Artificial Intelligence Review

https://doi.org/10.1007/s10462-026-11664-w

Article in Press

QFRS: quantitative finance reporting standards for forecasting, evaluation and trading claims

Received: 29 March 2026

Accepted: 24 July 2026

Published online: 08 August 2026

Cite this article as: Khushi M. QFRS: quantitative finance reporting standards for forecasting, evaluation and trading claims. *Artif Intell Rev* (2026). https://doi.org/10.1007/s10462-026-11664-w

Matloob Khushi

QFRS: Quantitative Finance Reporting Standards for Forecasting, Evaluation and Trading Claims

Matloob Khushi[1]*

[1]*Department of Computer Science, Brunel University of London, Uxbridge, London UK

matloob.khushi@brunel.ac.uk

Abstract. Financial time-series forecasting lies between AI and market microstructure, but most studies optimise generic error metrics instead of risk-adjusted economic value under realistic frictions. Unlike NLP and vision, the field lacks a shared, reviewer-enforced standard for data handling and evaluation, leading to persistent problems such as data leakage, backtest overfitting and metric-chasing on RMSE/MAE.

This paper introduces QFRS a novel, enforceable by reviewers and editors, seven-standard framework and checklist for evaluating and reporting financial asset forecasting and trading claims. QFRS covers quantitative studies on equities (stocks), forex, cryptocurrencies, rates, derivatives (futures, forwards, options, swaps), energy prices, and commodities (gold, oil and silver) and any other asset class. The 7 standards specify an end-to-end experimental pipeline, covering i) dataset construction, ii) labelling, iii) point-in-time feature engineering, iv) leakage-free scaling or normalisation, v) time-respecting data splits, vi) evaluation metrics and vii) cost and slippage-aware backtesting with explicit execution assumptions and decision rules mapping predictions to positions.

To validate the standard's diagnostic value, a compliance audit of Scopus-indexed forex forecasting papers published in 2025 is presented. None of these papers achieved full compliance across all seven standards, with economic backtesting (12.2%) and causal scaling (31.7%) recording the lowest pass rates. QFRS underpins a public state-of-the-art leaderboard, ensuring that only studies satisfying these standards are ranked, with the goal of shifting the literature from opaque, error-metric-driven results to transparent, economically meaningful and comparable benchmarks. The accompanying leaderboard is available and updated regularly at http://mkhushi.github.io

Keywords: QFRS Standards, AI for FinTech, Stock price prediction, cryptocurrency price prediction, Financial asset price forecast

1 Introduction

Machine learning in language and vision has matured by rigorously defined benchmarks such as GLUE and ImageNet where shared datasets, standardised metrics and clearly specified evaluation standards make progress both cumulative and auditable (Wang et al., 2018; Deng et al., 2009). In financial asset forecasting, by contrast, there is still no widely adopted, reviewer-enforceable standard for how data should be constructed, how information timing should be respected, or how models should be evaluated. There is extensive evidence that practices like random or K-fold splitting, full-sample preprocessing and the use of revised rather than real-time data can systematically leak future information and inflate reported performance (Bailey and López de Prado, 2014; Croushore and Stark, 2001; Kapoor and Narayanan, 2023; Yang et al., 2024). At the same time, much of the literature continues to equate “progress” with marginal reductions in error metrics such as RMSE or MAE, even though studies in empirical asset pricing and trading show that such improvements often fail to translate into robust, risk-adjusted economic gains once transaction costs, slippage and market frictions are accounted for (Gu et al., 2020; Hu, 2025; Poon and Granger, 2003).

These issues are not minor details. AI finance researchers often search over many models, feature sets and horizons. This increases the risk of discoveries that work only in the sample where they were found, a phenomenon known as backtest overfitting (Bailey and López de Prado, 2014; Bailey and López de Prado, 2021). Even when authors report out-of-sample results, using random splits or leaking future information through preprocessing can make weak signals look convincing on paper but fail in live trading. At the same time, most work still treats progress as a reduction in forecast error, for example lower RMSE or MAE. In

practice, the relevant objective is net, risk-adjusted profit and loss after realistic costs, slippage and constraints. Studies have underlined that better error scores do not guarantee profitable trading once these frictions are included (Hu, 2025). To empirically demonstrate the scale of this problem, this paper presents a systematic compliance audit of 41 Scopus-indexed forex AI forecasting papers published in 2025.

In response, this paper introduces the Quantitative Finance Reporting Standards (QFRS) framework, which is decomposed into seven numbered standards, denoted QFRS-1 to QFRS-7. For brevity, these are also referred to as S1 to S7 when the context is unambiguous. Each standard comprises one or more criteria (also referred to as sub-criteria), which are the individual, binary-coded checklist items against which a study is assessed. The word standard refers exclusively to one of the seven QFRS-k items; criterion or sub-criterion refers to an assessable item within a standard. The framework defines standards for evaluation and publishing financial asset forecasting and trading claims, applicable across equities, forex, cryptocurrencies, rates, derivatives, energy prices, and commodities. QFRS framework defines a single end-to-end pipeline covering data handling, information timing, feature engineering, scaling and normalisation, train/validation/test splits, model selection, and backtesting with transaction costs and slippage. Any study wishing to appear on the accompanying public leaderboard must satisfy this pipeline in full.

The contributions of this paper are fivefold. First, a full standard specification is provided for financial forecasting experiments, covering dataset construction, label alignment, feature design, and time-series splitting. Second, a set of standardised reporting tables is defined that every submission must complete, covering data periods, split dates, cost and slippage parameters, execution assumptions, and the mapping from predictions to positions. Third, a reference backtest setup is described so that profit and risk figures are computed consistently and always net of stated frictions. Fourth, leaderboard metrics and ranking rules are specified, based on risk-adjusted performance and drawdown behaviour rather than forecast error alone. Fifth, an empirical compliance audit of 41 recently published papers grounds the standard's design in documented real-world failures rather than theoretical concerns, providing direct evidence of the gap this standard is designed to close.

1.1 Scope and relationship to non-AI methodologies

QFRS was originally motivated by the failure modes arising in machine learning pipelines applied to financial time-series, but its core requirements are logically independent of the modelling paradigm. Nevertheless, the underlying data-quality requirements (QFRS-1), label-definition (QFRS-2), feature engineering (QFRS-3), evaluation metric (QFRS-6) and economic backtesting criteria (QFRS-7) are logically independent of the modelling paradigm. A study employing traditional statistical methods such as technical rule-based trading systems or econometric models still produces forecasts from financial time-series data and still claims trading value; the same data-leakage risks and the same need for cost-aware evaluation apply. Accordingly, QFRS is applicable, in principle, to any empirical study that (i) uses financial time-series data, (ii) constructs a predictive model or rule-based system, and (iii) claims forecasting or trading performance. In such studies, scaling (QFRS-4) and the train/test split chronological-order requirement (QFRS-5) do not apply and are coded as NA, and the study can still be considered compliant. The primary enforcement context, however, remains AI/ML-based forecasting research, which is the domain exhibiting the highest prevalence of the leakage and evaluation failures documented in Section 10.

2 QFRS-1: Dataset specification and data handling

2.1 Asset universe and markets

Supported asset classes: Each study must clearly state which assets are included so that results are not over-generalised across markets with very different microstructure. Assets may include: single stocks, exchange traded funds (ETFs), forex, cryptocurrencies, rates, derivatives (futures, forwards, options, swaps), energy prices (futures and spot indices), and commodities (gold, oil and silver) (Ante et al., 2021).

Venues and listing rules: State all trading venues and data vendors used for each asset class. Describe any filters on liquidity, market capitalisation or history length. For equities, clarify whether the dataset includes delisted securities and how delisting returns are handled to avoid survivorship bias. Look-ahead and survivorship biases have

been shown to produce large overstatements of performance if delistings and benchmark composition changes are ignored (Malevergne and Sornette, 2006; Bailey and López de Prado, 2014).

Bar frequency and data sampling: Fix bar frequency per track, for example: tick, one minute, five minute, one hour, one day. High-frequency tracks (limit order book data) must state whether they use event time or fixed-length bars (Ntakaris et al., 2019; Ntakaris et al., 2018). The choice between event-time and clock-time sampling is central in high-frequency forecasting (Briola et al., 2025).

2.2 Corporate events and instrument-specific rules

Equities and exchange traded funds: Document how corporate actions are processed, including stock splits, dividends and rights issues. Explain whether prices are adjusted for these events and how total return series are constructed (Ince and Porter, 2006; Malevergne and Sornette, 2006).

Futures and other derivatives: Specify contract selection and roll rules any price adjustments at rolls (Gorton and Rouwenhorst, 2006). Back-adjustment methods are widely discussed because they change the structure of historical returns and can affect trend-following performance estimates (Hurst et al., 2013).

Cryptocurrencies and stablecoins: Clarify how stablecoins and pegged assets are treated, including episodes of de-pegging, for example TerraUSD collapse. State any filters on venues or trading pairs to reduce noise from illiquid markets (Nurbaev et al., 2023; Makarov and Schoar, 2020).

2.3 Time, synchronisation and time zones

Cross-venue synchronisation: Mixing time zones without care can create artificial lead–lag relationships, a point highlighted in work on international market linkages and high-frequency co-movement (Andersen et al., 2000). When combining assets from multiple venues, align trades and quotes on the common time base. Describe resampling rules if different sources have different native frequencies (Anderson et al., 2003).

External information alignment: If macroeconomic announcements, news, or other external signals are used in the predictive model, use the edit time if one is available, that reflect when information was released to the market or updated (Anderson et al., 2003). Ensure that no feature uses external information before its publication time, consistent with practices in news-aware trading agents (Benhenda, 2025).

2.4 Missing data, halts and outliers

Missing and Non-trading periods: Gaps in trading arise from regular schedules and special closures, and models need a consistent time axis. Describe how exchange holidays, weekends and regular non-trading hours are represented. State whether missing bars are explicitly inserted or whether the time axis skips non-trading intervals.

Trading halts and suspensions: Halts coincide with large jumps and unusual spreads, and are treated as special events in market microstructure research. Explain how trading halts and suspensions are recorded in the data.

Forward fill and interpolation rules: Carrying forward last values is common but can create artificial predictability if applied across long gaps. For example if Monday and Wednesday prices are available but Tuesday's price is missed then forward filling with Monday's price should be justified. Forbid forward filling across long gaps unless clearly justified (Brownlees and Gallo, 2006).

Error and outlier handling: Describe filters for removing bad ticks (if there are any), duplicated records and obvious price error (bid-ask ordering), trades far outside recent ranges or volume errors, in line with high-frequency datasets used in the literature (Brownlees and Gallo, 2006; Ntakaris et al., 2018). State any winsorisation or outlier capping rules applied to returns or features (Brownen-Trinh, 2019).

2.5 Data sources and documentation

Named data providers and versions: Identify each data source (for example, specific exchanges, consolidated feeds, academic databases) and its version or download date. For equity studies, mention whether centre for research in security prices (CRSP) or equivalent datasets are used and how survivorship bias is controlled (Bailey and López de Prado, 2014).

Symbol lists and sample periods: Provide complete symbol lists and date ranges for each asset class and track. Follow the example of published benchmark datasets that list instruments and time spans so that later work can reproduce the sample (Briola et al., 2025; Brownlees and Gallo, 2006). Work on look-ahead benchmark bias has shown that failing to handle index membership and delistings correctly can materially overstate performance (Bailey and López de Prado, 2014).

2.6 Decision and execution timeline

Time points in the trading loop: A clear timeline prevents accidental use of future information and defines realistic reaction speed. The last timestamp at which features are known, for example close of the current bar or most recent bid/ask quote. Define fill time: when trades are assumed to execute (for example, next bar open, next bar mid, or within a defined window).

Alignment of features, labels and trades: Ensure that all features are constructed only from information available at or before the observation time. Follow the level of detail seen in high frequency forecasting work, where the sequence of order book events and the outcome interval is specified explicitly.

2.7 Fundamentals, macro and news fields

Publication-time fields: For fundamental data (for example, earnings, balance sheet items) record the official release date and time; only values released at or before the decision time may be used as features. For macroeconomic series, use the initial release values with their real-time publication timestamps rather than revised figures (Livnat and Mendenhall, 2006; Croushore and Stark, 2001).

News and text data: For news-based features, use datasets with clear release timestamps and align them to market data in the same way as macro announcements. The design should mirror existing work where news is tied to trading decisions through precise timing information (Benhenda, 2025; Tetlock, 2007).

3 QFRS-2: Labeling (ground truth construction)

Defining prediction targets (labels) is one of the most important tasks in an AI predictive pipeline. Labels are constructed from market prices, and how these labels logically map to trading actions must be made explicit. Explicit target definitions prevent hidden degrees of freedom, ensure comparability across machine learning architectures, and guarantee that the boundaries between feature windows and label windows remain strictly separated. Most studies fix a precise horizon h (for example, one bar ahead, five bars ahead, or next-day close) (Gu et al., 2020).

High-frequency limit order book (LOB) tracks may define h either in clock time (e.g., 10 seconds ahead) or in event time (e.g., the price after the next 50 order book ticks/messages). Event-time horizons naturally normalise periods of high and low market activity and are standard practice in modern LOB benchmark datasets (Xiao et al., 2025).

To avoid information leakage, the feature window must close entirely before the label window opens. If an order is simulated to fill at time t, no data generated at or after t may be used to construct the prediction features. This strict separation principle is non-negotiable in empirical asset pricing and automated trading research to ensure that reported profits represent genuine out-of-sample capability.

There are generally four methods for labelling.

3.1 Labels for regression tasks

Regression labels require a continuous-valued target that measures the magnitude of a future price movement or volatility level. The following sub-types are supported.

3.1.1 Log return and simple return regression

For return forecasting, the standard label is the log return over a future window from time $\boldsymbol{t}$ to $\boldsymbol{t+h}$, defined as:

$\boldsymbol{r_{t,t+h}^{\log} = \log P_{t+h} - \log P_t}$where $\boldsymbol{P_t}$ is the transaction price or mid-price at time $\boldsymbol{t}$. This target is foundational in empirical asset pricing because log returns aggregate additively across time, making them mathematically robust for long-horizon predictive regressions and cross-sectional machine learning studies (Gu et al., 2020). The simple return $\boldsymbol{r_{t,t+h}^{\text{simple}} = (P_{t+h}/P_t) - 1}$ may also be used; the choice must be stated explicitly.

