Journal of Contemporary Applied Mathematics
V. 17, No 2, 2027, January
ISSN 2222-5498, E-ISSN 3006-3183
https://doi.org/10.62476/jcam171.24

Filtration-Consistent Statistical Learning and Dependence-Aware Inference for Directional Forecasts of Iron Ore Futures

S. Gurbanov, S. Frumina, Z. F. Mamedov, L. Zotov, A. Borodin*, O. Kulikova

Abstract. We forecast the direction of iron ore futures returns on the Dalian Commodity Exchange using only information available at the forecast origin. Across 1,531 observations from 2020–2026, dependence-robust tests find no significant directional predictability for penalised logistic regression or three tree ensembles under daily-sign or triple-barrier labelling. Bootstrap upper confidence limits for the area under the ROC curve reach 0.54 and 0.60, respectively, leaving room for modest predictive effects. Changing only the predictor timestamp rule raises the realised Sharpe ratio from −0.06 to 0.66 for logistic regression and from −0.08 to 0.39 for the random forest. The formal tests yield the same conclusions under both timestamp rules. These results support reporting backtest performance alongside tests of predictive ability under an explicit information-availability rule.

Key Words and Phrases: filtration, look-ahead bias, statistical learning, dependence-robust inference, stationary bootstrap, Model Confidence Set, decision-tree ensembles, iron ore futures, directional forecasting

2010 Mathematics Subject Classifications: 62M10, 62P20, 68T09, 91G15

1. Introduction

Iron ore is a principal input to steel production, and the iron ore futures contract listed on the Dalian Commodity Exchange (DCE) since October 2013 is a central venue for its price discovery in China. Several studies report statistically or economically significant directional predictability in commodity futures, including Chinese steel-complex contracts, using machine-learning and related classification methods, often accompanied by profitable backtests [20, 28, 9]. Large

*Corresponding author.

 © *2010 jcam* All rights reserved.

out-of-sample gains are difficult to reconcile with weak-form market efficiency [11] and with the documented fragility of out-of-sample return predictability in equity markets [29, 5]. Two features of an empirical design can generate apparent gains in the absence of genuine predictability: predictors that are not yet observable at the forecast origin, and inference that ignores the serial dependence created by overlapping labels and persistent predictors. This paper addresses both.

Let $\mathcal{F}_t$ denote the information available to a decision-maker at the close of the DCE day session, c_t = 15:00 Beijing time, on trading day t, and let y_t denote the directional label attached to forecast origin t; in the simplest case $y_t = \mathbf{1}\{r_{t+1} > 0\}$, where $r_{t+1} = \ln(P_{t+1}/P_t)$ is the log return between consecutive DCE closes. A predictor is admissible only if it is $\mathcal{F}_t$-measurable. A large part of the conventional predictor set — the VIX, the US dollar index, the ten-year Treasury yield, the MOVE index and daily news-based uncertainty indices — is recorded at or after the US close, which falls at 04:00–05:00 Beijing time on date $t+1$ depending on daylight-saving time. The date-t value of such a variable therefore becomes known strictly inside the interval $(c_t, c_{t+1}]$ over which the target is realised. Because predictor and target nonetheless carry distinct date indices, the violation is easily overlooked in routine data pipelines.

The contribution of the paper is methodological and has three parts. First, we formulate filtration-consistent statistical learning as an explicit information-availability protocol: every predictor group is assigned a documented observation timestamp in Beijing time, and each observation enters the predictor vector only at the first forecast origin at which it is observable (Section 3.3, Proposition 1). Second, we combine this protocol with dependence-aware inference: purged and embargoed expanding-window estimation, uniqueness weighting under overlapping labels, stationary-bootstrap intervals for discrimination measures, heteroskedasticity- and autocorrelation-consistent (HAC) pairwise tests, and bootstrap multiple-comparison procedures. We show that statistics derived under independence can produce nominal rejections under overlapping labels that do not survive dependence-robust inference. Third, we use a controlled experiment in which only the timestamp rule changes to measure how filtration misspecification affects statistical and economic evaluation measures.

Applied to DCE iron ore futures over 2016–2026, the design does not reject the null hypothesis of no directional predictability for any classifier under the daily-sign label, and the nominal rejections that occur under triple-barrier labelling do not survive dependence-robust inference. We interpret this as an absence of statistically detectable predictability at the precision this sample affords, not as evidence that predictability is absent. Under the misspecified filtration, realised Sharpe ratios change markedly while the conclusions of the formal tests do not change.

2. Methodological background

2.1. Tests of directional and predictive ability

Directional accuracy cannot be evaluated by quadratic loss alone. Pesaran and Timmermann [25] derive a non-parametric statistic that compares the realised hit rate with the rate implied by independence of forecast and realisation; because the comparison is made against that independence rate, a rule that always predicts the same direction has an expected statistic of zero regardless of class imbalance. Pairwise comparison of forecast losses follows Diebold and Mariano [10], with the finite-sample correction of Harvey, Leybourne and Newbold [16] and the HAC variance estimator of Newey and West [24]. Because a specification search inflates the family-wise error rate, pairwise inference is insufficient. White [30] proposes a bootstrap Reality Check for the composite null that no model in a candidate set outperforms a benchmark.

Hansen [14] shows that the Reality Check is conservative when the candidate set contains poor models and proposes a studentised statistic with consistent recentring. Hansen, Lunde and Nason [15] construct the Model Confidence Set (MCS), a set of models that contains the best model(s) with a given asymptotic probability. All three procedures are implemented here with the stationary bootstrap of Politis and Romano [26].

For the AUC we use the asymptotic variance of DeLong, DeLong and Clarke-Pearson [8]. Because that variance is derived under independent sampling, which fails for overlapping financial labels, we supplement it with a stationary-bootstrap interval.

2.2. Statistical learning under dependent sampling

Gu, Kelly and Xiu [13] show that tree ensembles and neural networks improve cross-sectional prediction of equity returns, with the gains attributed to nonlinearity and predictor interactions. Their setting has a large cross-section; time-series prediction of a single contract has a far lower signal-to-noise ratio and an effective sample bounded by the length of the series.

López de Prado [22] catalogues the biases specific to this setting: overlap of labels when the horizon exceeds one step, the resulting need for purging and embargo in cross-validation, non-uniqueness of observations and the consequent overstatement of effective sample size, and backtest overfitting under repeated specification search [1]. The present paper adds a distinct requirement — that each predictor be measurable with respect to the filtration at the forecast origin — and treats it as part of the statistical design rather than as a data-cleaning step.

3. Data and information timing

3.1. Target

The target series is the closing price P_t of the DCE main continuous iron ore futures contract, in CNY per tonne, from 4 January 2016 to 30 April 2026 (N = 2,508 trading days). The annualised volatility of $r_t = \ln(P_t/P_{t-1})$ is 38.9%, and exactly zero returns account for 1.6% of observations, which gives no indication of stale-quote contamination.

3.2. Predictors

Table 1 summarises the data sources from which the 172 predictors are constructed.

Table 1: Data sources

Predictor group	Variables	Frequency	Primary source (data vendor)
Target	DCE iron ore futures, main continuous contract, close	daily	DCE
Steel complex	SHFE hot-rolled coil futures, main continuous (OHLC, volume, open interest, settlement price); composite steel, iron ore and coke price indices	daily	SHFE (via AKShare/Sina Finance); industry composite indices
Chinese equities	Shanghai Composite index; CSI 300 ETF (510300)	daily	SSE (via Yahoo Finance)
Global markets	US dollar index (DXY); 10-year US Treasury yield; VIX; MOVE; Hang Seng Properties index; Baltic Dry Index	daily	ICE; Cboe; ICE BofA; Hang Seng Indexes (via Yahoo Finance); Baltic Exchange (via Investing.com)
Foreign exchange	USD/CNY (onshore); USD/CNH and CNY/CNH (offshore)	daily	CFETS and offshore market (via Investing.com)
News-based uncertainty	Geopolitical risk indices GPR, GPRT, GPRA; US economic policy uncertainty (EPU)	daily	Caldara and Iacoviello [4]; Baker, Bloom and Davis [2]
Chinese macro	NBS and Caixin manufacturing PMI; NBS non-manufacturing PMI; CPI; PPI; M0–M2; total social financing; real-estate indicators	monthly	NBS; Caixin/S&P Global; PBoC
Chinese uncertainty	China EPU and trade policy uncertainty; China GPR	monthly	Baker, Bloom and Davis [2]; Caldara and Iacoviello [4]

Figure 1: Closing price of the DCE main continuous iron ore futures contract, January 2016 to April 2026. Blue shading marks the initial estimation window (2017–2019), which expands with each annual re-estimation; red shading marks the evaluation period (2020–2026) over which all reported statistics are computed. The evaluation window spans several distinct price regimes, including the 2020–2021 rally, the correction that coincided with the 2021 steel-output restrictions, and a subsequent decline during the contraction of the Chinese property sector, which makes it a demanding test of parameter stability.

