), and the difference is statistically significant ($p = 0.029$ for GBR, $p = 0.006$ for SVR). GBR does not forecast significantly better than SVR in the test block (RQ2; DM $p = 0.20$). Rules, clusters, and the decision tree separate low- from high-movement days no better than the majority-class benchmark out of sample (RQ3), although the rule on a low five-day average shows some lift (84.8% against 75.6%).

Recent absolute movements do carry information. The autocorrelations of AbsMov are about 0.2 up to lag 5; AbsMA5, PV, and AbsMov are the features most correlated with the target; HAR beats the historical mean significantly in both evaluations; and Rule 2 shows lift in both samples. This is the persistence documented in the volatility literature (Andersen et al., 2001). The R^2 values are nevertheless modest, for three reasons. First, daily absolute open-to-close changes are a noisy proxy for latent volatility, and measurement error in the proxy mechanically limits the R^2 of even a well-specified model (Andersen & Bollerslev, 1998). Second, the test block is calmer than the training period (mean AbsMov 0.67% against 0.86%): against the test-sample mean, which is not available at forecast time, every R^2 is negative, even that of the training-mean forecast (−0.126), whereas R^2OS against the training mean is positive for HAR, GARCH, and GBR. The rolling-origin evaluation, which covers several regimes, gives positive R^2OS for all models. Third, price limits truncate extreme moves, and the models lack information on macroeconomic, political, and geopolitical events.

Flexible models do not help in this setting. With about 1,200 training observations, eight correlated features, and a weak signal, GBR and SVR overfit (training R^2 of 0.21 and 0.23 against negative test R^2), and the decision tree falls from a training F1-score of 59.0% to 23.0% in the test sample. The advantages of tree ensembles reported in other markets (Lolic, 2025) may depend on richer information or longer samples; our data do not allow us to test this. The K-Means regimes illustrate the same problem: they differ in next-day AbsMov in the training sample but not in the test sample. The conclusions do not depend on the class threshold (Table 16).

A further lesson concerns evaluation design. Because the open-to-close move is bounded by the intraday range, same-day PV explains 92.8% of the variance of same-day AbsMov in the test sample. Any design that lets same-day information into the predictors will therefore report high R^2 values that are mechanical and not forecasts. Reported accuracy should be checked for information timing and compared with benchmarks such as HAR.

Uncorrected Proof

5.2. Practical Implications

For risk monitoring of next-day price movements in the TSE, the evidence favors simple persistence-based measures (recent averages of absolute movements, HAR-type regressions) over more complex machine-learning models, which add estimation noise without out-of-sample gains. A calm recent regime (low AbsMA5) is the most reliable signal, but it identifies low-movement days only, and no model gives dependable forecasts of large days. These implications concern forecast accuracy only; the study does not test portfolio returns, risk-adjusted performance, transaction costs, or trading profitability.

5.3. Limitations

Several limitations apply. The analysis covers one index at daily frequency and a one-day horizon. The final test block (303 observations) is calmer than the training period; the rolling-origin evaluation mitigates this, but it keeps hyperparameters fixed at values selected on the first 1,209 observations, which overlap parts of the forecast blocks and introduce a mild look-ahead in those results, and it covers one market. The price-limit regime changed during the sample (Section 3.3) and is not modelled explicitly. The target is a noisy proxy for volatility, and intraday data would allow better realized-volatility measures. The GARCH benchmark is a plain Gaussian GARCH(1,1); EGARCH, HAR-RV with intraday data, and other volatility models were not estimated. The predictors are price-based and exclude macroeconomic, political, and sentiment information. Price limits truncate the observed moves. The HAR benchmark uses the 22-day average AbsMA22, which the machine-learning models do not receive, so the finding that they do not beat HAR may partly reflect a smaller information set; giving them the same input was not tested. Several selected hyperparameters lie at the edge of the search grids (Table 3), so the machine-learning models may not have been given a fair chance, and the tree was not regularised beyond depth and leaf size; wider grids were not examined. The decision tree is evaluated with accuracy, precision, recall, and F1 at a single threshold; threshold-free measures (ROC-AUC, PR-AUC) and balanced accuracy were not computed, although Table 16 shows that the conclusion does not depend on the threshold. The Diebold–Mariano test of GARCH against HAR was computed in the test block but not in the rolling evaluation, and hyperparameters were not re-tuned within each rolling window. Sub-period results by price-limit regime were not estimated. The class threshold and the rule thresholds are fixed values and were not optimized, Rule 3 fires on only 10 test days, and class-based results depend on the class definition. Finally, the models are compared in parallel and not combined in an integrated architecture.