3.1.2 Mid-price change for high-frequency data

In limit order book (LOB) settings, transaction prices are often subject to bid-ask bounce. Consequently, the standard target is the change in the mid-price $m_t = (a_t + b_t)/2$, where a_t and b_t are the best ask and bid prices. The continuous mid-price change over a horizon of h events or time steps is defined as: $\Delta m_{t,t+h} = m_{t+h} - m_t$.Predicting this exact change is central to modelling order book imbalances and short-term liquidity shocks (Zhang et al., 2019).

3.1.3 Realised variance and volatility labels

For volatility prediction (often used in position sizing and risk management tracks), the target is typically realised variance over a fixed horizon $[t,t+h]$. It is defined as the sum of squared high-frequency intraday returns r_j:

$$RV_{t,t+h} = \sum_{j=1}^{N} r_j^2 .$$

where r_j are the N high-frequency intraday log-returns sampled within $[t,t+h]$ at a fixed frequency Δ, and $N = h/\Delta$. This model-free definition provides a consistent, observable proxy for latent volatility and is standard in econometric and machine learning volatility forecasting (Poon and Granger, 2003).

3.2 Labels for classification tasks

Classification labels discretise continuous returns or mid-price changes into a finite set of categories. Two principal constructions are supported.

3.2.1 Directional labels with thresholds (multi-class)

Classification tasks routinely discretise returns or mid-price changes. To separate meaningful market movements from microscopic noise, a directional target must include a no-change threshold θ:

$$y_{t,t+h} = \begin{cases} +1 & \text{if } \Delta m_{t,t+h} > \theta, \\ 0 & \text{if } |\Delta m_{t,t+h}| \le \theta, \\ -1 & \text{if } \Delta m_{t,t+h} < -\theta. \end{cases}$$

This multi-class construction is heavily utilised in deep learning for high-frequency finance, for example in the DeepLOB architecture, to ensure models are not penalised for failing to predict negligible price fluctuations that cannot be profitably traded due to spread costs (Zhang et al., 2019).

3.2.2 Binary classification

A widespread target in financial machine learning is predicting the simple binary direction (up or down) of the next bar. While intuitively simple, the standard requires this target to be rigorously defined to handle zero-return events (which are frequent due to discrete tick sizes). The binary label must be explicitly defined as:

$$y_{t,t+h} = \begin{cases} 1(\text{Up}) & \text{if } r_{t,t+h} > 0, \\ 0(\text{Down}) & \text{if } r_{t,t+h} \leq 0, \end{cases}$$

Exact zero-return could be excluded from the training/evaluation set entirely.
Some forecasting tasks target the occurrence of specific discrete events, for example, earnings surprises exceeding a threshold, flash crashes, circuit-breaker triggers, or regime-change points. Such events could also be labelled as binary. Another form of binary classification is Triple-barrier labelling where the binary labels [+1, -1], defined by a predefined trading strategy by incorporating profit-taking, stop-loss, and maximum holding-period barriers simultaneously (De Prado, 2018). For each event initiated at time t:

- An upper barrier is set at $P_t(1 + \tau_u \sigma_t)$, where τ_u is a profit-taking multiplier and σ_t is a rolling volatility estimate.
- A lower barrier is set at $P_t(1 - \tau_l \sigma_t)$, where τ_l is a stop-loss multiplier.
- A vertical barrier is placed at $t + h_{\max}$, representing the maximum holding period.

The label is determined by whichever barrier is touched first:

$$y_t = \begin{cases} +1 & \text{if upper barrier is touched first,} \\ -1 & \text{if lower barrier is touched first,} \\ \text{sign}(r_{t,t+h_{\max}}) & \text{if vertical barrier is reached.} \end{cases}$$

In cases where neither the profit-taking nor stop-loss barrier is hit before the maximum holding time (vertical barrier), the label is set to $\text{sign}(r_{t,t+h_{\max}})$, so that it encodes only the direction (profit vs loss) of the realised return over the full holding period. Zero-return outcomes are rare, but when they occur, they may be excluded from the training and evaluation set or retained as a third label.

Studies employing binary classification are highly susceptible to ignoring the usability of such models as predicting an up/down move of a very small fraction is economically meaningless but registers as statistically correct in accuracy metrics (Dixon et al., 2020).

3.3 Labels for Ranking and Selection Tasks

When the prediction task is to rank assets for portfolio construction (e.g., top-k long/short strategies), the relevant label is the cross-sectional ordering of future returns rather than their absolute magnitudes. At each rebalance date t, the label for asset i is its forward-looking excess return relative to the cross-sectional mean (Poh, 2021; Gu et al., 2020):

$$y_{i,t} = r_{i,t,t+h} - \bar{r}_{t,t+h}$$

where $\bar{r}_{t,t+h}$ is the equal- or value-weighted mean return across all assets in the universe over the horizon h. Alternatively, labels may be expressed as percentile ranks of future returns within the cross-section, or as binary indicators of membership in the top/bottom decile. The key requirement is that the label definition must match the evaluation standard: if the model is assessed by Spearman rank correlation or Precision@k, then the label must encode the cross-sectional ordering that these metrics measure (Gu et al., 2020; Kou and Lu, 2025).

3.4 Labels for Probabilistic Forecasting Tasks

When the model is required to output a full predictive distribution, a set of quantiles, or prediction intervals, the label is the realised value against which the forecast distribution is evaluated typically the realised return $r_{t,t+h}$, the realised mid-price change $\Delta m_{t,t+h}$, or the realised variance $RV_{t,t+h}$. No discretisation is applied; the continuous realised value serves as the ground truth (Meligkotsidou et al., 2014).

The distinction from regression labels described above lies not in the label itself but in the model output and evaluation: a probabilistic model produces a full conditional distribution $\hat{F}(y|x_t)$, and the evaluation uses strictly proper scoring rules rather than point-forecast error metrics (Gneiting and Raftery, 2007).

Under this standard, the label construction for probabilistic tasks must specify

- The continuous target variable (return, mid-price change, volatility, or other).
- The horizon h and any event-time versus clock-time convention.
- The quantile grid or interval coverage levels at which forecasts will be evaluated.

3.5 Labels for reinforcement learning

In reinforcement learning (RL) trading systems, the target is usually not expressed as an explicit supervised label such as an up/down class or next-period return. Instead, the agent interacts with a market environment and receives a reward determined by the realised consequence of its action. Under QFRS, this reward plays the same methodological role as a label: it must be precisely defined, strictly forward-looking, and fully aligned with the execution and backtesting assumptions of the study. RL trading and portfolio papers commonly optimise cumulative portfolio return, log-return, or a utility-adjusted reward rather than a conventional regression or classification target (Pippas et al., 2025; Meng and Khushi, 2019). Therefore, RL submissions must explicitly define the reward function at each decision time t, including: (i) the action space (e.g. buy, sell, hold, portfolio weights, or position changes), (ii) the reward horizon, whether one-step or episodic, (iii) whether the reward is based on simple return, log-return, portfolio value change, or a risk-adjusted utility, and (iv) whether transaction costs, slippage, borrowing costs, leverage constraints, and inventory penalties are included directly in the reward. A minimally acceptable one-step reward for a single-asset trading agent is the realised net portfolio return from t to $t+1$, for example

$$r_t = \log\left(\frac{V_{t+1}}{V_t}\right),$$

where V_t denotes portfolio value after action execution and transaction costs are deducted before V_{t+1} is computed. In multi-asset settings, the reward may instead be portfolio return or a utility-adjusted objective, but this must be stated exactly and justified relative to the claimed investment objective.

4 QFRS-3: feature engineering (anti-leakage by design)

Feature engineering is one of the most common sources of unintentional data leakage in financial machine learning. A model that appears profitable during validation will collapse in live trading if its input features secretly depend on information that was not available at the time of the prediction. Many published works in financial forecasting still apply full-sample preprocessing or use revised data without recognising the resulting bias (Lommers et al., 2021).

4.1 Allowed feature categories and the point-in-time rule

The standard permits a broad range of input features, provided every feature satisfies a strict "computable at decision time" constraint.

- **Market data (price, volume, order book)**

- Raw prices, volumes, and limit order book (LOB) snapshots are the most basic inputs to financial forecasting models. For high-frequency LOB features, the order book state must be captured at or before the decision event time. Work on deep learning for limit order books demonstrates this discipline by using strictly historical snapshots up to the prediction boundary (Zhang et al., 2019). The LOBCAST benchmark framework follows the same principle by extracting market observations up to time t and assigning labels strictly after t (Prata et al., 2024).

- **Technical indicators**

- Technical indicators such as moving averages, relative strength index, and stochastic oscillators are derived from historical price and volume series. A recent survey of feature selection and extraction techniques for stock market prediction documents the widespread use of these indicators and confirms that careful feature selection can materially change forecast accuracy (Htun et al., 2023). Under this standard, every indicator must be a causal function of past data only, and the lookback window must be fully defined.

- **Fundamental data**

- Earnings reports, balance sheet items and analyst estimates must carry a publication timestamp. Only values whose release date is at or before the decision time may be used. Real-time accounting and fundamental databases exist precisely because standard databases often overwrite initial releases with revised numbers, creating look-ahead bias (Livnat and Mendenhall, 2006).

- **Macroeconomic data**

- Macroeconomic indicators such as gross domestic product (GDP), employment figures and consumer prices are heavily revised after their initial release. Using the final revised figure in a backtest gives the model information it could not have had. The Federal Reserve Bank of Philadelphia maintains a real-time dataset specifically to address this problem, and any macro feature must use the initial release vintage available at decision time, not the ex-post revised value (Croushore and Stark, 2001).

- **News and text data**

- News-based features must be aligned to market data using the article or wire-service timestamp. Recent financial news datasets designed for machine learning forecasting include precise publication times to support this requirement (Saqur et al., 2024). Under this standard, a news item may contribute to a feature only after its stated release time.

4.2 Feature windowing rules

When creating rolling or lookback features (moving averages, rolling volatilities, momentum scores, or any cumulative statistic), the observation window must be strictly causal.

- **Lookback boundaries**

- Any rolling computation at time t must consume data only from the interval $[t - w,t]$, where w is the defined lookback window. Studies that compare different temporal input sizes for LOB-based deep learning confirm that varying the lookback window changes model behaviour and that input windows must remain within the historical boundary (Prata et al., 2024).

- **Causal filtering only**

- Feature pipelines sometimes use digital filters or signal processing to extract trends or smooth noise. Any filter applied must be causal (backward-looking only). Bidirectional filters or bidirectional recurrent layers that use both past and future data points to smooth a series violate the standard because they depend on values not yet observed at decision time. This is one of the data leakage vectors highlighted in the machine learning for finance literature (Lommers et al., 2021).

4.3 Prohibited operations

To prevent forward-looking bias, the following operations are prohibited:

- **Full-sample statistics (scaling leakage)**

- It is a frequent error to compute the mean, standard deviation, minimum or maximum of a feature across the entire dataset (including validation and test periods) and then use those statistics to standardise or normalise before splitting the data. Doing so injects future distributional information into the training window. This error has been identified as a common source of inflated results in financial forecasting studies (Dixon et al., 2020). Scaling parameters must be fitted exclusively on the training window and then applied forward to validation and test data.

- **Full-sample signal decomposition**

- Methods such as Wavelet denoising , Empirical Mode Decomposition (EMD), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Discrete Wavelet Transform (DWT) and Variational Mode Decomposition (VMD) are popular for pre-processing financial time series (Wen et al., 2024). However, applying these decompositions to the entire dataset at once causes serious information leakage because the decomposition algorithm uses data from the test period to shape historical components. A recent study demonstrated this problem explicitly: when CEEMDAN was applied to the full series before splitting, RMSE appeared roughly 100 times lower than when the decomposition was restricted to the training window, confirming that the apparent accuracy was illusory (Yang et al., 2024). Under this standard, any decomposition must be fitted within the training sliding window boundary (Yang et al., 2024).

- **Revised or backdated data**

- Many economic and fundamental variables are revised weeks or months after initial release. Using final revised values in a simulation that occurred before the revision was published is a form of look-ahead bias. The real-time data literature has long emphasised the gap between initial and revised macro releases and the resulting distortion in forecasting studies (Croushore and Stark, 2001). Under this standard, all macro and fundamental features must use the version available at the decision time.

- **Feature-target contamination**

- No feature may be computed from data that overlaps with the label window. If the label is the return from t to $t + h$, no feature may include any observation from time $[t + 1, \infty]$ onward. This strict separation is documented as standard practice in high-frequency LOB benchmarks, where the feature window and the label window are explicitly non-overlapping (Zhang et al., 2019; Prata et al., 2024)

5 QFRS-4: Scaling / Normalisation

Scaling parameters must be fitted only on the training window and then applied forward to validation and test data. Computing global statistics (mean, standard deviation, min, max) across the entire dataset before splitting is one of the most common sources of data leakage in machine learning, and it is particularly damaging in time-series financial forecasting because distribution parameters shift over time. When a scaler is fitted on the full dataset, test-period distributional information (future means, variances, ranges) leaks into training features.

Kapoor and Narayanan (2023) identify this as leakage type as [L1.2] "Pre-processing on training and test set" and show that it affected at least 294 papers across 17 scientific fields, often inflating reported model performance to the point of complete irreproducibility (Kapoor and Narayanan, 2023). Sasse et al. (2024), in a comprehensive pipeline-level analysis, confirm that "estimating the preprocessing parameters on the whole dataset invalidates the train/test separation" and that models can exploit the leaked statistics, yielding overly optimistic generalisation estimates (Sasse et al., 2023).