3.3. Information filtration and admissible predictors

For predictor series j, let $z_j(d)$ denote the observation with reference date d and let $a_j(d)$ denote its availability timestamp, that is, the earliest Beijing clock time at which the value is fixed and publicly observable. Forecast origins are indexed by DCE trading days, and c_t is the day-session close on day t.

The value of series j that enters the predictor vector at origin t is

$$x_{j,t} = z_j\big(d_j^*(t)\big), \qquad d_j^*(t) = \max\{d : a_j(d) \le c_{t-\Delta_j}\}, \tag{1}$$

where Δ_j is a safety margin in DCE trading days: $\Delta_j = 0$ for daily market series and $\Delta_j = 1$ for scheduled macroeconomic releases. Rule (1) makes every $x_{j,t}$ $\mathcal{F}_t$-measurable by construction. For a daily series observed after the DCE close, it implies that the date-t value is first used at origin $t+1$, that is, a lag $\ell(x) = 1$. Table 2 records, for each predictor group, the timestamp $a_j(\cdot)$ and the resulting lag.

Formally, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathcal{T}}, \mathbb{P})$ be a filtered probability space, where $\mathcal{T}$ is the set of DCE trading days, and let the information available to the forecaster at the day-session close c_t be

$$\mathcal{F}_t = \sigma\big\{z_j(d) : j = 1, \dots, p,\ a_j(d) \le c_t\big\} \vee \sigma\{P_s : s \le t\},$$

the σ-algebra generated by every observation published by c_t together with the DCE closing prices up to day t. Because c_t is increasing in t, $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \le t$, so $\{\mathcal{F}_t\}$ is a filtration. A label y_t defined over the horizon h is $\mathcal{F}_{t+h}$-measurable but in general not $\mathcal{F}_t$-measurable. A predictor vector x_t is admissible at origin t if $x_t \in L^0(\Omega, \mathcal{F}_t; \mathbb{R}^p)$, that is, if it is $\mathcal{F}_t$-measurable. The date index d of a series labels its reference period and carries no information about measurability; measurability is determined by the availability timestamp $a_j(d)$ alone.

Table 2: Observation timestamps and admissible lags by predictor group (Beijing time, UTC+8)

Predictor group	Series (vendor)	Availability timestamp $a_j(d)$ of the date-d observation	Before c_t on $d = t$?	Lag ℓ applied (DCE trading days)
Target and own-price technical	DCE iron ore main continuous close	15:00, day t (day-session close; trading day t includes the night session 21:00–23:00 of the preceding evening)	yes	0

Table 2 (continued)

Predictor group	Series (vendor)	Availability timestamp $a_j(d)$ of the date-d observation	Before c_t on $d = t$?	Lag ℓ applied (DCE trading days)
Steel complex: futures	SHFE HRC close, volume, open interest	15:00, day t (day-session close)	yes	0
Steel complex: futures	SHFE HRC settlement price	after 15:00, day t (computed after the close)	no	1
Steel complex: indices	Composite steel, iron ore and coke indices	evening, day t	no	1
Chinese equities	Shanghai Composite; CSI 300 ETF (Yahoo Finance)	15:00, day t (closing call auction 14:57–15:00)	yes	0
Hong Kong equities	Hang Seng Properties (Yahoo Finance)	16:00–16:10, day t (HKEX close and closing auction)	no	1
Onshore FX	USD/CNY end-of-day quote (Investing.com; vendor date = UTC calendar date)	after the 16:30 official close; onshore trading continues to 23:30 (before 3 January 2023) or 03:00 on day $t + 1$ (thereafter)	no	1
Offshore FX	USD/CNH, CNY/CNH (Investing.com)	end-of-day snapshot ≈ 17:00 New York = 05:00–06:00, day $t + 1$	no	1
US markets	DXY (ICE); 10-year UST yield (Cboe); VIX (Cboe); MOVE (ICE BofA)	US close 16:00–17:00 New York = 04:00–06:00, day $t + 1$ (depending on daylight-saving time)	no	1
Freight	Baltic Dry Index (Investing.com)	≈ 13:00 London = 20:00 (BST) or 21:00 (GMT), day t	no	1
News-based indices, daily	GPR, GPRT, GPRA; US EPU (final vintage)	end of news day t (US time), i.e. after c_t	no	1
Chinese macro: PMI	NBS manufacturing and non-manufacturing PMI	09:30 on the last calendar day of reference month M	release day	release + 1
Chinese macro: PMI	Caixin manufacturing PMI	09:45 on the first business day of month $M + 1$	release day	release + 1
Chinese macro: prices	CPI, PPI	09:30, ≈ 9th–10th of month $M + 1$	release day	release + 1
Chinese macro: money and credit	M0–M2; total social financing	≈ 10th–15th of month $M + 1$; time of day varies	release day	release + 1
Chinese macro: real estate	Investment, sales, climate index	10:00, ≈ 15th–17th of month $M + 1$ (January and February combined)	release day	release + 1

Table 2 (continued)

Predictor group	Series (vendor)	Availability timestamp $a_j(d)$ of the date-d observation	Before c_t on $d = t$?	Lag ℓ applied (DCE trading days)
Chinese uncertainty, monthly	China EPU and TPU; China GPR	published early in month $M + 1$; exact dates not archived, so release date set to the last calendar day of month $M + 1$	release day	release + 1

Note. c_t = 15:00 on DCE trading day t. "Lag ℓ" is the number of DCE trading days between the reference date and the first origin at which the observation is used, as implied by rule (1). For monthly series, "release + 1" means the first DCE trading day after the release date. Approximate release days are typical values; the protocol uses the release date of each individual release.

Proposition 1 (admissibility)**.** *Suppose that the availability timestamps $a_j(\cdot)$ are deterministic and $\Delta_j \geq 0$ for every j. Then:*

(i) for every j and t, the value $x_{j,t}$ defined by rule (1) *is $\mathcal{F}_t$-measurable, so x_t is admissible;*

(ii) if the classifier $\widehat{g}$ is estimated only from pairs (x_i, y_i) that are $\mathcal{F}_s$-measurable for some $s \leq t$, the forecast $\widehat{p}_t = \widehat{g}(x_t)$ is $\mathcal{F}_t$-measurable;

(iii) under the null hypothesis (3) *of Section 4.1, every forecast satisfying (ii) is independent of y_t, and its population AUC equals* 1/2.

Proof. (i) Since $c_{t-\Delta_j} \leq c_t$, every reference date d with $a_j(d) \leq c_{t-\Delta_j}$ satisfies $a_j(d) \leq c_t$, so $z_j(d)$ is $\mathcal{F}_t$-measurable by the definition of $\mathcal{F}_t$. Because a_j is deterministic, the index $d_j^*(t)$ in (1) is deterministic, and $x_{j,t} = z_j(d_j^*(t))$ is $\mathcal{F}_t$-measurable. (ii) The map $\widehat{g}$ is a measurable function of $\mathcal{F}_s$-measurable data and is therefore $\mathcal{F}_s$-measurable; since $\mathcal{F}_s \subseteq \mathcal{F}_t$ and x_t is $\mathcal{F}_t$-measurable by (i), so is $\widehat{p}_t$. The purging rule of Section 4.2, which restricts training indices to $i \leq s - \vartheta - e$, guarantees that every training label is realised by c_{s-e}, so the condition holds for the forecasts in this paper. (iii) Under (3), $\Pr(y_t = 1 \mid \mathcal{F}_t) = \pi_0$ almost surely for a constant π_0. By the tower property, $\Pr(y_t = 1 \mid \widehat{p}_t) = E\big[\Pr(y_t = 1 \mid \mathcal{F}_t) \mid \widehat{p}_t\big] = \pi_0$, so the binary variable y_t is independent of $\widehat{p}_t$. The conditional distributions of $\widehat{p}_t$ given $y_t = 1$ and $y_t = 0$ therefore coincide, and the population AUC, $\Pr(\widehat{p}^+ > \widehat{p}^-) + \frac{1}{2}\Pr(\widehat{p}^+ = \widehat{p}^-)$, equals 1/2.

Remark 1. *Part (iii) does not extend to non-admissible predictors. Suppose a series with* $a_j(t) \in (c_t, c_{t+1}]$ *— for example a variable recorded at the US close — enters the predictor vector at origin* t *with its date-*t *value, as under calendar-date alignment. Then* x_t *is* $\mathcal{F}_{t+1}$*- but in general not* $\mathcal{F}_t$*-measurable. Because the daily-sign label* $y_t = \mathbf{1}\{r_{t,t+1} > 0\}$ *is also* $\mathcal{F}_{t+1}$*-measurable, any dependence between the overnight movement of such a variable and the next DCE return can raise the AUC above* $1/2$ *even when* (3) *holds for every admissible predictor. Filtration misspecification therefore changes the hypothesis being tested, not merely the noise in its estimate; Section 5.3 measures the size of this effect in the data.*

Clock and daylight-saving time. All timestamps are converted to Beijing time (UTC+8, no daylight-saving time). The US close therefore falls at 04:00 Beijing time from March to November and at 05:00 otherwise, and the London-based Baltic Exchange publication at approximately 20:00 or 21:00. Because every such timestamp lies after c_t and before c_{t+1} throughout the year, daylight-saving transitions do not change the lag.