5.4. Conclusion and Future Research

This study evaluates five machine-learning approaches for forecasting the next-day absolute percentage change of the TSE index under a chronological, leakage-controlled design and compares them with five simple benchmarks. None of the machine-learning models improves on a HAR-type regression that uses the additional 22-day average, in a single test block or in a rolling-origin evaluation; the decision tree is less accurate than a majority-class forecast; and K-Means regimes carry no out-of-sample information on the next move. Recent absolute movements carry a weak, persistent signal that simple benchmarks already capture. The results support starting from simple volatility benchmarks and caution against reading high accuracy figures obtained with same-day information as forecasting skill.

Future research could estimate EGARCH and HAR-RV benchmarks with intraday data, add macroeconomic, political, and sentiment variables, study longer horizons such as weekly averages, test ensemble and stacked architectures, evaluate economic value with transaction costs, and extend the analysis to sector indices and individual stocks.

Uncorrected Proof

Abbreviations

Abbreviations used in the article are listed in Table 17.

Table 17. Abbreviations

Abbreviation	Meaning	Abbreviation	Meaning
TSE	Tehran Stock Exchange	DPC	Daily Percent Change
OPC	Overnight Percent Change	PV	Percentage Volatility (intraday range)
MA3, MA5	3- and 5-day moving averages of DPC	CD	Direction indicator
AbsMov	Absolute Percent Change, |DPC|	AbsMA5, AbsMA22	5- and 22-day averages of AbsMov
GBR	Gradient Boosting Regression	SVR	Support Vector Regression

HAR	Heterogeneous Autoregressive model	DM	Diebold–Mariano test
GARCH	Generalized Autoregressive Conditional Heteroskedasticity	R^2_{OS}	Out-of-sample R^2 relative to the historical mean
MSE, RMSE	Mean Squared Error, Root MSE	ARI, NMI	Adjusted Rand Index, Normalized Mutual Information

References

1. Adebayo, O., Mensah, N., & Adukpo, T. K. (2025). Beyond Traditional Cash Flow Management: How Machine Learning and Scenario Planning Drive Financial Resilience. Available at SSRN: https://ssrn.com/abstract=5174613.

2. Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4), 885–905.

3. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of financial economics*, 61(1), 43-76. https://doi.org/10.1016/S0304-405X(01)00055-1.

4. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

6. Campisi, G., Muzzioli, S., & De Baets, B. (2024). A comparison of machine learning methods for predicting the direction of the US stock market on the basis of volatility indices. *International Journal of Forecasting*, 40(3), 869-880. DOI: https://doi.org/10.1016/j.ijforecast.2023.07.002.

7. Christensen, K., Siggaard, M., & Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, 21(5), 1680-1727.

8. Corsi, F. (2009). A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7(2), 174–196.

9. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

10. EghtesadOnline. (2021, May 11). SEO rolls back decision on daily price spread. https://www.eghtesadonline.com/en/news/786444/seo-rolls-back-decision-on-daily-price-spread

11. Financial Tribune. (2021). Bourse council gives OK to raising daily price spread. https://financialtribune.com/node/107939

12. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

13. Giot, P. (2005). Relationships between implied volatility indexes and stock index returns. *Journal of Portfolio Management*, 31(3), 92. DOI: 10.3905/jpm.2005.500363.

Uncorrected Proof

14. Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., & Pandey, N. (2021). Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Computer Science, 7, e340.

15. Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281–291.