The scaling leakage is so pervasive that automated tooling now exists. Yang et al. (2022) developed a Datalog-based static analysis that detects preprocessing leakage in Python notebooks (Yang et al., 2022), and AlOmar et al. (2025) extended this into the LeakageDetector PyCharm plugin (AlOmar et al., 2025). The LeakageDetector static-analysis tool specifically flags Preprocessing Leakage as its most frequently detected category (55.6% of all detected instances), confirming that this is the single most common coding error in ML pipelines (AlOmar et al., 2025).

Under the QFRS standard the following rules apply to every form of feature scaling or normalisation:

- **Z-score (standardisation).** Compute the mean $\mu_{\text{train}}$and standard deviation σ_{train} exclusively on the training window. Transform all windows (train, validation, test) using those same training-derived parameters: $\hat{x} = (x - \mu_{\text{train}})/\sigma_{\text{train}}$.

- **Min–max scaling.** The minimum and maximum values used to scale features to a $[0, 1]$ or $[-1, 1]$ range must come from the training window only. If a test-period value falls outside the training range, it is clipped or left out-of-range. It must not be used to recompute the bounds.

- **Robust scaling (median / IQR).** The same rule applies: median and interquartile range are estimated on training data only.

- **Walk-forward / expanding-window updates.** In a rolling backtest, the scaler may be re-fitted each time the training window advances, but it must never include data beyond the current decision boundary. The Liu et al. (2022) leakage-suppression framework for time series applies exactly this logic by using overlapping slices that are decomposed and normalised independently within each historical window (Liu et al., 2022).

5.1 Instance-wise normalisation and non-stationarity

Recent deep-learning research has introduced instance-wise normalisation methods that remove per-window trend and seasonal components before forecasting and add them back afterward. These are permitted under this standard provided they remain strictly causal:

- **RevIN (Reversible Instance Normalization)** subtracts the per-instance (per-input-window) mean and divides by its standard deviation, then reverses this on the output. Because the statistics are derived from the input window alone (not from the label window or future data), RevIN is causal and permissible (Kim et al., 2021).

- **FAN (Frequency Adaptive Normalization)** extends instance normalisation into the frequency domain, removing the top-K Fourier components from each input window and predicting their evolution into the forecast horizon. Ye et al. (2024) show that FAN achieves 7–38% MSE improvement over backbone models precisely because it handles evolving seasonal non-stationarity that simple mean/variance normalisation misses (Ye et al., 2024).

- **Batch Normalization (BN) and Layer Normalization (LN)** in deep learning application to time-series forecasting models can introduce subtle leakage. Shen, Wei and Wang (2022) show that BN violates the scale-sensitivity property of time series (because batch statistics mix information across instances at different time steps), and that LN can leak future information in causal (decoder) architectures (Shen et al., 2022). They conclude that only normalisation methods that do not directly alter hidden-unit statistics (e.g. Weight Normalization) are safe for time-series forecasting (Shen et al., 2022).

6 QFRS-5: Train/Validation/Test split

The splitting standard aims to provide an unbiased estimate of out-of-sample performance while strictly preventing any information from test periods from influencing model training, feature construction, or hyperparameter selection. Splits must respect chronological order and market calendars and must remain reproducible across implementations.

Unlike IID data (independent and identically distributed) financial time series display temporal dependence, volatility clustering, and regime changes. These characteristics break the assumptions behind random shuffling and standard k-fold cross-validation, because past and future samples become statistically entangled in ways that artificially inflate performance estimates (Cerqueira et al., 2020).

Imagine flipping a fair coin 1,000 times: shuffling the sequence before splitting into train/validation/test sets is harmless, because each flip is independent and identically distributed. By contrast, in financial asset price forcast, clustering of high-volatility periods and structural breaks means that shuffling intermixes calm and crisis regimes,

letting the model see patterns from future crises while training, and thus contaminating the evaluation. Performance estimation methods for financial time series must therefore preserve temporal order, prevent any leakage of future observations into training or validation, and remain robust to non-stationarity and regime shifts (Hewamalage et al., 2023; Lones, 2024).

6.1 Single train–validation–test chronological split

A simple baseline is a **single chronological split**: the earliest segment is used for training, the middle segment for validation, and the most recent segment for testing. This is acceptable for large datasets with moderate non-stationarity, but it yields a single performance estimate with potentially high variance and is sensitive to the specific choice of split dates (Cerqueira et al., 2020; Sasse et al., 2025).

6.2 Rolling-origin / walk-forward evaluation

To obtain more stable estimates, we recommend rolling-origin (walk-forward) training and validation. The forecast origin moves forward in time, models are retrained (or updated) on data up to the origin and evaluated on the next block. Repeating this across multiple origins yields a distribution of out-of-sample errors. In this benchmark, each fold consists of a training window of earliest segment up to time, a validation window, a test window. All three are strictly ordered in time.

6.3 Design rules for data split

- **Chronological ordering and embargoing:** All splits must be strictly chronological at the bar level (tick, minute, day). When labels use a forward horizon h (e.g., future h-day return), we require an embargo period of at least h bars between the end of one window and the start of the next to avoid overlapping feature and label windows across train/validation/test.
- **Asset-level vs time-level splits (panel data):** For panel data (many assets over time), the default is time-based splits: the same calendar dates define train/validation/test across all assets, ensuring that all models learn only from the past relative to the evaluation period. Asset-based splits (train on a subset of assets, test on others) may be added as robustness checks for cross-sectional generalisation, but they do not replace time-based splits.
- **Handling multiple assets, markets, and regimes:** To mitigate overfitting to a single instrument or episode, experiments can cover multiple assets and multiple non-overlapping time segments, ideally spanning different regimes (bull, bear, sideways). Regime labels, if used, must be derived only from past data and must not peek into the test segment. Recent benchmarking work in time-series forecasting emphasises the importance of diverse datasets and careful evaluation design for fair comparison of methods (Qiu et al., 2024; Wang et al., 2024b; Cerqueira et al., 2024)

6.4 Prohibited operations

Random shuffling or standard k-fold cross-validation that mixes past and future observations are common leakage causes and are prohibited. These procedures violate temporal dependence and can produce severely optimistic error estimates for forecasting tasks.

7 QFRS-6: evaluation metrics and task types

Studies could be grouped in one of the four prediction task families. Each study must declare its task type at submission time and report the corresponding primary metrics in full.

7.1 Regression: returns, price or price changes

The model outputs a continuous forecast of future return or price change over a defined horizon h. Primary metrics: MAE, RMSE, Information Coefficient (IC), Information Coefficient Ratio (ICIR). Secondary (recommended): Pearson correlation, Spearman rank correlation, R^2, sMAPE (symmetric mean absolute percentage error). R^2 is often reported in forecasting and asset-pricing regressions as a familiar goodness-of-fit measure, but in financial return prediction it is typically small even for economically useful models and should therefore be interpreted alongside IC/ICIR and backtest performance rather than in isolation (Gu et al., 2020; Hyndman and Koehler, 2006; Kim et al., 2025).

Error-based metrics are often reported in the studies where given true values $y_t$and point forecasts $\hat{y}_t$over test index set T:

- Mean Absolute Error: $\text{MAE} = \frac{1}{|T|}\sum_{t\in T} |y_t - \hat{y}_t|$. MAE gives a natural scale of average absolute miss and is robust to outliers, making it a reliable primary metric for heavy-tailed financial returns .

- Mean Squared Error and Root MSE: $\text{MSE} = \frac{1}{|T|}\sum_{t\in T}(y_t - \hat{y}_t)^2$; $\text{RMSE} = \sqrt{\text{MSE}}$. MSE/RMSE penalise large errors more heavily, which is useful when large deviations carry disproportionate cost, but can be dominated by a small number of tail events that are characteristic of financial returns. Both MAE and RMSE must be reported for regression tasks.

sMAPE may be reported as an optional metric for price-level forecasting, but is unreliable when returns are close to zero. Studies should state explicitly when relative metrics are included and why (Hyndman and Koehler, 2006).

Error metrics assess pointwise accuracy but say little about cross-sectional information content, which is central for long-short portfolio construction (Gneiting and Raftery, 2007; Zhang et al., 2024; Nonejad, 2017).

- **Pearson and Spearman correlation.** Pearson measures linear association between $\hat{y}$and y; Spearman measures rank-based monotonic association and is closer to ranking utility in portfolio construction.

- **Information Coefficient (IC).** The Information Coefficient (IC) quantifies how close the predicted value is to the actual value by calculating the average Pearson correlation coefficient. For each time slice t, compute the cross-sectional Pearson correlation between predicted values and realised returns across assets i: $\text{IC}_t = \rho(\hat{y}_{i,t}, y_{i,t})$. The reported IC is the time-average: $\overline{\text{IC}} = \frac{1}{|T|}\sum_{t\in T} \text{IC}_t$(Grinold and Kahn, 2000).

- **Information Coefficient Ratio (ICIR).** The Information Coefficient Ratio (ICIR) is the ratio of IC to its standard deviation, calculated by dividing by the time standard deviation of IC: $\text{ICIR} = \overline{\text{IC}}/\sigma(\text{IC}_t)$. ICIR plays a role analogous to the information ratio: a high IC that fluctuates wildly across time is less useful than a modest but stable IC, and ICIR captures this distinction (Grinold and Kahn, 2000).

MAE/RMSE assess absolute level accuracy and are primary when the task is to predict magnitudes for position sizing, volatility forecasting, or price-level estimation. IC/ICIR assess cross-sectional ranking quality and are primary when the task is to generate tradable alpha signals where only the ordering of predictions matters.

7.2 Classification: directional or event

The model outputs a discrete label (up/down, hit barrier, or event/no-event) or a probability thereof. Primary metrics: Matthews Correlation Coefficient (MCC), Precision–Recall AUC (PR-AUC), Balanced Accuracy, confusion matrix with class prevalence. Secondary: F1, ROC-AUC, calibration metrics. MCC is strongly preferred as a primary metric because it remains informative under the severe class imbalance common in financial event prediction, whereas accuracy and F1 can produce dangerously over-optimistic scores when one class dominates (Chicco and Jurman, 2023a; Chicco and Jurman, 2023b) .

Confusion matrix and class prevalence: Submissions must report the full confusion matrix (TP, TN, FP, FN) and the class prevalence for each evaluation window. Reporting only aggregate scores without prevalence is misleading, particularly when the up/down ratio shifts across market regimes. Chicco and Jurman (2020) demonstrate that ignoring

prevalence leads to dangerously optimistic interpretations of accuracy and F1 in imbalanced settings (Chicco and Jurman, 2023b; Chicco and Jurman, 2023a).

Accuracy is reported but explicitly de-emphasised. In a market where 55% of days are "up", a naive classifier that always predicts "up" achieves 55% accuracy while providing no information. Precision, Recall, F1 summarise performance on the positive class. Balanced accuracy, mean of sensitivity and specificity; important when class ratios evolve across regimes.

Matthews Correlation Coefficient (MCC): MCC is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

MCC is the required primary classification metric because it produces a high score only when predictions are correct in all four confusion matrix quadrants, proportionally to both positive and negative class sizes. It has been shown to be more informative and reliable than accuracy, F1, balanced accuracy and ROC-AUC across a wide range of imbalanced settings.

When threshold-free metrics and calibration is required, use ROC-AUC and PR-AUC. ROC-AUC and PR-AUC integrate performance across all thresholds; PR-AUC is more informative than ROC-AUC when the positive class is rare, a common situation in event-driven trading (Davis and Goadrich, 2006).

7.3 Ranking and selection (top-k long/short)

The model outputs a score; evaluation concerns the quality of the top (or bottom) ranked assets for portfolio construction. Primary metrics: Spearman rank correlation, Precision@k and/or HitRate@k , selection stability across rebalance dates. Ranking metrics are central for long-short equity strategies where only the monotonic ordering of scores matters, not their absolute level (Gu et al., 2020; Kou and Lu, 2025; Petropoulos et al., 2022).

7.4 Probabilistic forecasting

The model outputs a full predictive distribution, a set of quantiles, or prediction intervals. Primary metrics: at least one strictly proper scoring rule (CRPS or WIS), plus pinball loss at a declared quantile grid. Secondary: empirical coverage at 50%, 80% and 95% levels, sharpness statistics. Proper scoring rules are required because they are minimised in expectation only by the true distribution, preventing distributional gaming (Gneiting and Raftery, 2007). The Weighted Interval Score (WIS), introduced in the epidemic forecasting literature, provides a practical approximation to CRPS for interval-format predictions and decomposes naturally into under-prediction, over-prediction and sharpness penalties (Bracher et al., 2021). These metrics are not very popular in financial markets prediction, however, due to their possibility of use are listed.

Table 1 Summary of evaluation metrics by task type in QFRS.