Calendar alignment. The predictor matrix is built on the DCE trading calendar. When a foreign market is closed on a DCE trading day, the most recent admissible observation is carried forward; when several foreign observations accumulate over a Chinese holiday, rule (1) selects the most recent one. Vendor date stamps (for example, the UTC calendar dates used by Investing.com) are mapped to the timestamps in Table 2 before rule (1) is applied.

Monthly releases. Monthly series are indexed by their release date rather than their reference month and held constant until the next release. The additional one-day margin ($\Delta_j = 1$) guards against releases made after the DCE close and against uncertainty about the exact time of release. Where exact historical release dates are not archived (the monthly China EPU, trade policy uncertainty and GPR indices), the release date is set conservatively to the last calendar day of month $M + 1$.

Data vintages. All series are taken in the vintage available at the time of download (May 2026). For market prices this is immaterial. Chinese PMI and CPI releases are, in practice, not revised; monetary aggregates, total social financing and real-estate statistics occasionally are, and the news-based indices are re-estimated by their compilers. The design is therefore real-time with respect to the timing of information but pseudo-real-time with respect to data vintages.

We retain $\ell = 0$ for the Chinese equity indices and the SHFE hot-rolled coil close, because the SSE and SHFE day sessions end at 15:00, the same time as the DCE day session. This is the least conservative element of the specification:

it treats information fixed at the same instant as the target price as available, and it therefore biases the analysis towards rejecting the null rather than towards accepting it.

Figure 2: Release times of the data sources relative to the forecast origin, in Beijing clock time. The red band is the interval $(c_t, c_{t+1}]$ over which the daily target is realised; it also contains the DCE night session (21:00–23:00), which is not shown. Blue markers denote the endpoints of the interval and grey markers the release times of other sources. The HKEX close, the Baltic Dry Index publication and the US close all lie strictly inside the interval, so the date-t values of these variables are not $\mathcal{F}_t$-measurable even though their date index precedes that of the target.

3.4. Unavailable data

The DCE series contains date and closing price only; contract volume, open interest and intraday information are unavailable. An alternative US-dollar-denominated iron ore series was examined and rejected: its volume is zero in 84.7% of observations and its high equals its low in 93.9%, which is consistent with settlement or indicative prices rather than transaction prices. Order-flow information is therefore proxied by the SHFE hot-rolled coil contract, for which volume, open interest and full OHLC data are available.

4. Statistical framework

4.1. The forecasting problem

Each label is indexed by its forecast origin t. Let $r_{t,t+h} = \ln(P_{t+h}/P_t)$. Three labelling operators are considered. Operator A (daily sign) sets $y_t = \mathbf{1}\{r_{t,t+1} > 0\}$. Operator B (weekly sign) sets $y_t = \mathbf{1}\{r_{t,t+5} > 0\}$. Operator C (triple barrier) uses the exponentially weighted volatility estimate σ_t with span 20 and the stopping time

$$\theta_t = \inf\bigl\{ s \in (t, t+T] : \bigl|\ln(P_s/P_t)\bigr| \geq k\,\sigma_t\sqrt{T} \bigr\} \wedge (t+T), \tag{2}$$

and sets $y_t = \mathbf{1}\{\ln(P_{\theta_t}/P_t) > 0\}$, with barrier multiplier $k = 0.75$ and vertical barrier $T = 10$ trading days.

Let $x_t \in \mathbb{R}^p$, $p = 172$, denote the $\mathcal{F}_t$-measurable predictor vector constructed by rule (1). A classifier is a measurable map $\widehat{g} : \mathbb{R}^p \to [0,1]$ delivering $\widehat{p}_t = \widehat{g}(x_t)$ as an estimate of $\pi_t = \Pr(y_t = 1 \mid \mathcal{F}_t)$, with induced directional forecast $\widehat{y}_t = \mathbf{1}\{\widehat{p}_t > 1/2\}$. The null hypothesis under test is

$$H_0 : \Pr(y_t = 1 \mid \mathcal{F}_t) = \Pr(y_t = 1) \quad \text{a.s.}, \tag{3}$$

that is, the conditional probability of an upward move is constant and the predictor vector carries no directional information. Under H_0 the population AUC of any $\mathcal{F}_t$-measurable forecast equals $1/2$, the Pesaran–Timmermann statistic is asymptotically standard normal under its independence assumptions, and no classifier has a smaller expected Brier loss than the correctly specified constant forecast.

4.2. Estimation window, purging and embargo

The estimation sample begins on 2 January 2017, after a warm-up period for rolling windows of up to 252 observations; the evaluation sample runs from 2 January 2020 to 29 April 2026 ($n = 1{,}531$).

Forecasts are generated by an expanding-window scheme with annual re-estimation, which yields seven disjoint evaluation blocks. Writing s for the first index of a block, the estimation sample is restricted to indices $i \leq s - \vartheta - e$, where ϑ is the label span parameter ($\vartheta = 1$ under operators A and B and $\vartheta = 10$ under operator C) and $e = 5$ is the embargo. Under operator B the combined gap $\vartheta + e = 6$ exceeds the five-day label span, and under operator C the gap of 15 exceeds the maximal label span of 10, so in every case no training label overlaps the evaluation block. This removes the mechanical dependence between training and evaluation errors that overlapping labels would otherwise induce.

Under operator C, labels overlap heavily. Let $\kappa_s = \sum_i \mathbf{1}\{i \leq s \leq \theta_i\}$ denote the number of labels concurrent at date s. The average uniqueness of observation i is

$$u_i = \frac{1}{\theta_i - i + 1} \sum_{s=i}^{\theta_i} \frac{1}{\kappa_s}. \tag{4}$$

Its sample mean is 0.135, so that a typical label is concurrent with roughly seven labels on average. Observations are weighted by u_i in all objective functions, and the effective sample size is $\sum_i u_i$ rather than the nominal count.

Hyperparameters are selected at each re-estimation step by inner expanding-window cross-validation with two splits, purged and embargoed in the same way, with the AUC as the selection criterion. All preprocessing operators (median imputation and standardisation) are estimated on training folds only.

4.3. Calibration of the triple-barrier operator

Table 3 and Figure 3 report the calibration of the triple-barrier operator. Widening the barriers raises the economic content of each label but lengthens the expected label span and intensifies overlap, which reduces the effective sample. The specification $k = 0.75$, $T = 10$ balances these effects: about 63% of labels are set by a horizontal barrier, while mean uniqueness remains above 0.13. The calibration uses only the distribution of the labels, not their association with the predictors.

4.4. Classifiers

Penalised logit. The fitted probability satisfies $\operatorname{logit}(\widehat{p}_t) = x_t' \widehat{\beta}$, where $\widehat{\beta}$ minimises the weighted ℓ_2-penalised negative log-likelihood

$$\widehat{\beta} = \arg\min_{\beta} \left[-\sum_i u_i \big\{ y_i \ln \pi_i(\beta) + (1 - y_i) \ln(1 - \pi_i(\beta)) \big\} + \frac{1}{2C} \, \|\beta\|_2^2 \right], \tag{5}$$

with $C \in \{0.01, 0.1, 1.0\}$ selected by inner cross-validation.

Figure 3: Trade-off governing the choice of barrier width. Blue curves (left axis) plot the share of labels set by a horizontal rather than the vertical barrier; orange curves (right axis) plot mean label uniqueness. Solid lines: $T = 10$ trading days; dashed lines: $T = 20$. The selected specification ($k = 0.75$, $T = 10$) is circled.

Table 3: Calibration of the triple-barrier operator

Barrier multiplier k	Vertical barrier T (trading days)	Share of labels set by a horizontal barrier	Share of positive labels, $\Pr(y_t = 1)$	Mean absolute log return at stopping, $\mathbb{E}\,\lvert r_{t,\theta_t}\rvert$	Mean label span $\mathbb{E}(\theta_t - t)$ (trading days)	Mean uniqueness $\bar{u}$
0.50	10	0.845	0.552	0.0476	4.82	0.182
0.75†	10†	0.632	0.559	0.0544	6.69	0.135
1.00	10	0.434	0.558	0.0574	7.99	0.113
1.50	10	0.198	0.554	0.0595	9.25	0.098
0.50	20	0.895	0.551	0.0627	8.22	0.120
0.75	20	0.688	0.556	0.0747	12.42	0.080
1.00	20	0.498	0.552	0.0797	15.23	0.064
1.50	20	0.255	0.552	0.0823	18.00	0.053

Note. † Selected specification. Barrier width is $k\sigma_t\sqrt{T}$, with σ_t the exponentially weighted daily volatility (span 20). $r_{t,\theta_t} = \ln(P_{\theta_t}/P_t)$. Mean uniqueness is defined in equation (4). The statistics depend only on the price series and not on the predictors, so the calibration does not use predictor–label associations from the evaluation window.