16. Hördahl, P., & Packer, F. (2007). Understanding asset prices: an overview. Bank for International Settlements. Available at SSRN: https://ssrn.com/abstract=1190962 or http://dx.doi.org/10.2139/ssrn.1190962.

17. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.

18. Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research*, 196(2), 819-825. DOI: https://doi.org/10.1016/j.ejor.2008.03.025.

19. Kumar, D., Sarangi, P. K., & Verma, R. (2022). A systematic review of stock market prediction using machine learning and statistical techniques. Materials Today: Proceedings, 49, 3187-3191. DOI: https://doi.org/10.1016/j.matpr.2020.11.399.

20. Letteri, I. (2024). Stock Market Forecasting Using Machine Learning Models Through Volatility-Driven Trading Strategies. In *FEMIB* (pp. 96-103). DOI: 10.5220/0012607200003717.

21. Letteri, I., Della Penna, G., De Gasperis, G., & Dyoub, A. (2022). DNN-ForwardTesting: A new trading strategy validation using statistical timeseries analysis and deep neural networks. arXiv preprint arXiv:2210.11532.

22. Lolic, M. (2025). Tree-based methods of volatility prediction for the S&P 500 index. *Computation*, 13(4), 84. https://doi.org/10.3390/computation13040084.

23. Mansilla-Lopez, J., Mauricio, D., & Narváez, A. (2025). Factors, Forecasts, and Simulations of Volatility in the Stock Market Using Machine Learning. *Journal of Risk and Financial Management*, 18(5), 227. DOI: https://doi.org/10.3390/jrfm18050227.

24. Mensah, N., Oforiwaa Agbeduamenu, C., Nyarkoah Obodai, T., Kwame Adukpo, T. (2025). Leveraging machine learning techniques to forecast market volatility in the u.s. *EPRA International Journal of Economics, Business and Management Studies (EBMS)*. DOI: https://doi.org/10.36713/epra20998.

25. Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & Shahab, S. (2020). Deep learning for stock market prediction. Entropy, 22(8), 840. https://doi.org/10.3390/e22080840.

26. Pattanayak, A. M., Swetapadma, A., & Sahoo, B. (2024). Exploring different dynamics of recurrent neural network methods for stock market prediction-a comparative study. *Applied Artificial Intelligence*, 38(1), 2371706. DOI: https://doi.org/10.1080/08839514.2024.2371706.

27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Uncorrected Proof

28. Petneházi, G., & Gáll, J. (2019). Exploring the predictability of range-based volatility estimators using recurrent neural networks. *Intelligent Systems in Accounting, Finance and Management*, 26(3), 109-116. DOI: https://doi.org/10.1002/isaf.1455.

29. Ramos-Pérez, E., Alonso-González, P. J., & Núñez-Velázquez, J. J. (2019). Forecasting volatility with a stacked model based on a hybridized Artificial Neural Network. *Expert Systems with Applications*, 129, 1-9. DOI: https://doi.org/10.1016/j.eswa.2019.03.046.

30. Rodriguez, F. S., Norouzzadeh, P., Anwar, Z., Snir, E., & Rahmani, B. (2024). A machine learning approach to predict the S&P 500 absolute percent change. *Discover Artificial Intelligence*, 4(1), 8. https://doi.org/10.1007/s44163-024-00104-9.

31. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

32. Sadhukhan, S., Gopaliya, S. M., & Jain, P. (2022). A novel approach to quantify volatility prediction. arXiv preprint arXiv:2211.00528. DOI: https://doi.org/10.48550/arXiv.2211.00528.

33. Seo, M., Lee, S., & Kim, G. (2019). Forecasting the volatility of stock market index using the hybrid models with google domestic trends. Fluctuation and Noise Letters, 18(01), 1950006. DOI: https://doi.org/10.1142/S0219477519500068.

34. Srivastava, A. K., Srivastava, A., Singh, S., Sugandha, S., Tripta, & Gupta, S. (2022). Design of Machine-Learning Classifier for Stock Market Prediction. SN Computer Science, 3(1), 88. DOI: https://doi.org/10.1007/s42979-021-00970-5.