Task Type	Primary Metrics	Secondary Metrics	Key Interpretation Guidelines
Regression Continuous forecasts of returns, price movements, or volatility over horizon h	- MAE (Mean Absolute Error) - RMSE (Root Mean Squared Error) - IC (Information Coefficient) - ICIR (IC Ratio: IC/σ(IC))	- Pearson correlation - Spearman rank correlation - R^2 (with caveats) - sMAPE (price-level only)	- MAE/RMSE assess magnitude accuracy (position sizing, level estimation) - IC/ICIR assess cross-sectional ranking quality (long-short alpha) - R^2 typically small in finance; interpret alongside IC/ICIR and backtest P&L - Report all four primary metrics
Classification	- MCC (Matthews Correlation Coefficient) - PR-AUC (Precision-Recall AUC)	- F1 score - ROC-AUC	- MCC required primary metric (robust to severe

Binary (up/down) or multi-class (up/flat/down) direction; rare event detection	- Balanced Accuracy - Full confusion matrix + class prevalence	- Brier score - ECE (Expected Calibration Error)	class imbalance) - Must report confusion matrix (TP, TN, FP, FN) with class prevalence for each window - Accuracy alone is misleading when class distribution skewed - PR-AUC > ROC-AUC when positive class is rare
Ranking / Selection (top-k long-short) Model outputs scores for ranking assets; evaluation focuses on top/bottom deciles for portfolio construction	- Spearman rank correlation - Precision@k / HitRate@k - Selection stability across rebalance dates	- Kendall's τ - NDCG (Normalized Discounted Cumulative Gain) - Portfolio turnover	- Only monotonic ordering matters, not absolute score levels - Central for long-short equity strategies - Stability metrics prevent excessive churning
Probabilistic (distributions, quantiles, intervals) Full predictive distributions, quantile forecasts, or prediction intervals for uncertainty quantification	- CRPS (Continuous Ranked Probability Score) or WIS (Weighted Interval Score) - Pinball loss at ≥3 declared quantiles - Empirical coverage at 50%, 80%, 95% confidence levels	- Sharpness statistics (interval width) - Reliability diagrams / calibration curves - Quantile score decomposition	- Strictly proper scoring rules required (CRPS/WIS) - WIS decomposes into under-prediction, over-prediction, sharpness penalties - Verify calibration: nominal 95% intervals should contain 95% of outcomes

8 QFRS-7: Backtest metrics (economic performance)

Forecast accuracy alone does not determine whether a model creates economic value. A strategy that reduces RMSE by a fraction of a basis point may be statistically detectable yet worthless once transaction costs, slippage and capital constraints are accounted for (Hu, 2025; Chen et al., 2025; Dinler, 2025; Zhang et al., 2025). Conversely, a model with moderate error scores may generate substantial risk-adjusted profit because its errors are uncorrelated with costly tail events. Therefore, this standard requires every submission to be mapped into concrete trading rules that determine when positions are opened and closed, how stop-loss and take-profit levels are placed, and how position size (volume or lot size) is scaled as a function of risk. These rules, together with explicit assumptions about transaction costs, slippage, and execution venues, produce a path of portfolio equity from which returns, drawdowns, and risk-adjusted ratios are derived.

The following sections formalise this mapping and items are labelled (M) Mandatory which are required for any study making trading-performance claims; absence of any mandatory item constitutes a Fail on QFRS-7. Items labelled (R) Recommended are strongly encouraged as good practice but are not gatekeeping criteria.

8.1 Trade Statistics and Risk (M)

Return and risk must be reported at the strategy level, conditional on the actual trades taken under a fully specified set of execution rules. For each backtest, the standard requires authors to report not only aggregate portfolio returns but also the underlying trade statistics that explain how those returns were generated.
At a minimum, the following quantities must be disclosed:

i. Testing period market regimes must have bull (rise of ≥ 20% from a recent market low), bear (drop of ≥ 20% from a recent market high), side ways market conditions during the testing period. Bull/bear market threshold must be defined for tested asset classes.
ii. Number of trades executed over the evaluation period, separately for long and short sides if both are permitted.
iii. Trade volume / lot size policy, including the nominal position size per trade (e.g., fixed lot, fixed notional, volatility-scaled) and any maximum gross or net exposure constraints.
iv. Win rate, defined as the fraction of closed trades with non-negative net profit after costs.
v. Average profit and loss per trade, reported separately for winning and losing trades and net of all stated frictions.
vi. Holding-period distribution, summarised by median and interquartile range of trade duration, since risk profiles differ sharply between intraday, swing, and multi-week positions.
vii. Total profit or loss over the evaluation period and its annualised equivalent.
viii. Report **Maximum drawdown (MDD)**: the largest peak-to-trough decline in the equity curve. Maximum drawdown is one of the most widely used risk indicators in the fund management industry because it directly measures the worst historical loss an investor would have experienced (Chekhlov, Uryasev and Zabarankin, 2005). A strategy that reports an attractive Sharpe ratio of 1.5 but suffered a 60% drawdown during a crisis period would have been abandoned by most investors long before the Sharpe was realised, because the psychological and margin-call pressure of such a drawdown is unsustainable. Drawdown captures this survival risk in a way that volatility alone cannot: two strategies with identical annualised volatility can have dramatically different drawdown profiles if one concentrates its losses in a single streak while the other spreads them evenly. For this reason, MDD is a required leaderboard metric and not merely optional.

8.2 Trading frictions and implementability (M)

The standard treats backtest performance as credible only if it survives realistic trading frictions and is supported by a sufficiently rich sample of trades. Any strategy that reports gross returns or Sharpe ratios without explicit assumptions on transaction costs, slippage and position constraints is considered non-implementable, as even modest frictions can erase apparent alpha in systematic strategies (Polbennikov et al., 2021; Mansouri and Eterovic, 2023; Almgren and Chriss, 2001). In addition, the framework explicitly guards against three risk-management failure modes: regime overfitting, metric over-optimisation, and under-sampled trading activity (McNeil et al., 2015; Petropoulos et al., 2022). Regime overfitting arises when a model is tuned to a single environment such as a post-crisis bull market and then collapses when volatility or liquidity conditions change; rigorous walk-forward studies in algorithmic trading show that performance can flip sign across regimes even when in-sample metrics look stable (Kercheval and Zhang, 2015; Bahoo et al., 2024). Metric over-optimisation occurs when hyperparameters and trading rules are repeatedly adjusted to maximise a single indicator (for example, Sharpe) across many backtests; Monte Carlo studies of backtest selection demonstrate that the best-observed Sharpe under such searching can substantially exceed the true out-of-sample Sharpe, even when no real edge exists (Harvey and Liu, 2015; Openja et al., 2024). Finally, strategies that achieve attractive risk-adjusted returns on the basis of only a handful of trades or an unusually high win rate on a tiny sample are treated with particular suspicion: empirical work on trend-following and microstructure-based strategies shows that stable performance requires hundreds to thousands of independent trades across varying market conditions, and that apparent “perfect” win rates are almost always a sign of overfitting or unmodelled costs (Singh et al., 2022; López de Prado, 2022)

8.3 Backtesting tools and implementation details (M)

Implementing backtests correctly is technically demanding, particularly when orders include stop-loss and take-profit levels that may be hit within the same bar, partial fills, spread modelling, intrabar volatility and liquidity issues (Bernardini and de Castro, 2021; Gort et al., 2022). Many research studies implement backtesting, if any, in ad hoc Python scripts that operate only on bar-level data and implicitly assume a simplistic fill model. Firstly, the implementation of a proper trade execution, stop loss, take profit, draw down, balance/equity and profit are not technically easy, moreover addressing all above challenges make it hard to correctly implement. For this reason, QFRS strongly encourages, not reinventing the wheel and the use of a mature backtesting engines or trading platforms rather than bespoke one-off scripts (Löw et al., 2015).

There are various free and commercial backtesting engines are available such as Feqtrade (freqtrade.io), MetaTrader, AmiBroker, MultiCharts, QuantConnect or Backtrader (de Castro, 2021). MetaTrader and other similar platforms connect broker fetch live data allows strategy testing by integrating Python and other ML frameworks, visual inspection of trades, manages spread, trade commission, draw down, winrate, total rate, Sharpe Ratio and profit/loss.

Under QFRS, authors remain free to implement backtests in any language, but they are expected either (i) to rely on a documented, widely used backtesting platform whose execution model is clearly specified, or (ii) to provide a precise description of intrabar execution rules and validation tests comparable to those proposed in the backtest-correctness literature (Bhuiyan et al., 2025; Zhong et al., 2025; Zu et al., 2023).

8.4 Risk-adjusted performance (R)

Risk-adjusted performance is fundamental in quantitative investing: the goal is not only to maximise returns, but to do so while explicitly controlling risk. While many classical measures treat risk primarily as return volatility, investors are often more concerned with sustained capital losses, which are better captured by drawdowns. Two of the most widely used risk-adjusted metrics are the Sharpe Ratio (Brinza et al., 2023), which evaluates excess return per unit of total volatility, and the Sortino Ratio (Izadi and Yaghoobi, 2024), which refines this idea by penalising only downside volatility (Basile, 2024; Zabolotskyy et al., 2025). Within this landscape, Zhang and Khushi (Zhang and Khushi, 2020) proposed a new perspective by jointly modelling volatility and drawdown risk in a single objective; similarly, there are many other risk-adjusted metrics exist (K.C and Laha, 2021; Xue and Ye, 2026; Long et al., 2026; Arnold et al., 2025).

Below we briefly review several risk-adjusted performance ratios commonly used by quantitative fund managers, including the Sharpe, Sortino, Sterling, and Calmar ratios. These ratios are not mandated by QFRS, but recommended as they provide useful optional diagnostics that authors may report alongside the required economic performance and drawdown statistics.

Sharpe ratio:

The Sharpe ratio measures the excess return of a strategy per unit of total return volatility:

$$S = \frac{\mathbb{E}[r_t - r_f]}{\sigma(r_t)}$$

where r_t is the strategy return at time t, r_f is the risk-free rate, $\mathbb{E}[\cdot]$ denotes the time-series expectation, and $\sigma(r_t)$ is the standard deviation of strategy returns. The Sharpe ratio is the most widely reported risk-adjusted metric in quantitative finance. However, it treats upside and downside volatility symmetrically, which is inappropriate for strategies with skewed return distributions. A strategy that occasionally produces very large gains (positive skew) is penalised just as much as one that occasionally produces very large losses. Moreover, the Sharpe ratio is easily inflated by backtest overfitting: Bailey and López de Prado (2014) (Bailey and López de Prado, 2014) show that when many strategy configurations are tried, the best in-sample Sharpe can be dramatically higher than the true out-of-sample Sharpe, hence they suggested Deflated Sharpe Ratio to be reported.

Sortino ratio: The Sortino ratio modifies the Sharpe ratio by replacing the total return standard deviation with the downside deviation and the standard deviation computed only over returns that fall below a minimum acceptable return (MAR), typically set to the risk-free rate or zero:

$$\text{Sortino} = \frac{\mathbb{E}[r_t - r_f]}{\sigma_d}$$

where the downside deviation σ_d is defined as:

$$\sigma_d = \sqrt{\frac{1}{|T|}\sum_{t\in T} \min(r_t - r_f, 0)^2}$$

and $|T|$ is the number of observations in the evaluation window. The Sortino ratio addresses the key limitation of the Sharpe by penalising only downside volatility. In practice, investors care far more about avoiding losses than about dampening gains. A strategy that earns 20% per year with occasional sharp drawdowns looks identical to one that earns 20% per year with smooth, steady gains under the Sharpe ratio, but the Sortino correctly distinguishes them. For trading strategies in volatile asset classes such as cryptocurrencies or small-cap equities, where return distributions are frequently left-skewed, the Sortino provides a more realistic picture of risk-adjusted performance.

Sterling ratio: The Sterling ratio is a drawdown-based risk-adjusted performance metric that measures average annual return in excess of a risk-free (or target) rate relative to average maximum drawdown over a specified period. Unlike the Sharpe ratio, which scales excess return by return volatility, and the Sortino ratio, which scales by downside volatility only, the Sterling ratio directly penalises strategies that experience large or repeated peak-to-trough losses, making it particularly sensitive to path-dependent capital erosion. In practice, this means two portfolios with similar Sharpe or Sortino ratios can exhibit very different Sterling ratios if one suffers deeper or more persistent drawdowns, which is often more aligned with an investor's tolerance for risk (Maier-Paape and Zhu, 2018; Goldberg and Mahmoud, 2017).

A common definition of the Sterling ratio over horizon T is

$$\text{Sterling Ratio} = \frac{\grave{R} - R_f}{\text{AvgMaxDD} + 1\%}$$

where $\grave{R}$is the average annual (or period) return, $R_f$is the risk-free (or target) rate, and AvgMaxDDis the average of historical maximum drawdowns over the evaluation window.

Calmar ratio: The Calmar ratio measures the average annual return of a strategy per unit of maximum drawdown over a given evaluation horizon. It focuses directly on the trade-off between growth and worst-case capital loss, making it particularly relevant for leveraged or trend-following strategies where drawdowns are the binding risk constraint for investors.

A common definition of the Calmar ratio over horizon T is

$$\text{Calmar} = \frac{\bar{R}_T}{|\text{MDD}_T|},$$

where $\bar{R}_T$ is the compound annual growth rate (or average annual excess return) over the evaluation window and MDD_T is the maximum peak-to-trough drawdown observed over the same period.

Compared with the Sharpe and Sortino ratios, which scale excess return by volatility (total or downside), the Calmar ratio scales return by realised worst-case loss, directly reflecting the path-dependence of capital erosion. Two strategies with similar Sharpe and Sortino ratios can therefore exhibit very different Calmar ratios if one experiences deeper or more prolonged drawdowns, which is often more aligned with institutional investors' concerns about risk of ruin and capital lock-ups in drawdown-constrained mandates.

9 Reviewer and Editor Checklist

Peer review is the primary enforcement mechanism for any field-wide reporting standard. Reviewers and handling editors are asked to evaluate a submitted manuscript against the seven QFRS criteria before making an acceptance recommendation. Each criterion is designed to be binary and unambiguous: a study either provides explicit, verifiable evidence of compliance or it does not. Vague descriptions, implicit assumptions, or appeals to common practice do not constitute a pass. The checklist in Table 3 is applied as a gatekeeping step before a manuscript proceeds to acceptance; the standard applicability rules in Sections 9.1–9.3 govern how to handle criteria that are not relevant to a particular study design.

9.1 Standard Applicability

Not every standard criterion is relevant to every study. A purely rule-based algorithmic trading strategy involves no statistical model fitting and therefore has no meaningful scaling/normalisation step (S4) or train/validation/test split

(S5). Similarly, a single-asset, single-venue study has no cross-venue synchronisation requirement (a sub-criterion of QFRS-1), and a spot foreign exchange study has no futures roll rules (another sub-criterion of QFRS-1). Therefore for such studies the standard will be coded as Not Applicable (N/A) or when the sub-criterion is logically inapplicable to the study's design. When the standard and sub-criterion is applicable to the study's design but the manuscript provides insufficient, vague, or contradictory evidence. Absence of evidence is treated as Fail; the burden of proof rests with the authors. The standard and sub-criterion is pass when explicit, verifiable evidence of compliance appears in the manuscript.

The following three rules apply uniformly across all seven standards and are formalised mathematically in Section 9.2.