Decision trees. A classification tree partitions the predictor space into rectangles $R_1, \ldots, R_M$ by recursive binary splitting. Let $\widehat{\pi}_m$ denote the weighted proportion of positive labels in node m and $G(m) = 2\widehat{\pi}_m(1-\widehat{\pi}_m)$ the Gini index. At each node the split (j, v) on variable j at threshold v solves

$$\max_{j,\,v}\left[G(m) - \frac{N_L}{N_m}G(m_L) - \frac{N_R}{N_m}G(m_R)\right], \tag{6}$$

where the node masses $N_m = \sum_{i\in m} u_i$ are sums of uniqueness weights. The fitted probability in a terminal node is $\widehat{\pi}_m$. A single tree is a low-bias, high-variance estimator: it can represent arbitrary interactions among predictors, but its partition is unstable under resampling of the training set. This motivates the two ensemble constructions below.

Random forest [3]. $B = 200$ trees are grown on bootstrap resamples, with the candidate split variables at each node restricted to a random subset of size $\lceil 0.3p \rceil$. The forecast is $\widehat{p}_t = B^{-1}\sum_{b=1}^{B}\widehat{\pi}^{(b)}(x_t)$, whose variance is

$$\operatorname{Var}(\widehat{p}_t) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2, \tag{7}$$

where σ^2 is the variance of a single-tree forecast and ρ is the pairwise correlation between trees. Randomising the split candidates reduces variance through ρ rather than through B alone. Tree depth is restricted to $\{3, 6\}$ and the minimum terminal-node mass to 30, which regularises the partition given the low signal-to-noise ratio.

Gradient boosting [12], XGBoost [6] and LightGBM [18]. Trees are fitted sequentially to the functional gradient of the logistic loss $L(\cdot,\cdot)$:

$$F_m(x) = F_{m-1}(x) + \nu f_m(x), \qquad f_m = \arg\min_{f\in\mathcal{T}} \sum_i u_i \, L\big(y_i, F_{m-1}(x_i) + f(x_i)\big), \tag{8}$$

with shrinkage $\nu = 0.03$ and $M = 250$ iterations, $\mathcal{T}$ denoting the class of admissible trees. XGBoost uses a second-order approximation to L with explicit ℓ_2 regularisation of leaf weights; LightGBM grows trees leaf-wise under a bounded leaf count. Row and column subsampling fractions are 0.8 and 0.5. Maximum depth is restricted to $\{2, 3\}$ for XGBoost and the leaf count to $\{7, 15\}$ for LightGBM. Boosting reduces bias sequentially and is correspondingly more exposed than bagging to fitting idiosyncratic variation; shallow base learners and small shrinkage limit that exposure.

Tree ensemble. The unweighted average of the random forest, XGBoost and LightGBM probability forecasts.

Benchmark. The expanding-window unconditional positive-class frequency

$$\widehat{p}_t^{\,b} = \frac{1}{|I_t|} \sum_{i\in I_t} y_i, \qquad I_t = \{i : \tau_i \le t\}, \tag{9}$$

where τ_i is the trading day on which label y_i is determined ($\tau_i = i + 1$ under operator A, $\tau_i = i+5$ under operator B and $\tau_i = \theta_i$ under operator C). Restricting the average to labels realised by the forecast origin makes $\widehat{p}_t^{\,b}$ $\mathcal{F}_t$-measurable; under operator A, $I_t = \{i : i \le t - 1\}$. The benchmark induces the always-long directional rule while remaining a probability forecast that can be evaluated under quadratic loss. Because $\widehat{p}_t^{\,b}$ varies with t, the pooled out-of-sample AUC of the benchmark is not identically 1/2 (Tables 4 and 5 report 0.4915 and 0.5113); the deviation reflects only drift of the running base rate and carries no information from the predictors.

4.5. Test battery

Let $L_{k,t} = (\widehat{p}_{k,t} - y_t)^2$ denote the Brier loss of model k and $L_{b,t}$ that of the benchmark. The Brier loss is a strictly proper scoring rule, so comparisons of probability forecasts under it are well posed. Two differentials of opposite orientation are used: $d_{k,t}^{DM} = L_{k,t} - L_{b,t}$ in pairwise comparisons, so that positive statistics count against the model, and $d_{k,t} = L_{b,t} - L_{k,t}$ in the composite nulls, so that positive values count in its favour.

Directional accuracy. With $\widehat{P} = n^{-1}\sum_t \mathbf{1}\{\widehat{y}_t = y_t\}$, $\widehat{P}_y = n^{-1}\sum_t y_t$, $\widehat{P}_x = n^{-1}\sum_t \widehat{y}_t$ and $\widehat{P}^* = \widehat{P}_y\widehat{P}_x + (1-\widehat{P}_y)(1-\widehat{P}_x)$ the hit rate implied by independence, the Pesaran–Timmermann statistic [25] is

$$S_{PT} = \frac{\widehat{P} - \widehat{P}^*}{\left\{\widehat{V}(\widehat{P}) - \widehat{V}(\widehat{P}^*)\right\}^{1/2}} \xrightarrow{d} N(0,1) \quad \text{under } H_0 \text{ and serial independence.} \tag{10}$$

Discrimination. The AUC is tested against 1/2 with the DeLong variance [8]. Let n_1 and n_0 be the numbers of positive and negative labels, and let s_{10}^2 and s_{01}^2 be the sample variances of the structural components V_{10} (computed over positive cases) and V_{01} (computed over negative cases). Then

$$\widehat{V}(\widehat{AUC}) = \frac{s_{10}^2}{n_1} + \frac{s_{01}^2}{n_0}, \qquad z = \frac{\widehat{AUC} - 1/2}{\widehat{V}(\widehat{AUC})^{1/2}}. \tag{11}$$

Because this variance presumes independent sampling, we additionally report a percentile interval from the stationary bootstrap with mean block length 10.

Pairwise loss comparison. With $\bar{d}$ the sample mean of $d_{k,t}^{DM}$, $\widehat{V}_{NW}$ the Newey–West long-run variance estimator [24] with Bartlett kernel and bandwidth $h-1$, and h the forecast horizon, the corrected statistic [10, 16] is

$$DM = \frac{\bar{d}}{\left\{\widehat{V}_{NW}/n\right\}^{1/2}} \cdot \left\{\frac{n+1-2h+h(h-1)/n}{n}\right\}^{1/2} \sim t_{n-1}. \tag{12}$$

Here $\widehat{V}_{NW} = \widehat{\gamma}_0 + 2\sum_{j=1}^{\ell}\left(1 - \frac{j}{\ell+1}\right)\widehat{\gamma}_j$ is the Bartlett-kernel estimate of the long-run variance of the loss differential $d_{k,t}^{DM}$, with $\widehat{\gamma}_j$ its sample autocovariance at lag j and bandwidth $\ell = h-1$. Under operator A, $h = 1$ and $\ell = 0$, so $\widehat{V}_{NW}$ reduces to the sample variance of the loss differential. This choice is valid only if loss differentials are serially uncorrelated, which persistent predictors may violate, and a zero bandwidth then understates the variance. We therefore treat the DM statistics as supplementary evidence; the stationary-bootstrap procedures below, which do not rely on this assumption, form the primary basis for inference.

Composite nulls. Let $K = 5$ be the number of competing classifiers. White's Reality Check [30] tests $H_0^{RC} : \max_k \mathbb{E}[d_k] \le 0$ through

$$V = \sqrt{n}\max_{1\le k\le K}\bar{d}_k, \tag{13}$$

with the bootstrap distribution of $\sqrt{n}(\bar{d}_k^* - \bar{d}_k)$. Hansen's studentised statistic with consistent recentring [14] is

$$T^{SPA} = \max\left(0,\ \max_{1\le k\le K} \frac{\sqrt{n}\,\bar{d}_k}{\widehat{\omega}_k}\right), \qquad \widehat{\mu}_k = \bar{d}_k \cdot \mathbf{1}\left\{\bar{d}_k \le -\frac{\widehat{\omega}_k}{\sqrt{n}}\sqrt{2\ln\ln n}\right\}, \quad (14)$$

where $\widehat{\omega}_k^2$ is the bootstrap variance of $\sqrt{n}\,\bar{d}_k$. Both procedures use 2,000 stationary-bootstrap replications with mean block length 10.

Model Confidence Set. Let $\mathcal{M}$ be the current set of m models, including the benchmark, and $\bar{L}_k$ the sample mean loss of model k. Define $\bar{d}_{k\cdot} = \bar{L}_k - m^{-1}\sum_{l\in\mathcal{M}} \bar{L}_l$, the loss of model k relative to the average over the set. The MCS [15] is constructed by iterative elimination based on

$$T_{\max} = \max_{k\in\mathcal{M}} \frac{\bar{d}_{k\cdot}}{\widehat{\mathrm{se}}(\bar{d}_{k\cdot})}, \quad (15)$$

with bootstrap critical values. At each step the model attaining $T_{\max}$ is eliminated if the equivalence hypothesis is rejected. Each model receives an MCS p-value, and model k belongs to the estimated MCS at confidence level $1-\alpha$ if and only if its MCS p-value is at least α. We use $1-\alpha = 0.90$.