35. Tehran Stock Exchange. (2021, May 11). Tehran Stock Exchange to establish symmetric price fluctuation range. Mondo Visione. https://m.mondovisione.com/news/tehran-stock-exchange-to-establish-symmetric-price-fluctuation-range-2021511

36. Tehran Stock Exchange. (2022a, April 16). Tehran Stock Exchange extends daily price changes. Mondo Visione. https://mondovisione.com/media-and-resources/news/tehran-stock-exchange-extends-daily-price-changes-2022416/

37. Tehran Stock Exchange. (2022b, August 27). Tehran Stock Exchange: Price fluctuation to increase. Mondo Visione. https://mondovisione.com/media-and-resources/news/tehran-stock-exchange-price-fluctuation-to-increase-2022827/

38. Tehrani, R., Mohammadi, S. and Porebrahimi, M. (2011). Modeling and forecasting the volatility of Tehran Exchange Dividend Price Index (TEDPIX). *Financial Research Journal*, 12(30), 23-36. DOI: 20.1001.1.10248153.1389.12.30.1.1.

39. Yan, K., & Li, Y. (2024). Machine learning-based analysis of volatility quantitative investment strategies for American financial stocks. *Quant. Financ. Econ*, 8, 364-386. DOI: 10.3934/QFE.2024014.

Uncorrected Proof

40. Yao, W., Gu, Y., Chang, S., Li, J., Zhao, Q., & Ge, F. (2022, December). Stock price analysis and forecasting based on machine learning. *In Third International Conference on Computer Science and Communication Technology* (ICCSCT 2022) (Vol. 12506, pp. 1503-1510). SPIE. DOI: https://doi.org/10.1117/12.2662176.

41. Zhang, C., Zhang, Y., Cucuringu, M., & Qian, Z. (2024). Volatility forecasting with machine learning and intraday commonality. *Journal of Financial Econometrics*, 22(2), 492-530. DOI: https://doi.org/10.1093/jjfinec/nbad005.

Appendix A. Illustrative Daily Observations

Table A1 shows the first ten feature rows (the first observations after the 21-day warm-up) for the variables used in the analysis. Volume is not included because the original volume column was found to be invalid (Section 3.3).

Table A1. Illustrative Daily Observations and Engineered Variables

Date	OP	HP	LP	CP	DPC	OPC	PV	MA3	MA5	CD	AbsMov	AbsMA5
2019-04-29	210,440.0	214,867.0	210,440.0	214,867.6	2.104	0.572	2.104	1.377	0.810	1	2.104	0.981
2019-04-30	216,899.0	219,086.0	216,637.0	216,642.0	-0.118	0.945	1.130	1.159	0.872	-1	0.118	0.919
2019-05-01	217,222.0	220,108.0	217,222.0	220,109.0	1.329	0.268	1.329	1.105	1.068	1	1.329	1.116
2019-05-04	221,546.0	224,589.0	221,546.0	224,589.6	1.374	0.653	1.374	0.861	1.236	1	1.374	1.283
2019-05-05	226,475.0	226,952.0	224,057.0	224,057.3	-1.068	0.839	1.292	0.545	0.724	-1	1.068	1.199
2019-05-06	223,368.0	223,368.0	215,818.0	215,820.6	-3.379	-0.308	3.498	-1.024	-0.372	-1	3.379	1.454
2019-05-07	213,710.0	213,710.0	207,966.0	208,197.1	-2.580	-0.978	2.762	-2.342	-0.865	-1	2.580	1.946
2019-05-08	207,779.0	211,344.0	207,779.0	211,344.1	1.716	-0.201	1.716	-1.414	-0.787	1	1.716	2.023
2019-05-11	211,301.0	212,479.0	209,154.0	209,154.9	-1.016	-0.020	1.590	-0.626	-1.265	-1	1.016	1.952
2019-05-12	208,660.0	208,660.0	203,489.0	204,069.3	-2.200	-0.237	2.541	-0.500	-1.492	-1	2.200	2.178

Source: Authors' calculations. OP, HP, LP, CP are the opening, highest, lowest and closing prices.

Uncorrected Proof