- **Rule 1 (N/A escalation).** A standard is coded N/A at the standard level if the standard or all of its sub-criteria are N/A (the whole standard is structurally irrelevant to the study). An N/A standard will be considered QFRS compliance.
- **Rule 2 (No partial credit).** There is no partial credit, no weighted average, and no minimum threshold. This conservative rule reflects the cumulative nature of information leakage: a single compromised step can invalidate all downstream results including the backtest.
- **Rule 3 (Pass condition).** A standard passes if and only if every applicable sub-criterion (i.e. every sub-criterion not coded N/A) is coded Pass. A single applicable Fail is sufficient to fail the entire standard.

The aggregation rules above are stated in a mathematical form below.

Let:

- $k \in \{1,...,7\}$ index standards,
- $j \in \{1,...,m_k\}$ index sub-criteria within standard k,
- $S_{k,j} \in \{\text{Pass, Fail, N/A}\}$ be the code for sub-criterion $C_{k,j}$.

Define the set of *applicable* sub-criteria for standard k as

$$A_k = \{j \in \{1,...,m_k\} : S_{k,j} \neq \text{N/A}\} \quad \textit{(Eq. 1)}$$

Then the standard-level status of P_k is:

$$\text{Status}(P_k) = \begin{cases} \text{N/A}, & \text{if } A_k = \emptyset, \\ \text{Pass}, & \text{if } A_k \neq \emptyset \text{ and } \forall j \in A_k,\ S_{k,j} = \text{Pass}, \\ \text{Fail}, & \text{otherwise.} \end{cases} \quad \textit{(Eq. 2)}$$

Equivalently, using an indicator function for applicable sub-criteria:

$$\text{Pass}(P_k) = \begin{cases} 1, & \text{if } A_k \neq \emptyset \text{ and } \prod_{j \in A_k} \mathbf{1}\{S_{k,j} = \text{Pass}\} = 1, \\ 0, & \text{otherwise,} \end{cases} \quad \textit{(Eq. 3)}$$

and $\text{Status}(P_k) = \text{N/A}$ when $A_k = \emptyset$.

This means if applicable, (i) a sub-criterion must be coded Pass or Fail; N/A is not permitted for applicable items. (ii) A standard passes only when all applicable sub-criteria pass. (iii) If no sub-criteria are applicable, the standard itself is N/A and does not penalise overall compliance. The formulation eliminates ambiguity and provides a machine-readable specification suitable for future automated screening tools (Section 12.2).

9.2 Mandatory Sub-Criteria Per Standard

Table 2 lists the mandatory sub-criterion for each standard that can never be coded N/A for a study making a forecasting or trading performance claim within the stated applicability scope.

Overall QFRS compliance requires that a study receives a status of either Pass or N/A for all seven standards $k \in \{1,...,7\}$, meaning for all k Status$(P_k) \neq$ Fail. Furthermore, to prevent scope-evasion, standards S1, S2, S3, S6, and S7 are non-waivable and must be evaluated strictly to Pass for any paper claiming trading performance within the framework's baseline applicability scope.

Table 2: Mandatory sub-criteria per standard. Failure on a mandatory sub-criterion constitutes standard failure regardless of the status of other sub-criteria. * Non-waivable for studies within the stated applicability scope.

Standard	Mandatory sub-criterion (never N/A)*	Applicability scope
QFRS-1* Dataset specification	Named data source(s), date range, and bar frequency must be explicitly stated.	All studies
S2* Labelling	Prediction target and forecast horizon precisely defined; strict feature/label window separation demonstrated.	All studies
S3* Feature engineering	All features satisfy the point-in-time rule (no look-ahead bias of any form) even where no new features are engineered.	All studies
S4 Scaling	Scaling parameters fitted on the training window only.	Studies applying any form of feature scaling
S5 Split	Splits strictly chronological; no random shuffling of time-series data.	Studies with any estimated or trained model
S6* Evaluation metrics	Task type declared; corresponding primary metrics reported in full.	All studies
S7* Backtesting	Fully specified trading rule and at least one cost-aware simulation provided.	All studies claiming trading performance

9.3 Seven-Item Reviewer Checklist

Table 3 is the operational checklist applied during peer review. For each standard the reviewer records a status code (Pass / Fail / N/A) and, where the code is Fail or N/A, a brief supporting rationale citing the relevant section of the manuscript. A manuscript is recommended for acceptance and leaderboard inclusion only when all seven standards are coded Pass or N/A, with standards S1, S2, S3, S6, and S7 all coded Pass.

Table 2: Reviewer checklist. Each standard is assessed as Pass, Fail, or N/A according to the aggregation rules in Section 9.1. Standards marked * contain at least one mandatory sub-criterion that cannot be coded N/A (see Table 2). Leaderboard eligibility requires all seven standards to be Pass or N/A (S4/S5).

#	Standard	Reviewer questions
1*	Dataset specification & data handling	Are the asset universe, venue sources, bar frequency, corporate events, time zones, missing-data treatment, and outlier handling described in sufficient detail to rule out survivorship bias, look-ahead bias, and sampling bias? Are data providers and download dates named? Are symbol lists and sample periods stated in full?
2*	Labelling (ground truth)	Are prediction targets (returns, price changes, directional labels, volatility, events) precisely defined with explicit horizons? Is there strict, demonstrable separation between the feature window and the label window with no overlap at any step of the pipeline?
3*	Feature engineering (anti-leakage)	Are all features specified and demonstrably point-in-time? Is there explicit evidence of no full-sample preprocessing, no use of revised or backdated fundamental/macro data, and no signal decomposition (EMD, CEEMDAN, wavelet, VMD) or filter applied to data beyond the current

		training boundary?
4	Scaling and normalisation	Are all scaling and normalisation steps (Z-score, min-max, robust scaling, instance-wise methods, etc.) fitted exclusively on the training window or via strictly causal walk-forward updates? Is there explicit evidence that test- and validation-period distributional information was not used to compute scaling parameters? [N/A: no scaling applied, e.g. purely rule-based strategy with no model fitting.]
5	Train/validation/test split & model selection	Are splits strictly chronological with no random shuffling or standard K-fold cross-validation applied to time-series data? Are split dates stated explicitly? Is an embargo applied where the label horizon creates overlap? Is hyperparameter tuning confined to train/validation with no implicit model shopping on the test set? [N/A: no parameters estimated, e.g. algorithmic trading strategies.]
6*	Evaluation metrics	Is the prediction task type declared (regression / classification / ranking / probabilistic)? Are required primary metrics reported in full? - Regression: MAE and RMSE both reported; IC/ICIR reported for cross-sectional tasks. - Classification: MCC and/or PR-AUC with full confusion matrix and class prevalence. - Probabilistic: at least one proper scoring rule (CRPS or WIS) reported.
7*	Backtesting Risk, robustness & investor readiness	Is there a fully specified trading rule (entry, exit, position sizing, execution timeline) consistent with the predictive setup? Does the backtest incorporate realistic transaction costs and slippage? Are annualised returns, maximum drawdown, win rate, and total trades reported? Are robustness checks provided across time segments, assets, market regimes, and cost assumptions? Are implementation details sufficient for an informed investor to assess live deployability and for the study to qualify for the QFRS leaderboard?

10 Empirical validation

To demonstrate the diagnostic value of QFRS and to ground its design in observed practice rather than theoretical concerns alone, this section presents a systematic compliance audit of 2025 papers retrieved from Scopus using the search term "Forex Prediction". 50 papers were retrieved, of which 9 survey papers and off-topic papers were excluded.

Each paper was evaluated independently against all seven standard criteria (QFRS-1–S7) using a binary pass/fail judgement. A standard criterion receives a pass (✓) only when the paper provides explicit, verifiable evidence of compliance; absence of evidence, vague descriptions, or implicit assumptions are coded as fail (✗).

10.1 Audit procedure for the reviewed studies

The compliance audit of the 41 reviewed studies was conducted using a structured standard-based coding procedure applied uniformly across all papers. Initially, two coders (including the author) independently assessed each study against all seven QFRS standards (S1–S7), assigning each standard a binary judgement of Pass (✓) or Fail (✗). Across the resulting 287 standard-level judgements, the coders agreed on 284 and disagreed on 3, yielding a Cohen's κ of approximately 0.98, which indicates almost perfect inter-rater reliability. On re-reading the underlying papers, all three mismatches were traced to simple human error in applying the decision rules. After correction, the final codes reflected the agreed decision in each case, with a standard coded Pass only when the paper provided explicit, verifiable evidence of compliance; where the paper did not clearly address a criterion, the standard was coded Fail.

A standard received Pass only when the manuscript contained explicit, verifiable evidence that the relevant "*applicable*" sub-criteria had been satisfied. If the authors did not state the required information clearly, the criterion was coded Fail. In other words, absence of evidence was treated as evidence of non-compliance for audit purposes. This conservative rule was adopted to ensure reproducibility and to avoid giving credit on the basis of implicit assumptions, reviewer inference, or common but unstated practice. For example, if a paper reported that data were split into training and test sets but did not specify that the split was chronological, S5 was coded as Fail. Similarly, if a paper stated that normalisation was applied but did not specify that scaling parameters were estimated on the training window only, S4 was coded as Fail. This rule was particularly important for detecting hidden leakage in scaling, decomposition, and data-splitting procedures. As explained in Section 9, if any "applicable" sub-criterion fails, the whole standard is coded as failed.

Ambiguous cases were resolved using the same conservative principle. Where the text of a paper was vague, incomplete, or internally inconsistent, the criterion was coded as Fail unless the necessary methodological safeguard was stated explicitly enough to be independently verified. This applied, for example, to papers that mentioned backtesting without specifying the trading rule, transaction costs, slippage, or execution timing, and to papers that reported accuracy metrics without clearly defining the forecast horizon or target construction. The purpose of this rule was not to penalise authors for brevity, but to reflect the central premise of QFRS: a result cannot be considered reproducible or deployable unless the underlying method is reported with sufficient precision.

For transparency, the reason for each failed standard was recorded in the Comment column of the Table 3. The comments identify which standard failed and briefly state the basis for failure, such as non-chronological splitting, missing evidence of training-only scaling, absence of trading rules, or lack of economic backtesting. This comment field serves as an audit trail linking each binary judgement to a specific reporting deficiency in the underlying paper. As a result, Table 3 is not only a summary of compliance outcomes but also a concise justification record for the coding decisions applied across the 41-study sample.

Of the 41 Forex-relevant papers (excluding surveys or off-topic papers: total 9) not a single paper out of 41 achieved full compliance across all seven standards (Figure 1). Table 3 presents the full compliance matrix. The mean number of standards passed per paper was 4.22 out of 7 (60.3%). Partial compliance (five or more standards passed) was achieved by only 15 papers (36.6%). These results provide direct empirical evidence for the standard's core claim: that the existing literature systematically fails to meet the minimum standards required for results to be economically meaningful, reproducible, and bias-free.

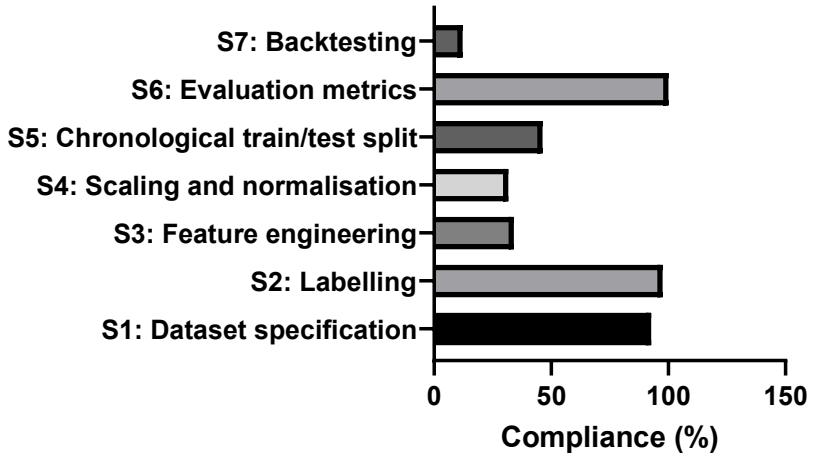

Fig. 1. QFRS compliance rate of Scopus-indexed 'forex prediction' eligible papers published in 2025. Pass requires explicit evidence in the manuscript. S = QFRS Standard.

10.2 Standard-Level Analysis

S1 and S2 are near-universally satisfied. Dataset specification (S1: 92.7%) and label definition (S2: 97.6%) have become standard practice in the field, likely because reviewers and authors alike recognise the need to describe what data and what prediction target is used. The S1 failure and single S2 failure involve papers that have used ambiguous statements, did not provide date ranges, data sources, did not name the asset being predicted, using phrases such as "a specific stock" without identifying the ticker symbol.

S3 (Anti-leakage feature engineering) fails in 63.4% of papers. This is the most technically complex standard criterion and its failure rate confirms the concern raised in Section 4. The dominant failure modes are: (i) application of decomposition methods (EMD, CEEMDAN) to the full dataset before splitting (e.g., Papatsimpas et al.; Xu et al.); (ii) use of full-sample Prophet decomposition with residuals passed to LSTM without point-in-time guarantees (Haider et al.); and (iii) bidirectional or non-causal feature pipelines with no explicit statement of causality (Liu et al., KAN-GARCH-MIDAS; Song et al.).

S4 (Scaling and normalisation) fails in 68.3% of papers. The most prevalent single failure mode is the application of Min-Max scaling to the entire dataset, including validation and test periods, before train/test splitting. Multiple papers explicitly state this in their methodology, e.g., "we applied Min-Max Scaling to normalize our inputs... The data underwent final processing by splitting it into training and testing sections" (Rafi et al.); and "preprocessed with MinMaxScaler and divided into 80% training and 20% testing sets" (Rahat et al.). These explicit statements confirm that scaling leakage is not merely an inference from vague descriptions but a documented pipeline error in a substantial fraction of the literature.

S5 (Chronological train/validation/test split) fails in 51.2% of papers. Failure modes include K-fold cross-validation without time-series awareness (Dakalbab et al.; Xu et al.), an 80/20 or 70/30 split with no explicit statement that the split is chronological (multiple papers), and the absence of a separate validation set for hyperparameter selection (most papers use only a binary train/test split, with no description of how model selection was performed). The use of standard K-fold cross-validation on financial time series is particularly damaging because it allows future periods to inform training, producing severely optimistic performance estimates.