Properties and limitations of the battery. Five points bear on interpretation.

First, the benchmark is nested within every classifier: a constant probability forecast is the special case in which all slope coefficients vanish or every tree collapses to its root. Under nesting the Diebold–Mariano statistic has a non-standard limiting distribution, and under H_0 the estimated larger model is expected to have a higher out-of-sample loss than the benchmark because it fits parameters whose population values are zero [7]. Positive DM statistics are therefore what H_0 predicts; they indicate the estimation cost of the classifiers, not that the predictors carry information with the wrong sign. Conversely, the DM test is conservative against the alternative that the larger model is better, which works against rejection of H_0.

Second, White's Reality Check is conservative when the candidate set contains models that are far from the benchmark, because the maximum in the statistic is taken over all K models including irrelevant ones; Hansen's recentring removes such models from the null distribution, which is why we report both.

Third, an MCS p-value is not the probability that a model is best. A value of 1 means only that the model was never eliminated, and the order in which retained models would be eliminated is not itself an inferential result: two models that are both in the MCS are statistically indistinguishable at the chosen level.

Fourth, the stationary bootstrap is preferred to the moving-block bootstrap because the resampled series is stationary by construction, which matters when loss differentials are computed from overlapping labels. The mean block length of 10 was fixed a priori at the maximal label span under operator C; the conclusions are unchanged for mean block lengths between 5 and 20.

Fifth, the statistics differ in their robustness to dependence. The Pesaran–Timmermann and DeLong variances are derived under independence. Under operator A, labels do not overlap, but persistent predictors can still induce serial correlation in hits. Under operators B and C, labels overlap and these variances are understated. We therefore treat the stationary-bootstrap results (AUC intervals, Reality Check, SPA and MCS) as the primary basis for inference, the DM statistic as supplementary because its bandwidth is zero under operator A, and independence-based statistics as descriptive whenever labels overlap.

Decomposition of the Brier score. To separate calibration from discrimination we use the decomposition of Murphy [23]. Partitioning the forecasts into J bins with n_j observations, mean forecast $\bar{p}_j$ and realised frequency $\bar{o}_j$, and writing $\bar{o}$ for the unconditional frequency,

$$\overline{BS} \approx \underbrace{\frac{1}{n}\sum_{j=1}^{J} n_j\big(\bar{p}_j - \bar{o}_j\big)^2}_{\text{reliability}} - \underbrace{\frac{1}{n}\sum_{j=1}^{J} n_j\big(\bar{o}_j - \bar{o}\big)^2}_{\text{resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{uncertainty}}. \tag{16}$$

The identity is exact when forecasts are constant within bins; with binned continuous forecasts a small within-bin variance term remains. Reliability measures miscalibration and is minimised at zero; resolution measures the ability to separate observations into groups with different outcome frequencies and is the component that reflects discrimination; uncertainty depends only on the sample and is common to all forecasters. A model can therefore attain a low Brier score by predicting close to the base rate, and the decomposition distinguishes that case from genuine discrimination. We use $J = 10$ equiprobable bins.

4.6. Economic evaluation

The position is $w_t = 2 \cdot \mathbf{1}\{\widehat{p}_t > 1/2\} - 1$ under operators A and B. Under operator C it is the concurrency-normalised average of open positions, $w_t = \kappa_t^{-1} \sum_{i:\, i \le t \le \theta_i} \operatorname{sign}(\widehat{p}_i - 1/2)$. Net returns are

$$\Pi_t = w_t\, r_{t+1} - \gamma\, |w_t - w_{t-1}|, \qquad \gamma = 5 \text{ basis points}. \tag{17}$$

We report the annualised mean, standard deviation, Sharpe ratio, maximum drawdown and Calmar ratio. The passive long position delivered an annualised

mean return of 3.20% at 38.89% volatility (Sharpe ratio 0.082) over the evaluation window.

Equation (17) values positions at the DCE closing price on which the signal is conditioned. Because the signal uses information fixed at the 15:00 close, that price is not attainable in practice: the returns in (17) are signal returns, which measure the economic content of the forecasts, rather than implementable trading returns. An implementable backtest would fill at the next available price (the night session or the next day-session open) and account for slippage and contract rolls; it would change realised returns but not the statistical tests of Section 4.5. The Sharpe ratios and break-even costs reported below are interpreted in this sense.

Sharpe ratios are reported as point estimates. Under independently and identically distributed returns, the standard error of an annualised Sharpe ratio SR estimated over Y years is approximately $\{(1 + SR^2/2)/Y\}^{1/2}$ [21]; with $Y \approx 6.3$ this is between 0.40 and 0.44 for every Sharpe ratio reported in this paper. Serial correlation modifies the approximation, and a formal comparison of the Sharpe ratios of two strategies requires a paired HAC or bootstrap test [19].

4.7. Minimum detectable effects and what the data exclude

A non-rejection is informative only in relation to the effect sizes the design could have detected. We report two complementary quantities: design-stage minimum detectable effects, and confidence limits computed from the realised data. Only the second supports statements about which effect sizes are inconsistent with the evidence.

Minimum detectable effects. Consider a classifier issuing balanced directional forecasts, $\widehat{P}_x \approx 1/2$, against the observed $\widehat{P}_y = 0.522$ under operator A. The independence-implied hit rate is then $\widehat{P}^* = 0.500$, with standard error $\{\widehat{V}(\widehat{P}) - \widehat{V}(\widehat{P}^*)\}^{1/2} = 0.0128$. For a single pre-specified two-sided test at the 5% level with power 0.8, the conventional minimum detectable effect is

$$\Delta_{\min} = (z_{0.975} + z_{0.80}) \cdot 0.0128 = 2.802 \times 0.0128 = 0.0358, \tag{18}$$

which corresponds to a hit rate of approximately 53.6% against the independence rate of 50.0%. The corresponding thresholds for the AUC are 0.542 with the DeLong standard error under operator A and 0.584 with the stationary-bootstrap standard error under operator C.

These figures are approximations with a limited purpose. They assume serial independence, which, as noted in Section 4.5, understates the variance under

overlapping labels. They refer to one pre-specified comparison: with five classifiers and a Bonferroni adjustment, the hit-rate threshold rises to about 54.4% and the operator A AUC threshold to about 0.551. They describe the design, not the realised effect, and power computed from observed effect sizes is uninformative [17]. Above all, a minimum detectable effect is not an exclusion bound: failing to reject H_0 does not show, at any stated confidence level, that effects larger than $\Delta_{\min}$ are absent. Such a statement requires a confidence limit or a formal equivalence test [27].

The distinction between nominal and effective sample size makes the dependence correction concrete. Under operator C the nominal evaluation sample is $n = 1{,}531$, but with the mean uniqueness of 0.135 reported in Section 4.2 the uniqueness-weighted sample $\sum u_i$ corresponds to only about $0.135 \times 1{,}531 \approx 207$ non-overlapping labels. The stationary-bootstrap standard error of the penalised-logit AUC under operator C (0.0298) is about twice its DeLong value (0.0147), a variance ratio of about 4.1, which corresponds to an effective sample for the AUC of roughly $1{,}531/4.1 \approx 370$. The two measures differ because uniqueness counts concurrent labels, whereas the bootstrap variance also reflects how strongly concurrent labels are correlated; both imply that the operator C sample carries the information of a few hundred independent observations rather than 1,531. Under operator A, where labels do not overlap, the bootstrap and DeLong standard errors are close (0.0142 and 0.0148) and the effective and nominal sample sizes approximately coincide. For this reason the operator C thresholds above are computed with the bootstrap standard error.

What the data exclude. The stationary-bootstrap 95% intervals in Tables 4 and 5 provide dependence-robust confidence limits. Under operator A, the upper limits for the classifiers' AUCs lie between 0.518 and 0.541; under operator C they reach 0.598 for the penalised logit. The evidence is therefore compatible with discrimination of up to about 0.54 under operator A and about 0.60 under operator C for the best classifier, and effects of that size can be neither detected nor excluded with this sample. In terms of hit rates under operator A, the largest Pesaran–Timmermann statistic (-0.009) implies an approximate, independence-based 95% upper limit for the excess hit rate $\widehat{P} - \widehat{P}^*$ of about 2.5 percentage points. Hit rates of 60% or more would correspond to excess hit rates far outside these limits.

5. Results

5.1. The daily-sign label under the adapted filtration

Table 4 reports the full battery under operator A and the adapted filtration.