S6 (Statistical evaluation metrics) is satisfied universally (100%). All 41 papers report at least one statistical metric (typically MAE, RMSE, or accuracy), confirming that reporting of forecast error metrics has become a universal norm. The standard does not credit this as a success; rather, the universal compliance on S6 combined with the near-universal failure on S7 highlights the core problem the standard addresses: the literature has converged on reporting statistical metrics while almost entirely ignoring economic metrics.

S7 (Economic backtesting) fails in 87.8% of papers. Only five papers (12.2%) provide any form of economic backtesting. Of these five, several provide only a Sharpe ratio without specifying trading rules, transaction costs, fill assumptions, or position-sizing methodology (e.g., Al-Quraishi et al. provides a Sharpe ratio but "no trading rules, no

transaction costs. and no economic performance metrics" beyond the ratio itself). Only Iswara et al., López-Herrera et al., Uygun et al., Wangchailert et al., and one further paper include some form of trading simulation, and none provides the complete set of backtest reporting elements mandated by QFRS-7. The practical implication is unambiguous: the overwhelming majority of published forex AI forecasting claims have no empirical basis for asserting that the proposed model would generate profit in live trading.

Table 3: *QFRS compliance audit of 41 'forex prediction papers' indexed in Scopus in 2025 (order by first author last name). S1: Dataset specification and data handling, S2: Labelling / ground-truth construction, S3: Feature engineering - anti-leakage, S4: Scaling and normalisation, S5: Train / validation / test split, S6: Evaluation metrics, S7: Economic backtesting*

Reference	S1	S2	S3	S4	S5	S6	S7	Comment
Abbaszadeh et. al. "A Synergistic Hybrid Architecture... for Robust Hour-Ahead Forex Forecasting" (Abbaszadeh et al., 2025)	✗	✗	✗	✗	✗	✗	✗	Fails S7: No economic backtesting, trading rules, or risk-adjusted metrics (Sha provided.
Ahmad et. al. "Foreign Exchange Prediction and Trading Using Low-Shot Machine Learning"(Ahmad et al., 2025)	✗	✗	✗	✗	✗	✗	✗	Fails on look-ahead leakage, random splitting, and lack of economic backtestin
Ahmed, "Leveraging Machine Learning... BRICS Economies"(Ahmed, 2025)	✗	✗	✗	✗	✗	✗	✗	Significant failures in dataset frequency, target labeling, anti-leakage design, a backtesting.
Alqodri et. al. "Comparative Analysis... Dollar Index Prediction"(Alqodri et al., 2025)	✗	✗	✗	✗	✗	✗	✗	Fails on target labeling, look-ahead leakage via global min-max scaling, and no backtesting.
Al-Quraishi et al. "Forecasting... using GRU and SVR"(Al-Quraishi et al., 2025)	✗	✗	✗	✗	✗	✗	✗	S1: No data source/provider named and no dataset date range stated. S3: No po guarantee for features; news sentiment temporal alignment not addressed. S4: Normalization performed but no method specified and no statement of training fitting. S5: No chronological split confirmed. S7: no backtesting, no trading rules, no transaction cost but Sharpe Ratio given.
Chi et al. "Advanced Predictive Modeling... Using Ensembles Learning" (Chi et al., 2025b)	✗	✗	✗	✗	✗	✗	✗	Fails on data handling documentation, leakage prevention, splitting logic, and backtesting.
Chi et al. "Enhancing forex market forecasting... multivariate LSTM models"(Chi et al., 2025a)	✗	✗	✗	✗	✗	✗	✗	Fails S7: Lacks economic backtesting, explicit trading rules, and risk-adjusted (Sharpe/MDD).
Dakalbab et al. Advancing Forex prediction through multimodal text-driven model and attention mechanisms(Dakalbab et al., 2025)	✗	✗	✗	✗	✗	✗	✗	Fails S4, S5, and S7. Uses standard K-fold cross-validation; lacks training-only proof and economic backtesting.
Das et al. ELM-GA-ERWCA: A Tailored Forex Exchange Trading Model...(Das et al., 2025)	✗	✗	✗	✗	✗	✗	✗	S7: economic backtest omitted.
Dave et. al. Predicting Forex Prices (Dave et al., 2025b)	✗	✗	✗	✗	✗	✗	✗	Fails on 6/7 criteria; lacks leakage prevention, chronological split verification, economic backtesting.
Dave et al., "Forex Price Prediction: A Multi-model Approach Integrating Sentiment Analysis Using LLMs with LSTM, XGBoost, Transformer Models" (Dave et al., 2025a)	✗	✗	✗	✗	✗	✗	✗	Fails point-in-time data alignment, causal scaling, and chronological splitting r backtesting.
Duong et al., Investigating Market Strength Prediction with CNNs on Candlestick Chart Images (Duong et al., 2025)	✗	✗	✗	✗	✗	✗	✗	No description of any normalization or scaling (or an explicit statement that no for image inputs or any derived features. Train/validation/test split procedure is specified; no dates, proportions, or confirmation that chronological splitting wa backtesting reported.
Gadalay et al., Enhancing Forex Market Predictive Analytics with ELM and Whale Optimization Algorithm(Gadalay et al., 2025)	✗	✗	✗	✗	✗	✗	✗	Fails strictly point-in-time; no mention of avoiding look-ahead bias, no descrip data normalization/scaling or an explicit statement that none is applied, no bac
Ghahremani et al. - Forex prediction with ALFA (Ghahremani and Nguyen, 2025)	✗	✗	✗	✗	✗	✗	✗	Normalization seesm to be globalisation, as author state, 'We first preprocess th Forex data, i.e., clean, format, and normalize it'. No backtesting.

Haider et al. – Enhancing Forex Market Predictions with a Hybrid Prophet-LSTM Model (Haider et al., 2025)	□	□	□	□	□	□	□	Prophet is a full-sample decomposition model; residuals then fed to LSTM, wi explicit point-in-time / anti-leakage guarantees. “Normalize and reshape data f but no details on scaler type or fitting only on training window: Only “split int T (80%) and test Z (20%)”; no statement that the split is strictly chronological.
Hassaan et al. – FPGA-Accelerated ESN with Chaos Training for Financial Time Series Prediction (Hassaan et al., 2025)	□	□	□	□	□	□	□	Min–max normalization is defined but done on “input features” with no statem min/max are computed only on training data and then applied forward. No exp mention that split is chronological rather than shuffled.
Iswara et al., Integration of Conformal Prediction in Automated Forex Trading Systems: Development and Evaluation (Iswara et al., 2025)	□	□	□	□	□	□	□	No explicit point-in-time or anti-leakage description for feature computation. Normalization described but not localized to training set. No mention of chron split.
Khansalar., The Usefulness of the Currency Constraint for Predicting Forex (Khansalar, 2025)	□	□	□	□	□	□	□	Failed missing point-in-time guarantees, scaling details, and lack of transactio backtesting.
Khansama et al., Predicting FOREX trend using incremental spiking neural network (Khansama et al., 2025)	□	□	□	□	□	□	□	Algorithm steps 4–7 describe encoding using Ifmin, Ifmax, μi, α and creation o neurons, but they do not address temporal causality or look-ahead. No statemer scaling/encoding parameters (Ifmin, Ifmax, μ, α, β, T, N) are fitted only on trai nor any mention of train-only normalization. no explicit chronological ordering boundaries beyond referencing hold-out method.
Lee et al., Analysis of Foreign Exchange Behavior using Sequential Minimal Optimization and Linear Regression (Lee et al., 2025)	□	□	□	□	□	□	□	S3: No explicit statement that features/targets are constructed in a strictly point way or that no future information is used in model fitting or evaluation (Sectio Methodology; Section III A–B, pages 2–4). S4: No description of any scaling/normalization or explicit confirmation that no scaling was done; Weka mentioned but scaler fitting on training only is never specified (Section II, pag Although a 70/30 split is described and K-fold is rejected for time-series reason is not explicitly stated to be chronological and no separate validation set or tun standard is defined (Section II, page 2). S7: No trading rules, no transaction co bid-ask spreads, and no economic performance metrics (return, MDD, Sharpe/ are reported; only forecast error metrics and plots are given (Section III, pages Section IV Conclusions, page 6).
Liu & Wang, Treeformer: Deep Tree-Based Model with Two-Dimensional Information Enhancement for Multivariate Time Series Forecasting (Liu and Wang, 2025)	□	□	□	□	□	□	□	The methodology describes a hybrid approach using tree-based encoders and Transformer-based forecasting but no demonstration or statement that all featu "point-in-time". Normalisation vaguely discussed and methods not described.
Liu et al., Dynamic Financial Stress Index Modeling with IGSA-MLPNN Fusion (Liu et al., 2025a)	□	□	□	□	□	□	□	No statement that all inputs to IGSA-MLPNN are point-in-time or that dynami weighting and forecasting avoid use of future information in feature constructi training. No description of any scaling/normalization, nor any assurance that, i parameters are fit only on the training window.
Liu et al., Combining deep learning with econometric models: volatility forecasting using the KAN-GARCH-MIDAS framework (Liu et al., 2025b)	□	□	□	□	□	□	□	No explicit point-in-time assurance for KAN inputs or MIDAS long-run compo explicit statement that only past macro and return data are used when forecasti description of any scaling/normalization of returns or macro variables or confi that, if used, statistics are fit only on the training window.
López-Herrera et al., Directional forecasting for eight forex pairs against the US dollar using machine learning techniques (López-Herrera et al., 2025)	□	□	□	□	□	□	□	author use macroeconomic indicators (Treasury bill rates, yield curve slope). H the paper does not explicitly prohibit or state that it avoided the use of revised/ macroeconomic data. No evidence that scaling parameters (mean, SD, min, ma fitted exclusively on the training window
Mohamed et al., Enhancing EUR/USD Exchange Rate Predictions Using AIRS and ESN Models with Stacking (Mohamed et al., 2025)	□	□	□	□	□	□	□	No explicit guarantee that lagged indicator, PCA, and SMOTE features are stri point-in-time and do not use future information when forming training exampl description of how features are scaled/normalized or whether any scaler/PCA p are fit only on the training window and then applied to validation/test: Reports ratios but does not specify trading rules, frictions, or full backtest standard.
Narmandakh et al., A 2D-CNN-LSTM-Based Deep Learning Model for Forex Price Prediction using Lag Features (Narmandakh et al., 2025)	□	□	□	□	□	□	□	No explicit guarantee that lag-k features, technical indicators, and input matric constructed strictly point-in-time with only past data, or that label horizons do into feature computation. Scaling over the “training features” assumed that the statistics are not recomputed including test data or during multi-asset experime
Nayak et al., Predicting Forex Trends: A Comprehensive Analysis of Supervised learning	□	□	□	□	□	□	□	No description of any scaling/normalization of inputs, therefore, it is assumed was applied, nor that any scaler was fit. S5: Data split is 70/30 but split is not e by date; no statement that shuffling is disabled, so chronological nature of the s

in Exchange Rate Prediction (Nayak et al., 2025)								ambiguous.
Njoki et al., Ensemble Learning for FX Trend Prediction (Njoki et al., 2025)	?	?	?	?	?	?	?	Authors state 'Important features are extracted using CNN.' without stating that done point-in-time without using future information. No description of scaling/normalization considered pass here. Train/test splits are described only proportions; no validation set described; not clear how they used k-fold validat
Papatsimpas et al. Hybrid decomposition and deep learning approach for data-driven FOREX forecasting optimization (Papatsimpas and Parsopoulos, 2025)	?	?	?	?	?	?	?	EMD is run on all 5000 data including the forecast day, so decomposition uses data relative to earlier timestamps. Min–max scaling is defined on each full IM without stating that min/max are estimated only on the training window or via forward updates.
Rafi et al. AI-Driven Currency Arbitrage Optimization System and Correlation Risk Using Statistical Analysis (Rafi et al., 2025)	?	?	?	?	?	?	?	Authors state, "The machine learning models require scaled data thus we appli Max Scaling to normalize our inputs for creating stable numerical data. The data underwent final processing by splitting it into training and testing sections.", which is future information leak
Rahat et al. Advancing Exchange Rate Forecasting: Leveraging Machine Learning and AI for Enhanced Accuracy in Global Financial Markets (Rahat et al., 2025)	?	?	?	?	?	?	?	Authors state, Yahoo Finance provided the historical USD/BDT exchange rate data from 2018 to 2023, which was then preprocessed with MinMaxScaler and divided into 80% training and 20% testing sets."
Saghafi et al. Forecasting Forex EUR/USD Closing Prices Using a Dual-Input Deep Learning Model with Technical and Fundamental Indicators (Saghafi et al., 2025)	?	?	?	?	?	?	?	No backtesting simulation
San et al., A Hybrid Deep Learning Algorithm for Forecasting of Global Currency Market (San et al., 2025b)	?	?	?	?	?	?	?	The two San et al. papers are highly similar in dataset, problem formulation, ar experimental design, but this manuscript makes an explicit design choice to av data scaling/normalization.
San et al., A Machine Learning Model for Foreign Exchange Prediction (San et al., 2025a)	?	?	?	?	?	?	?	Preprocessing mentions normalization only in a generic description of data mi there is no explicit description of any scaling/normalization actually applied to inputs, nor of fitting scalers only on training data. Data are split 80:20, but ther explicit statement that the split is strictly chronological (by date) rather than ra cross-validation is described without specifying a walk-forward or time-series- scheme.
Sanaei et al., Enhancing Forex Market Forecasting with ConvLSTM2D (Sanaei and Daneshvar, 2025)	?	?	?	?	?	?	?	The dataset is split into 75% training, 12.5% validation, and 12.5% test but fail explicitly say whether the split was chronological or random.
Saurabh et al., Autoencoder-LSTM Framework for Accurate Foreign Exchange Volatility Prediction (Saurabh et al., 2025)	?	?	?	?	?	?	?	Autoencoder was trained but no mention of fit on training window or full datas statement that Min-Max (or any scaler) is fit only on the training set and then a validation/test sets.This paper provides no percentages for the split and suggest "possible to randomize the training data," which risks compromising the tempo structure
Song et al., Advanced Forecasting of Financial Data with Neural Networks (Song and Yang, 2025)	?	?	?	?	?	?	?	The study uses “a specific stock,” but does not name the ticker or symbol. Lab infer from the statement, "forcast the upcoming value B". Discussion of feature engineering and normalization are vague. No chron split described.
Uygun et al., Financial asset price prediction with graph neural network-based temporal deep learning models (Uygun and Sefer, 2025)	?	?	?	?	?	?	?	normalization is used but never explicitly restricted to training/causal window
Wangchailert & Paireekreng, Enhancing FOREX Market Predictions: A Comparative Study of Candlestick Patterns and the MIDDAM Patterns (Wangchailert and Paireekreng, 2025)	?	?	?	?	?	?	?	Features are candle patterns, MIDDAM was designed looking at full datatset. N machine learning implemented and hence no normalisation was performed. Ba performed based on candle patterns.
Wanna & Porntrakoon, Multi-Source Financial Data Integration via LSTM-GRU-	?	?	?	?	?	?	?	The paper uses macroeconomic data (GDP, CPI, interest rates) There is no exp demonstration that all features are "point-in-time." Specifically, the paper does that the model avoids revised/backdated macroeconomic data. The paper fails t

Attention Models for Robust Forex Predictions (Wanna and Porntrakoon, 2025)								that normalization parameters (min/max) were fitted exclusively on the training chronological split not confirmed.
Xu et. al., Forex forecasting: The critical role of feature selection (Xu et al., 2025)	☐	☐	☐	☐	☐	☐	☐	Fails on outlier handling, full-sample PCA decomposition, non-causal scaling, splitting, and lack of economic backtesting.