Table 4: Operator A (daily sign), adapted filtration. Evaluation window 2 January 2020 – 29 April 2026, $n = 1{,}531$; unconditional positive-class frequency 0.5219

Model	Hit rate	S_{PT}	p_{PT}	AUC	DeLong z	p_{DeLong}	Bootstrap 95% CI for AUC	Brier score	DM vs benchmark	p_{DM}	MCS p-value
Penalised logit	0.5042	−0.574	0.566	0.5119	0.803	0.422	[0.4856, 0.5413]	0.3195	9.582	< 0.001	< 0.001
Random forest	0.5088	−0.093	0.926	0.4911	−0.602	0.547	[0.4670, 0.5182]	0.2519	0.745	0.456	0.454
XGBoost	0.4905	−1.073	0.284	0.4981	−0.127	0.899	[0.4746, 0.5253]	0.2611	3.508	0.001	< 0.001
LightGBM	0.5042	−0.009	0.993	0.4979	−0.142	0.887	[0.4738, 0.5260]	0.2674	4.585	< 0.001	< 0.001
Tree ensemble	0.5003	−0.414	0.679	0.4960	−0.269	0.788	[0.4724, 0.5235]	0.2578	2.794	0.005	< 0.001
Benchmark	0.5219	–	–	0.4915	−0.575	0.565	[0.4631, 0.5164]	0.2509	–	–	1.000

Note. All p-values are two-sided. S_{PT}: Pesaran–Timmermann statistic, equation (10). DeLong z: test of H_0: AUC $= 1/2$, equation (11). Bootstrap CI: stationary bootstrap, mean block length 10, 2,000 replications. DM: Diebold–Mariano statistic with the Harvey–Leybourne–Newbold correction, equation (12); positive values indicate higher Brier loss than the benchmark. MCS p-value: Model Confidence Set, equation (15); models with MCS p-value ≥ 0.10 belong to the 90% MCS. Composite nulls: White Reality Check $V = -0.039$, $p = 0.894$; Hansen SPA $T^{SPA} = 0.000$, $p = 0.736$.

No classifier attains the hit rate of the always-long rule (0.5219). The Pesaran–Timmermann statistic is negative for all five classifiers, with a minimum p-value of 0.284. No DeLong statistic is significant (minimum $p = 0.422$), and every stationary-bootstrap 95% interval for the AUC contains 1/2. Four of the five classifiers have significantly higher out-of-sample Brier loss than the benchmark (DM $p \leq 0.005$). As explained in Section 4.5, this is the pattern expected under H_0 when estimated models are compared with a nested constant forecast, and it measures the estimation cost of the classifiers rather than an inverse signal. Neither composite null is rejected (Reality Check $p = 0.894$; SPA $p = 0.736$). The SPA statistic is exactly zero, which means that no classifier has a positive studentised mean loss differential relative to the benchmark.

The 90% MCS contains the benchmark (MCS $p = 1.000$) and the random forest (MCS $p = 0.454$) and excludes the other four classifiers. The benchmark's MCS p-value of 1 indicates only that it was not eliminated; the procedure does not distinguish it from the random forest. The random forest is retained because its probability forecasts remain close to the base rate (Table 7), not because it discriminates: its AUC is 0.4911.

Taken together, the operator A results provide no statistically detectable evidence of directional predictability. They do not establish that predictability is absent: as shown in Section 4.7, AUC values up to about 0.54 remain compatible with the data.

Figure 4: Cumulative log returns of the long–short strategies under the adapted filtration, net of 5 basis points per unit of turnover (operator A); the dashed line is the passive long position. LOGIT: penalised logit; RF: random forest; XGB: XGBoost; LGBM: LightGBM; ENS_TREE: tree ensemble. All five strategies end the evaluation window below zero. The random forest accumulated a large gain in 2020–2021 that was subsequently reversed; the other strategies spent most of the window below zero. No strategy shows a persistent advantage over the passive position.

5.2. Other labelling operators

Under operator B (not tabulated), the qualitative conclusions are unchanged. The tree ensembles display nominally significant negative Pesaran–Timmermann statistics. Because operator B labels overlap over five days, the independence-based Pesaran–Timmermann variance is understated, so these nominal rejections are not reliable evidence of anti-predictability, although they are consistent with overfitting of training-sample variation.

Table 5 reports operator C.

Table 5: Operator C (triple barrier, $k = 0.75$, $T = 10$), adapted filtration. Evaluation window as in Table 4, $n = 1{,}531$; unconditional positive-class frequency 0.5421

Model	Hit rate	S_{PT}	p_{PT}	AUC	DeLong z	p_{DeLong}	Bootstrap 95% CI for AUC	Brier score	DM vs benchmark	p_{DM}	MCS p-value
Penalised logit	0.5434	1.686	0.092	0.5396	2.696	0.007	$[0.4815, 0.5984]$	0.3007	3.790	< 0.001	0.018
XGBoost	0.5369	1.305	0.192	0.4988	-0.083	0.934	$[0.4316, 0.5614]$	0.2669	2.306	0.021	0.026
LightGBM	0.5082	-0.966	0.334	0.4797	-1.370	0.171	$[0.4174, 0.5461]$	0.3097	4.420	< 0.001	0.001
Tree ensemble	0.4990	-1.965	0.050	0.4735	-1.793	0.073	$[0.4089, 0.5426]$	0.2748	3.019	0.003	0.018
Random forest	0.4847	-3.455	0.001	0.4392	-4.145	< 0.001	$[0.3756, 0.5118]$	0.2672	3.045	0.002	0.026
Benchmark	0.5421	–	–	0.5113	0.766	0.443	$[0.4553, 0.5638]$	0.2493	–	–	1.000

Note. Definitions as in Table 4. Under operator C labels overlap (mean uniqueness 0.135), so S_{PT} and the DeLong z, which assume independence, are descriptive; inference rests on the bootstrap interval, the DM statistic and the composite procedures (Section 4.5). Composite nulls: White Reality Check $V = -0.690$, $p = 0.998$; Hansen SPA $T^{SPA} = 0.000$, $p = 0.752$.

The penalised logit attains a hit rate of 0.5434 against an unconditional positive-class frequency of 0.5421 and an AUC of 0.5396, the largest in the study. The DeLong statistic rejects H_0: AUC $= 1/2$ at the 1% level ($z = 2.696$, $p = 0.007$). This rejection does not survive relaxation of the independence assumption: the stationary-bootstrap interval $[0.4815, 0.5984]$ contains $1/2$. The bootstrap standard error (0.0298) is about twice the DeLong value (0.0147), so the DeLong variance is understated by a factor of about four; with the bootstrap standard error the statistic becomes $z = 1.33$ ($p \approx 0.18$). Under operator A, where labels do not overlap, the two standard errors are close (0.0142 against 0.0148). This comparison is consistent with label overlap being an important source of the understatement. Operators A and C, however, differ in more than the degree of overlap — in label horizon, class balance and the persistence of the label sequence — so the present design does not separately identify the contribution of overlap and of other dependence mechanisms.

The 90% MCS contains the benchmark alone: every classifier has an MCS p-value of 0.026 or less. The random forest has a hit rate of 0.4847, an AUC of

0.4392 and $S_{PT} = -3.455$ ($p = 0.001$), and its DeLong statistic is -4.145. Both statistics are derived under independence; the bootstrap interval $[0.3756, 0.5118]$ contains 1/2, so the evidence that the random forest performs worse than an uninformative rule is not robust to dependence. The same applies to the tree ensemble ($S_{PT} = -1.965$, $p = 0.050$). Below-1/2 point estimates of this kind are consistent with a flexible estimator fitting training-sample structure that does not persist out of sample, but the design cannot distinguish that explanation from sampling variation.

5.3. The cost of filtration misspecification

Table 6 contrasts the adapted and the misspecified filtrations. The two runs differ only in the timestamp rule: under the misspecified filtration, global daily variables enter with $\ell = 0$ and monthly releases are indexed by reference period rather than release date. The predictor set, re-estimation scheme, hyperparameter grids, random seed and transaction costs are identical.

Table 6: Adapted versus misspecified filtration, operator A

	Hit rate		AUC		DeLong p	Sharpe ratio, net of 5 bp	
Model	Adapted	Misspecified	Adapted	Misspecified	Misspecified	Adapted	Misspecified
Penalised logit	0.5042	0.5121	0.5119	0.5269	0.069	−0.06	0.66
Random forest	0.5088	0.5232	0.4911	0.5168	0.258	−0.08	0.39
XGBoost	0.4905	0.5147	0.4981	0.5143	0.333	−0.18	0.12
LightGBM	0.5042	0.5003	0.4979	0.5149	0.313	−0.14	−0.15
Tree ensemble	0.5003	0.5127	0.4960	0.5156	0.293	−0.17	−0.01

Note. Misspecified filtration: global daily variables enter with $\ell = 0$ and monthly releases are indexed by reference period. DeLong p-values under the adapted filtration are reported in Table 4. Approximate standard error of each annualised Sharpe ratio: 0.40–0.44 (Section 4.6). Composite nulls under misspecification: White Reality Check $p = 0.721$; Hansen SPA $p = 0.672$. 90% Model Confidence Set under misspecification: random forest (MCS $p = 1.000$) and benchmark (MCS $p = 0.676$).

Statistical measures. The hit rate changes by between -0.4 and $+2.4$ percentage points (it rises for four of the five classifiers), and the AUC rises by between 0.015 and 0.026 for all five. The Pesaran–Timmermann test remains insignificant in every case (minimum $p = 0.243$); the DeLong test rejects at the 10% level in one case out of five (penalised logit, $p = 0.069$); and neither composite null is rejected (Reality Check $p = 0.721$; SPA $p = 0.672$).