11 Comparison with other frameworks

Work on financial time-series evaluation has begun to converge on shared datasets and model benchmarks, but there is still no field-wide standard that governs how arbitrary AI-based forecasting and trading papers should design experiments, report results, and control for leakage. Existing efforts fall into three main categories: (i) financial time-series benchmarks such as FinTSB (Hu, 2025) and FinTSBridge (Wang et al., 2025) curate datasets and model baselines; (ii) domain-specific microstructure benchmarks such as LOBCAST (Prata et al., 2024) and DeepLOB-style suites (Zhang et al., 2019); and (iii) general time-series benchmarks and generation frameworks such as TSFM-Bench (Li et al., 2025) and TSGBench (Ang et al., 2023). All of these are dataset-centred benchmarks: they tell you how to run models on their benchmark, but they do not prescribe a reviewer-enforceable standard for financial ML papers, nor do they couple compliance to a public leaderboard. QFRS differs from all of these in that it is a standard and reporting standard intended to be enforced by reviewers and editors across the entire AI finance assets price forecasting literature, independent of any particular benchmark dataset (Hu, 2025; Wang et al., 2025).

On the trading side, several benchmarks now report some risk-adjusted performance, but none elevate backtesting to a universal publishing requirement. FinTSB and FinTSBridge include risk-oriented evaluation, yet their primary focus remains forecast accuracy and financial IC/IR-type metrics within their own suites rather than a general backtesting schema with mandated annualised returns, drawdowns and Sharpe ratios for arbitrary strategies (Wang et al., 2025; Zhang et al., 2025). LOBCAST and DeepLOB mostly report predictive metrics (e.g. accuracy, F1, AUC) and task-level profitability proxies, without a standard layer for costed portfolio backtests across assets. Broader time-series benchmarks such as TSFM-Bench and TSGBench rarely require portfolio-level backtests outside a few finance tracks, while finance-focused benchmarks on crude-oil volatility and cryptocurrency generation incorporate trading performance only within their specific datasets (Pasula, 2023; Ang et al., 2025; Ang et al., 2023). In contrast, QFRS explicitly mandates cost- and slippage-aware backtests with cumulative and annualised returns, volatility, maximum drawdown, and at least one risk-adjusted ratios (Sharpe Ratio) as part of the eligibility criteria for its leaderboard, and ties these requirements to structured reporting tables and a reviewer checklist.

QFRS is not another benchmark dataset, but a domain-specific, leakage-aware, economically grounded standard that can sit on top of FinTSB, LOBCAST, TSFM-Bench or any custom dataset, and that defines what must be done and reported before backtest-based claims can be trusted, compared and placed on a public leaderboard.

12 Discussion

There is now substantial evidence that such a standard is necessary to prevent the literature from being cluttered with fragile or non-replicable results. Research on ML pipelines shows that preprocessing leakage, especially fitting scalers on train+test, affects hundreds of published papers and can invalidate reported performance (Kapoor and Narayanan, 2023; Sasse et al., 2025).

In financial forecasting, full-sample signal decompositions and random or K-fold splits have been shown to produce dramatic RMSE improvements (up to 100×) that disappear under proper chronological splits and window-bounded preprocessing (Yang et al., 2024; Wen et al., 2024). Benchmarks such as FinTSB, TSFM-Bench and TSGBench explicitly complain that ad-hoc datasets, non-standard splits and inconsistent metrics make it impossible to tell whether a new model is genuinely better or just evaluated under more favourable conditions (Wang et al., 2025; Ang et al., 2023; Liu et al., 2024). Other fields have responded to analogous problems by adopting formal reporting standards: in clinical prediction, TRIPOD and TRIPOD+AI now specify minimum items that must be reported for AI-based risk-prediction models and are increasingly required by journals (Collins et al., 2015; Collins et al., 2024).

Section 9.1 makes explicit that once a sub-criterion is applicable, it must be coded either Pass or Fail; partial compliance is not permitted. This rule is stated both in the explanatory text and in the mathematical formulation of

the aggregation logic. The rationale for this conservative design is practical rather than cosmetic. In quantitative finance, apparently strong theoretical or backtested results often fail when exposed to live trading conditions, particularly once transaction costs, slippage, and execution frictions are properly incorporated. Bailey and López de Prado have shown that backtest overfitting can easily produce false discoveries that do not survive real-market deployment (Bailey and de Prado, 2021; Lopez de Prado, 2022)., while studies on transaction costs document that many apparently profitable strategies become untradeable after realistic frictions are included (Novy-Marx and Velikov, 2016; DETZEL et al., 2023).

This issue is not hypothetical. In prior work, Zhang and Khushi (2020) (Zhang and Khushi, 2020) reported very strong backtested profitability on 5-minute AUD/USD data for 2019, but transaction costs were not incorporated in the reported results. When the same strategy logic was subsequently tested in live market conditions, it lost money rather than generating the theoretical annualised profit suggested by the backtest. This illustrates why partial compliance is not an acceptable standard for methodological auditing in quantitative finance. A study that omits a single critical safeguard such as cost-aware backtesting, leakage-free scaling, or chronologically valid splitting may still look impressive in theory while being invalid in practice. For that reason, the framework treats any failed applicable sub-criterion as sufficient to fail the corresponding criterion as a whole. A paper is therefore either compliant or non-compliant with an applicable criterion; there is no intermediate category that would risk legitimising practically non-deployable results.

12.1 Rationale for a unified leaderboard

Leaderboards have become a standard way to organise progress in other areas of AI: language and vision now rely on public rankings such as artificialanalysis.ai, Chatbot Arena and similar platforms to give practitioners a single, continuously updated view of which models perform best under common conditions (Chiang et al.; Vertogradov and Shchelokova, 2025). In contrast, AI-based financial asset prediction has no equivalent: every year, thousands of papers get published with non-comparable metrics, leaving both researchers and investors without a clear answer to a basic question, '*which model, evaluated under realistic risk and cost assumptions, is best?* '. QFRS leaderboard precisely to fill this gap: to provide a transparent, standard-compliant ranking where models are evaluated on fully specified, cost-aware backtests with risk-adjusted performance and drawdown statistics, so that investors and practitioners can make informed allocation and deployment decisions rather than navigating an opaque, cluttered literature.

QFRS-compliant papers will be ranked solely on backtesting performance metrics, and only studies that provide a verifiable dataset and executable code (or equivalent reproducible implementation) will be eligible for inclusion on the leaderboard.

12.2 Future Work

A domain-specific, LLM-based screening tool could be used to support the maintenance of the leaderboard in a scalable way. Recent work on LLM-assisted systematic reviews (Syriani et al., 2024) and automated peer review shows that large models can extract structured methodological information from scientific articles and evaluate them against pre-specified checklists, providing a template for finding potential compliant papers. Building on these advances, we envisage fine-tuning a domain-specialised LLM that reads financial ML papers (and their code when available), reconstructs the experimental pipeline (data, labels, features, scaling, splits, metrics, and costs), and flags standard complaint papers (Wang et al., 2024a; Quttainah et al., 2024; Yu et al., 2024; Zhu et al., 2025). The scoring specification in Section 9.1, particularly Equations (1)–(3), provides the machine-readable input required for such an LLM-based screening tool. Because each sub-criterion is coded as a ternary value (Pass, Fail, NA) and standard-level compliance is derived deterministically from these codes, the underlying logic can be directly translated into a structured prompt template or a classification head in a fine-tuned model without requiring subjective interpretation. Papers that pass human verification of all mandatory checks will have their reported metrics and backtest results featured on the leaderboard.

Acknowledgement

MK is supported by UKRI–NERC grant UKRI4338. The author thanks Professor Marcos López de Prado, Professor George Ghinea, and the anonymous reviewers for their feedback, which helped improve the quality of this paper. The author also thanks Sara Muzaffar Qureishi (Brunel MSc student) for validating the empirical testing.

References

ABBASZADEH, A., HOSSEINI, S. A. & AKBARZADEH-T, M. R. A Synergistic Hybrid Architecture with Residual Attention and Mixture-of-Experts for Robust Hour-Ahead Forex Forecasting. 2025 15th International Conference on Computer and Knowledge Engineering, ICCKE 2025, 2025.

AHMAD, F., SUN, Y., WANG, L. & WANG, Y. Foreign Exchange Prediction and Trading Using Low-Shot Machine Learning. Lecture Notes in Networks and Systems, 2025. 111-147.

AHMED, H. 2025. Leveraging Machine Learning for Exchange Rate Prediction: Fresh Insights from BRICS Economies. *International Journal of Economics and Financial Issues,* 15**,** 72-79.

AL-QURAISHI, Y., MANHOSH, M. M., AL-KHAMEES, H. A. A., SRAYYIH, F. H., ABED, M. K., AL-SHARIFY, T. A., HAMAD, M. T., KADHIM, R. I. & HASHIM, W. A. AI-Enhanced Models for Investment Risk Prediction in Financial Markets. 3rd International Conference on Business Analytics for Technology and Security, ICBATS 2025, 2025.

ALMGREN, R. & CHRISS, N. 2001. Optimal execution of portfolio transactions. *Journal of Risk,* 3**,** 5-40.

ALOMAR, E. A., DEMARIO, C., SHAGAWAT, R. & KREISER, B. Leakagedetector: An open source data leakage analysis tool in machine learning pipelines. 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), 2025. IEEE, 844-849.

ALQODRI, F., WIDIYANINGTYAS, T. & PRASETYA, D. D. Comparative Analysis of ARIMA, Random Forest, and BiLSTM for Dollar Index Prediction in the 21st Century. 2025 9th International Conference on Electrical, Electronics and Information Engineering, ICEEIE 2025, 2025.

ANDERSEN, T., BOLLERSLEV, T., DIEBOLD, F. & LABYS, P. 2000. Great realisations. *RISK-LONDON-RISK MAGAZINE LIMITED-,* 13**,** 105-109.

ANDERSON, T. G., BOLLERSLEV, T., DIEBOLD, F. X. & VEGA, C. 2003. Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American economic review,* 93**,** 38-62.

ANG, Y., HUANG, Q., BAO, Y., TUNG, A. K. H. & HUANG, Z. TSGBench: Time Series Generation Benchmark. Proceedings of the VLDB Endowment, 2023. 305-318.

ANG, Y., WANG, Q., HUANG, Q., BAO, Y., XI, X., TUNG, A. K. H., JIN, C. & HUANG, Z. 2025. CTBench: Cryptocurrency Time Series Generation Benchmark.

ANTE, L., FIEDLER, I. & STREHLE, E. 2021. The impact of transparent money flows: Effects of stablecoin transfers on the returns and trading volume of Bitcoin. *Technological Forecasting and Social Change,* 170**,** 120851.

ARNOLD, T., FARIZO, J. & NORTH, D. S. 2025. Facilitating Portfolio Construction Using Ex-Ante Information Ratios and Maximum Sharpe Ratios. *Journal of Wealth Management,* 28**,** 35-41.

BAHOO, S., CUCCULELLI, M., GOGA, X. & MONDOLO, J. 2024. Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics,* 4**,** 23.

BAILEY, D. H. & DE PRADO, M. L. 2021. How “Backtest Overfitting” in Finance Leads to False Discoveries. *Significance,* 18**,** 22-25.

BAILEY, D. H. & LÓPEZ DE PRADO, M. 2014. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *Journal of Portfolio Management,* 40**,** 94-107.

BAILEY, D. H. & LÓPEZ DE PRADO, M. 2021. How "Backtest Overfitting" in Finance Leads to False Discoveries. *Significance,* 18**,** 22-25.

BASILE, I. 2024. Hedge funds. *Asset Management and Institutional Investors: Methods and Tools for Asset Allocation, Portfolio Management and Performance Evaluation.* Springer.

BENHENDA, M. 2025. FinRL-DeepSeek: LLM-infused risk-sensitive reinforcement learning for trading agents. *arXiv preprint arXiv:2502.07393*.

BERNARDINI, M. S. & DE CASTRO, P. A. L. 2021. Is it a great Autonomous FX Trading Strategy or you are just fooling yourself. *arXiv preprint arXiv:2101.07217*.

BHUIYAN, M. S. M., RAFI, M. A., RODRIGUES, G. N., MIR, M. N. H., ISHRAQ, A., MRIDHA, M. & SHIN, J. 2025. Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies. *Array,* 26**,** 100390.