Figure 5: AUC by classifier under three specifications: daily-sign label with the adapted filtration ("Daily return sign"), triple-barrier label with the adapted filtration ("Triple-barrier labelling") and daily-sign label with the misspecified filtration ("Naive alignment"). The dashed line marks the value $1/2$ implied by H_0. Under the daily-sign label and the adapted filtration all estimates lie within 0.012 of $1/2$, four of the five marginally below it. Triple-barrier labelling raises the AUC point estimate only for the penalised logit and lowers it for the tree-based models. Classifier abbreviations as in Figure 4.

Figure 6: Hit rate on a rolling 252-observation window (operator A, adapted filtration) against the contemporaneous unconditional positive-class frequency (dashed). Classifier abbreviations as in Figure 4. The trajectories fluctuate within approximately $[0.44, 0.58]$. Periods in which a classifier exceeds the base rate, such as LightGBM from mid-2024 to early 2025 when the base rate itself fell to about 0.50, are not sustained across years and are not synchronised across classifiers. This pattern is consistent with sampling variation around a weak or zero conditional signal.

Model Confidence Set. Under the misspecified filtration the 90% MCS again contains both the random forest (MCS $p = 1.000$) and the benchmark (MCS $p = 0.676$). Because the benchmark is not eliminated, the MCS provides no evidence that the random forest outperforms it. Relative to Table 4, what changes is only which of the two retained models has the smaller sample loss, and hence the order in which they would be eliminated; their joint membership of the MCS is unchanged. As noted in Section 4.5, the ordering of models within an MCS is not an inferential result, so the MCS conclusions are the same under both filtrations.

Economic measures. The realised Sharpe ratio changes sign for three of the five classifiers: from -0.06 to 0.66 for the penalised logit, from -0.08 to 0.39 for the random forest and from -0.18 to 0.12 for XGBoost. Because the two runs share data, seed and hyperparameter grids, these differences are attributable to the timestamp rule within this pipeline. They are not, however, statistically established. With approximately 6.3 years of evaluation data the standard error of an annualised Sharpe ratio is about 0.4 (Section 4.6), so none of the individual Sharpe ratios in Table 6 differs significantly from zero at the 5% level; the largest, 0.66, has an approximate t-statistic of 1.5. A paired test of the Sharpe differences [19], which would exploit the strong correlation between the positions of the two runs, would be needed to establish that the misspecified rule earns a higher population Sharpe ratio.

A possible mechanism. One explanation consistent with these patterns is that misspecification matters mainly on dates on which the overnight move of US markets carries information about the following Chinese session. If such dates are relatively few but concentrated in the upper tail of the absolute-return distribution, the hit rate would change little while the return-weighted contribution could be large. We have not tested this explanation directly. A direct check would compare the misspecified-minus-adapted difference in hit rates across quantiles of $|r_{t+1}|$.

Implication. In this experiment the economic measure moved far more, relative to its usual magnitude, than the test statistics did, yet that movement was not itself statistically significant. A study that reports only economic performance therefore cannot distinguish a filtration error from genuine predictability, while a study that reports only predictive-ability statistics may not reveal a filtration error at all. We therefore recommend reporting both, together with an explicit information-availability protocol such as that in Section 3.3.

Figure 7: Cumulative log returns under the two filtrations for the penalised logit (panel a) and the random forest (panel b), operator A, net of 5 basis points. Blue ("Strict alignment"): adapted filtration; red ("Naive alignment"): misspecified filtration. The paths do not diverge monotonically — for the random forest the adapted run is ahead during 2021–2022 — but by the end of the window the cumulative log-return gap is about 1.7 for the penalised logit and about 1.1 for the random forest.

5.4. Decomposition of the Brier score

Table 7 applies the decomposition (16) and resolves an apparent tension in Table 4: the random forest attains the lowest Brier score among the classifiers although its AUC lies below one half.

Three observations follow. First, resolution, the component that reflects discrimination, never exceeds 0.0064 and is typically below 0.0020, against an uncertainty of approximately 0.249. The share of outcome variance that the classifiers resolve is thus below 0.6% under operator A and below 2.6% under operator C. Second, under operator A the benchmark's resolution (0.0023) exceeds that of every classifier, which is the decomposition's expression of the absence of detectable discrimination: the drift of the expanding base rate tracks the realised frequency marginally better than any fitted model. Third, differences in Brier scores across classifiers are driven mainly by reliability, that is, by calibration rather than information. The penalised logit issues dispersed forecasts, $sd(\widehat{p}) = 0.258$, and incurs a reliability penalty of 0.0704; the random forest issues forecasts concentrated near the base rate, $sd(\widehat{p}) = 0.045$, and incurs only 0.0037. The random forest is a more cautious forecaster rather than a more informative one, and its retention in the MCS of Table 4 should not be read as evidence of skill.

Table 7: Murphy decomposition of the Brier score, adapted filtration

	Operator A (daily sign)				Operator C (triple barrier)			
Model	Brier score	Reliability	Resolution	$sd(\widehat{p})$	Brier score	Reliability	Resolution	$sd(\widehat{p})$
Penalised logit	0.3195	0.0704	0.0014	0.258	0.3007	0.0520	0.0033	0.221
Random forest	0.2519	0.0037	0.0015	0.045	0.2672	0.0243	0.0064	0.091
XGBoost	0.2611	0.0117	0.0008	0.109	0.2669	0.0239	0.0058	0.126
LightGBM	0.2674	0.0193	0.0015	0.137	0.3097	0.0616	0.0014	0.214
Tree ensemble	0.2578	0.0092	0.0007	0.092	0.2748	0.0285	0.0026	0.133
Benchmark	0.2509	0.0031	0.0023	0.023	0.2493	0.0022	0.0013	0.040
Uncertainty (all models)		0.2495				0.2482		

Note. Components from equation (16) with $J = 10$ equiprobable bins. $sd(\widehat{p})$: standard deviation of the probability forecasts over the evaluation window. Because forecasts vary within bins, Brier score $\approx$ reliability $-$ resolution $+$ uncertainty holds up to a within-bin term of at most 0.004.

5.5. Stability across sub-periods and volatility regimes

Table 8 reports hit rates by calendar year. Under operator A no classifier exceeds the contemporaneous frequency of upward moves in more than three of

the seven sub-periods, and the ranking of classifiers is unstable: the penalised logit leads in 2020, the random forest in 2021 and 2023, and LightGBM in 2024. Year-to-year dispersion of roughly ±4 percentage points is of the magnitude implied by binomial sampling variation alone for sub-samples of about 243 observations (standard error ≈ 0.032), so the data do not support year-specific conclusions.

Table 8: Hit rate by calendar year, operator A, adapted filtration

Year	n	Frequency of $y_t = 1$ (hit rate of the always-long rule)	Hit rate of classifier				
			Penalised logit	Random forest	XGBoost	LightGBM	Tree ensemble
2020	243	0.556	0.551	0.531	0.502	0.502	0.490
2021	243	0.523	0.486	0.547	0.473	0.498	0.510
2022	242	0.541	0.492	0.471	0.483	0.488	0.479
2023	242	0.558	0.533	0.545	0.492	0.508	0.529
2024	242	0.483	0.479	0.496	0.504	0.541	0.508
2025	243	0.481	0.498	0.481	0.486	0.494	0.490
2026	76	0.487	0.461	0.447	0.500	0.487	0.487
2020–2026	1,531	0.522	0.504	0.509	0.491	0.504	0.500

Note. 2026 covers 2 January – 29 April. The binomial standard error of an annual hit rate is approximately 0.032 for $n \approx 243$ and 0.057 for $n = 76$. The last row reproduces the full-sample values of Table 4.

Table 9 partitions the evaluation window into terciles of trailing 21-day realised volatility. Volatility is a natural candidate for state dependence: if predictability were concentrated in turbulent states, an unconditional test would dilute it. Because the tercile boundaries are computed over the evaluation window as a whole, the partition is a descriptive device rather than a conditioning rule that could be used in real time.

Within a tercile ($n \approx 510$) the independence-based standard error of an AUC is roughly $\sqrt{3}$ times its full-sample value, about 0.026 under operator A, which implies a minimum detectable AUC of about 0.57; under operator C, with the bootstrap standard error, the corresponding value is about 0.64. No regime-specific estimate reaches these thresholds. The largest value, 0.595 for the penalised logit in the middle tercile under operator C, exceeds the full-sample threshold of Section 4.7 but not the tercile threshold, and it is not accompanied by elevated values in the adjacent terciles.

No adjustment for the eighteen cells of Table 9 has been made; a Bonferroni adjustment for eighteen comparisons would raise the tercile thresholds to about 0.60 under operator A and 0.69 under operator C. The random forest under

Table 9: AUC by volatility regime, adapted filtration

Regime (tercile of trailing 21-day realised volatility)	n	Operator A (daily sign)			Operator C (triple barrier)		
		Penalised logit	Random forest	Tree ensemble	Penalised logit	Random forest	Tree ensemble
Low volatility	510	0.534	0.484	0.483	0.504	0.452	0.463
Middle volatility	510	0.510	0.484	0.498	0.595	0.426	0.443
High volatility	511	0.494	0.510	0.509	0.497	0.436	0.488

Note. Entries are AUCs within each tercile. Realised volatility at origin t uses returns up to and including day t; tercile boundaries are computed over the full evaluation window. Approximate minimum detectable AUC within a tercile (5% level, power 0.8): 0.57 under operator A (DeLong standard error) and 0.64 under operator C (bootstrap standard error). Because the tercile boundaries use the whole evaluation window, the regime classification is ex post: it is not available at the forecast origin, serves only to describe where in the sample the forecasts performed better or worse, and enters no forecast or trading rule.

operator C is below one half in all three terciles (0.452, 0.426 and 0.436), which suggests that its below-1/2 full-sample AUC is not driven by a single volatility state.

5.6. Turnover and break-even costs

The economic results in Sections 5.1 and 5.3 were computed at 5 basis points. Table 10 reports the cost at which each strategy's gross return is exhausted, which separates the question of whether there is a signal from the cost assumption.

Table 10: Turnover and break-even transaction costs, operator A, long–short strategies

Model	Annualised turnover ($\sum \lvert w_t - w_{t-1} \rvert$ per year)	Annualised gross return	Break-even cost (bp per unit of turnover)	Number of position changes
Penalised logit	39.0	−0.4%	−1.07	119
Random forest	92.0	+1.6%	+1.73	280
XGBoost	131.5	−0.5%	−0.42	400
LightGBM	147.3	+1.9%	+1.32	448
Tree ensemble	126.2	−0.3%	−0.21	384
Benchmark	0.2	+3.0%	1,834.7	1

Note. Break-even cost = annualised gross return / annualised turnover; a negative value means that the gross return is already negative. Evaluation window as in Table 4.

Break-even costs range from −1.07 to +1.73 basis points. Three of the five classifiers — the penalised logit, XGBoost and the tree ensemble — have negative gross returns before any cost is charged, and the gross return of the best (the

random forest) is exhausted at a cost of less than two basis points, about a third of the assumed level and below any realistic execution cost on the contract. The evidence therefore does not indicate that transaction costs eliminate an otherwise exploitable signal; rather, none of the gross returns is significantly different from zero. Annualised gross returns lie between -0.5% and $+1.9\%$ against an unconditional volatility of 38.9%; over about 6.3 years even the largest corresponds to an approximate t-statistic of 0.12 (under an i.i.d. approximation; with serially correlated strategy returns a HAC or bootstrap standard error would be required, but a statistic this small is far from any conventional critical value). The benchmark, which changes position once over the evaluation window, is not a trading strategy in any meaningful sense, and its break-even cost is reported only for completeness.

5.7. Predictor importance

Normalised split-gain importance aggregated across the tree ensembles is distributed as follows: Chinese macroeconomic releases 24.0%, global financial variables 22.4%, own-price technical variables 16.7%, steel complex 15.5%, order-flow variables 9.3%, Chinese equities 6.1%, news-based indices 4.1% and calendar variables 2.0% (shares sum to 100% up to rounding).

Because the estimators generating these quantities show no detectable out-of-sample discrimination, the importance shares should not be given an economic interpretation; they describe how the fitted models used the training data. One observation is nonetheless informative: the order-flow block comprises 17 of the 172 predictors (9.9% of the set) and receives 9.3% of total importance, close to its proportional share, which is what would be expected if the block carried no more signal than the average predictor. Augmenting the predictor set with adjacent-market volume and open interest did not improve any statistic in the battery (results not tabulated).

6. Limitations

Several limitations qualify the results. First, the design is pseudo-real-time: data are taken in their May 2026 vintage, and some monthly release dates are approximate (Section 3.3). Second, the economic evaluation assumes execution at the DCE closing price on which the signal is conditioned; an implementable strategy would trade in the night session or at the next open, which would change realised returns but not the statistical tests. Third, the analysis concerns a single main continuous contract, whose returns may contain discontinuities at roll dates. Fourth, the hyperparameter grids are deliberately small and results are reported

for a single random seed. Fifth, the composite tests are based on Brier loss; tests based on directional loss could rank models differently. Sixth, and most importantly, the sample has limited power for small effects: discrimination of up to about 0.54 (operator A) and 0.60 (operator C) in AUC terms cannot be excluded.

7. Conclusion

We have proposed a filtration-consistent design for statistical learning of directional forecasts, in which each predictor enters only after its documented availability timestamp, and combined it with dependence-aware inference. Proposition 1 shows that, under this design, the null hypothesis of no directional predictability implies a population AUC of one half for every admissible forecast, whereas forecasts built on calendar-aligned predictors observed after the forecast origin carry no such guarantee. We applied the design to DCE iron ore futures, using a penalised logit and three tree ensembles with 172 predictors, three labelling operators and a 2020–2026 evaluation window.

Under the daily-sign label, no element of the test battery rejects the null hypothesis of no directional predictability: the best hit rate (0.5088) is below the unconditional frequency of upward moves (0.5219); no AUC differs significantly from 1/2; White's Reality Check and Hansen's SPA test give p-values of 0.894 and 0.736; and the 90% Model Confidence Set contains the unconditional benchmark together with the random forest, whose forecasts stay close to the base rate. Under the triple-barrier label, nominal rejections arise only from independence-based statistics and do not survive dependence-robust bootstrap inference; under the weekly-sign label the qualitative conclusions are unchanged. These findings indicate that any directional predictability is too small to be detected with this sample; they do not establish its absence. Bootstrap confidence limits leave room for AUC values of up to about 0.54 under the daily-sign label and 0.60 under the triple-barrier label.

A controlled experiment that changes only the timestamp rule raises the realised Sharpe ratio from -0.06 to 0.66 for the penalised logit and from -0.08 to 0.39 for the random forest, while the conclusions of all formal tests, including the Model Confidence Set, remain unchanged. With a standard error of about 0.4, these Sharpe ratios are themselves imprecise. Within this experiment, filtration misspecification therefore moved economic performance measures substantially without producing statistically significant evidence of predictability. Economic backtests and formal tests of predictive ability carry complementary information about specification error, and we recommend that studies in this area report both, together with an explicit information-availability protocol.

Our results suggest that, for directional forecasts of a single commodity contract, the rigour of the evaluation protocol matters at least as much as the flexibility of the estimator. Features of the conditional return distribution other than its sign, such as volatility, may be more promising objects of study.

References

[1] Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society, 61*(5), 458–471.

[2] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics, 131*(4), 1593–1636.

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

[4] Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review, 112*(4), 1194–1225.

[5] Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies, 21*(4), 1509–1531.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery.

[7] Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*(1), 291–311.

[8] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845.

[9] Deng, S., Zhu, Y., Huang, X., Duan, S., & Fu, Z. (2022). High-frequency direction forecasting of the futures market using a machine-learning-based method. *Future Internet, 14*(6), Article 180.

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

[11] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417.

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

[13] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies, 33*(5), 2223–2273.

[14] Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics, 23*(4), 365–380.

[15] Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica, 79*(2), 453–497.

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

[17] Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19–24.

[18] 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. In *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154).

[19] Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance, 15*(5), 850–859.

[20] Liu, J., Papailias, F., & Quinn, B. (2021). Direction-of-change forecasting in commodity futures markets. *International Review of Financial Analysis, 74*, Article 101677.

[21] Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal, 58*(4), 36–52.

[22] López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.

[23] Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595–600.

[24] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708.

[25] Pesaran, M. H., & Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics, 10*(4), 561–565.

[26] Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association, 89*(428), 1303–1313.

[27] Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680.

[28] Waldow, F., Schnaubelt, M., Krauss, C., & Fischer, T. (2021). Machine learning in futures markets. *Journal of Risk and Financial Management, 14*(3), Article 119.

[29] Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies, 21*(4), 1455–1508.

[30] White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), 1097–1126.

Sadig H. Gurbanov
Baku Engineering University, AZ0101, Baku, Azerbaijan
E-mail: sadiq.kamal1310@gmail.com
ORCID: 0009-0003-1427-4408

Svetlana V. Frumina
Plekhanov Russian University of Economics, 115054, Moscow, Russia;
Financial University under the Government of the Russian Federation, 125167, Moscow, Russia
E-mail: frumina@mail.ru

Zahid F. Mamedov
Azerbaijan State University of Economics (UNEC), AZ1001, Baku, Azerbaijan;
E-mail: z.mamedov@unec.edu.az

Lev M. Zotov
Financial University under the Government of the Russian Federation, 125167, Moscow, Russia
E-mail: levzotov2005@mail.ru

Alex I. Borodin
Plekhanov Russian University of Economics, 115054, Moscow, Russia;
Azerbaijan State University of Economics (UNEC), AZ1001, Baku, Azerbaijan
E-mail: aib-2004@yandex.ru

Oksana M. Kulikova
Vladivostok State University, 690014, Vladivostok, Russia
E-mail: ya.aaaaa11@yandex.ru

Received 24 March 2026
Accepted 20 October 2026