BRACHER, J., RAY, E. L., GNEITING, T. & REICH, N. G. 2021. Evaluating epidemic forecasts in an interval format. *PLoS computational biology,* 17**,** e1008618.

BRINZA, A., IOAN, V. & LAZARESCU, I. 2023. Critical analysis of the sharpe ratio: Assessing performance and risk in financial portfolio management. *Ovidius University Annals, Economic Sciences Series,* 23**,** 633-639.

BRIOLA, A., BARTOLUCCI, S. & ASTE, T. 2025. HLOB–Information persistence and structure in limit order books. *Expert Systems with Applications,* 266**,** 126078.

BROWNEN-TRINH, R. 2019. Effects of winsorization: The cases of forecasting non-GAAP and GAAP earnings. *Journal of Business Finance & Accounting,* 46**,** 105-135.

BROWNLEES, C. T. & GALLO, G. M. 2006. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational statistics & data analysis,* 51**,** 2232-2245.

CERQUEIRA, V., ROQUE, L. & SOARES, C. Forecasting with deep learning: Beyond average of average of average performance. International Conference on Discovery Science, 2024. Springer, 135-149.

CERQUEIRA, V., TORGO, L. & MOZETIČ, I. 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning,* 109**,** 1997-2028.

CHEN, L., LIU, S., YAN, J., WANG, X., LIU, H., LI, C., JIAO, K., YING, J., LIU, Y. V. & YANG, Q. 2025. Advancing financial engineering with foundation models: progress, applications, and challenges. *Engineering*.

CHI, D. T. K., KIEN, H. N. T. & NGUYEN, T. Q. 2025a. Enhancing forex market forecasting with feature-augmented multivariate LSTM models using real-time data. *Knowledge-Based Systems,* 330.

CHI, D. T. K., LAN, L. T. H. & LE MINH HOA, N. Advanced Predictive Modeling of Forex Market Fluctuations Using Ensembles Learning. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 2025b. 202-211.

CHIANG, W.-L., ZHENG, L., SHENG, Y., ANGELOPOULOS, A. N., LI, T., LI, D., ZHU, B., ZHANG, H., JORDAN, M. & GONZALEZ, J. E. Chatbot arena: An open platform for evaluating llms by human preference. Forty-first International Conference on Machine Learning.

CHICCO, D. & JURMAN, G. 2023a. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData mining,* 16**,** 4.

CHICCO, D. & JURMAN, G. 2023b. A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index. *Journal of Biomedical Informatics,* 144**,** 104426.

COLLINS, G. S., MOONS, K. G., DHIMAN, P., RILEY, R. D., BEAM, A. L., VAN CALSTER, B., GHASSEMI, M., LIU, X., REITSMA, J. B. & VAN SMEDEN, M. 2024. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj,* 385.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery,* 102**,** 148-158.

CROUSHORE, D. & STARK, T. 2001. A real-time data set for macroeconomists. *Journal of econometrics,* 105**,** 111-130.

DAKALBAB, F., KUMAR, A., TALIB, M. A. & NASIR, Q. 2025. Advancing Forex prediction through multimodal text-driven model and attention mechanisms. *Intelligent Systems with Applications,* 26.

DAS, S. R., MOHANTY, A. K., MISHRA, D. & PANDA, J. K. 2025. ELM-GA-ERWCA: A Tailored Forex Exchange Trading Model Combining Extreme Learning Machine with Genetic Algorithm and Evaporation Based Water Cycle Algorithm. *International Journal of Computational Intelligence Systems,* 18.

DAVE, Y., VARASTEHPOUR, S. & SHAKIBA, M. Forex Price Prediction: A Multi-model Approach Integrating Sentiment Analysis Using LLMs with LSTM, XGBoost, Transformer Models. Lecture Notes in Networks and Systems, 2025a. 229-243.

DAVE, Y., VARASTEHPOUR, S. & SHAKIBA, M. Predicting Forex Prices: An Evaluation of Long Short-Term Memory, XGBoost and Transformer Architectures. EECT 2025 - 2025 5th International Conference on Advances in Electrical, Electronics and Computing Technology, Proceeding, 2025b.

DAVIS, J. & GOADRICH, M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, 2006. 233-240.

DE CASTRO, P. A. L. 2021. mt5se: An Open Source Framework for Building Autonomous Trading Robots. *arXiv preprint arXiv:2101.08169*.

DE PRADO, M. L. 2018. *Advances in financial machine learning*, John Wiley & Sons.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, 2009. Ieee, 248-255.

DETZEL, A., NOVY-MARX, R. & VELIKOV, M. 2023. Model Comparison with Transaction Costs. *The Journal of Finance,* 78**,** 1743-1775.

DINLER, A. 2025. A Review of Balancing Price Forecasting in the Context of Renewable-Rich Power Systems, Highlighting Profit-Aware and Spike-Resilient Approaches. *Energies,* 18**,** 6460.

DIXON, M. F., HALPERIN, I. & BILOKON, P. 2020. *Machine learning in finance*, Springer.

DUONG, N. T., HOANG, K. T., DUONG, K. Q., DINH, D. Q., LE, H. D., NGUYEN, T. H., DUONG, B. X., TRAN, B. Q. & BUI, A. N. Investigating Market Strength Prediction with CNNs on Candlestick Chart Images. ACMLC 2024 - 2024 6th Asia Conference on Machine Learning and Computing, 2025. 120-127.

GADALAY, R., SAI SHASANK, V. C., CHAITRA, N. S. & DAS, S. R. Enhancing Forex Market Predictive Analytics with ELM and Whale Optimization Algorithm. ICoICC 2025 - 3rd International Conference on Intelligent and Cloud Computing, 2025.

GHAHREMANI, S. & NGUYEN, U. T. 2025. Prediction of foreign currency exchange rates using an attention-based long short-term memory network. *Machine Learning with Applications,* 20.

GNEITING, T. & RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association,* 102**,** 359-378.

GOLDBERG, L. R. & MAHMOUD, O. 2017. Drawdown: from practice to theory and back again. *Mathematics and Financial Economics,* 11**,** 275-297.

GORT, B. J. D., LIU, X.-Y., SUN, X., GAO, J., CHEN, S. & WANG, C. D. 2022. Deep reinforcement learning for cryptocurrency trading: Practical approach to address backtest overfitting. *arXiv preprint arXiv:2209.05559*.

GORTON, G. & ROUWENHORST, K. G. 2006. Facts and fantasies about commodity futures. *Financial Analysts Journal,* 62**,** 47-68.

GU, S., KELLY, B. & XIU, D. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies,* 33**,** 2223-2273.

HAIDER, M. R. A., SONI, S. & SAH, S. Enhancing Forex Market Predictions with a Hybrid Prophet-LSTM Model. Lecture Notes in Networks and Systems, 2025. 197-205.

HARVEY, C. R. & LIU, Y. 2015. Backtesting. *The Journal of Portfolio Management,* 42**,** 13-28.

HASSAAN, Z. A., YACOUB, M. H. & SAID, L. A. 2025. FPGA-Accelerated ESN with Chaos Training for Financial Time Series Prediction. *Machine Learning and Knowledge Extraction,* 7.

HEWAMALAGE, H., ACKERMANN, K. & BERGMEIR, C. 2023. Forecast evaluation for data scientists: common pitfalls and best practices: H. Hewamalage et al. *Data Mining and Knowledge Discovery,* 37**,** 788-832.

HTUN, H. H., BIEHL, M. & PETKOV, N. 2023. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation,* 9**,** 26.

HU, Y. 2025. FinTSB: A Comprehensive and Practical Benchmark for Financial Time Series Forecasting. *arXiv preprint arXiv:2502.18834*.

HURST, B., OOI, Y. H. & PEDERSEN, L. H. 2013. Demystifying managed futures. *Journal of Investment Management,* 11**,** 42-58.

HYNDMAN, R. J. & KOEHLER, A. B. 2006. Another look at measures of forecast accuracy. *International journal of forecasting,* 22**,** 679-688.

INCE, O. S. & PORTER, R. B. 2006. Individual equity return data from Thomson Datastream: Handle with care! *Journal of Financial Research,* 29**,** 463-479.

ISWARA, B. D., SUGIARTAWAN, P. & KUSUMA, A. S. Integration of Conformal Prediction in Automated Forex Trading Systems: Development and Evaluation. Digest of Technical Papers - IEEE International Conference on Consumer Electronics, 2025. 389-394.

IZADI, M. & YAGHOOBI, M. A. 2024. Portfolio optimization based on bi-objective linear programming. *Rairo-Operations Research,* 58**,** 713-739.

K.C, M. & LAHA, A. K. 2021. A Robust Sharpe Ratio. *Sankhya B,* 83**,** 444-465.

KAPOOR, S. & NARAYANAN, A. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y),* 4**,** 100804.

KERCHEVAL, A. N. & ZHANG, Y. 2015. Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance,* 15**,** 1315-1329.

KHANSALAR, E. 2025. The Usefulness of the Currency Constraint for Predicting Forex. *International Journal of the Economics of Business,* 32**,** 393-409.

KHANSAMA, R. R., PRIYADARSHINI, R. & NANDA, S. K. Predicting FOREX trend using incremental spiking neural network. ESIC 2025 - 5th International Conference on Emerging Systems and Intelligent Computing, Proceedings, 2025. 185-188.

KIM, J., KIM, H., KIM, H., LEE, D. & YOON, S. 2025. A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review,* 58**,** 216.

KIM, T., KIM, J., TAE, Y., PARK, C., CHOI, J.-H. & CHOO, J. Reversible instance normalization for accurate time-series forecasting against distribution shift. International conference on learning representations, 2021.

KOU, G. & LU, Y. 2025. FinTech: a literature review of emerging financial technologies and applications. *Financial Innovation,* 11**,** 1.

LEE, J. E. L., CABALLERO, A. R., CABALLERO, J. M., ALFARO, F. A. & DELA CRUZ, F. L. Analysis of Foreign Exchange Behavior using Sequential Minimal Optimization and Linear Regression. Proceedings - IC2IE 2025: 8th 2025 International Conference of Computer and Informatics Engineering: Human-Machine Synergy Brings Together the Physical and Digital Worlds, 2025.

LI, Z., QIU, X., CHEN, P., WANG, Y., CHENG, H., SHU, Y., HU, J., GUO, C., ZHOU, A., JENSEN, C. S. & YANG, B. 2025. TSFM-Bench: A Comprehensive and Unified Benchmark of Foundation Models for Time Series Forecasting. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2.* Toronto ON, Canada: Association for Computing Machinery.

LIU, F., CHEN, L., ZHENG, Y. & FENG, Y. 2022. A prediction method with data leakage suppression for time series. *Electronics,* 11**,** 3701.

LIU, M., WEN, S., WANG, Z. & ZHANG, Q. Dynamic Financial Stress Index Modeling with IGSA-MLPNN Fusion. Procedia Computer Science, 2025a. 1088-1095.

LIU, T., CHOO, W., XINPING, H. & LI, L. 2025b. Combining deep learning with econometric models: volatility forecasting using the KAN-GARCH-MIDAS framework. *Journal of Applied Economics,* 28.

LIU, X., HU, J., LI, Y., DIAO, S., LIANG, Y., HOOI, B. & ZIMMERMANN, R. Unitime: A language-empowered unified model for cross-domain time series forecasting. Proceedings of the ACM Web Conference 2024, 2024. 4095-4106.

LIU, X. & WANG, W. 2025. Treeformer: Deep Tree-Based Model with Two-Dimensional Information Enhancement for Multivariate Time Series Forecasting. *Mathematics,* 13.

LIVNAT, J. & MENDENHALL, R. R. 2006. Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. *Journal of accounting research,* 44**,** 177-205.

LOMMERS, K., HARZLI, O. E. & KIM, J. 2021. Confronting machine learning with financial research. *arXiv preprint arXiv:2103.00366*.

LONES, M. A. 2024. Avoiding common machine learning pitfalls. *Patterns,* 5.

LONG, X., KAMPOURIDIS, M. & PAPASTYLIANOU, T. 2026. Multi-objective genetic programming-based algorithmic trading, using directional changes and a modified sharpe ratio score for identifying optimal trading strategies. *Artificial Intelligence Review,* 59.

LÓPEZ-HERRERA, F., JIMÉNEZ, J. G. M. & SANTIAGO, A. R. 2025. Directional forecasting for eight forex pairs against the US dollar using machine learning techniques. *Discover Artificial Intelligence,* 5.

LOPEZ DE PRADO, M. 2022. Causal factor investing: Can factor investing become scientific? *Available at SSRN 4205613*.

LÓPEZ DE PRADO, M. 2022. Machine Learning for Econometricians: The Readme Manual. *Journal of Financial Data Science,* 4.

LÖW, R., MAIER-PAAPE, S. & PLATEN, A. 2015. Correctness of backtest engines. *arXiv preprint arXiv:1509.08248*.

MAIER-PAAPE, S. & ZHU, Q. J. 2018. A general framework for portfolio theory. Part II: Drawdown risk measures. *Risks,* 6.

MAKAROV, I. & SCHOAR, A. 2020. Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics,* 135**,** 293-319.

MALEVERGNE, Y. & SORNETTE, D. 2006. *Extreme financial risks: From dependence to risk management*, Springer.

MANSOURI, S. M. & ETEROVIC, D. S. 2023. Machine learning and the cross-section of emerging market corporate bond returns. *Available at SSRN 4632924*.

MCNEIL, A. J., FREY, R. & EMBRECHTS, P. 2015. *Quantitative risk management: concepts, techniques and tools-revised edition*, Princeton university press.

MELIGKOTSIDOU, L., PANOPOULOU, E., VRONTOS, I. D. & VRONTOS, S. D. 2014. A quantile regression approach to equity premium prediction. *Journal of Forecasting,* 33**,** 558-576.

MENG, T. L. & KHUSHI, M. 2019. Reinforcement Learning in Financial Markets. *Data,* 4.

MOHAMED, E. B., BRAHIM, R., SAMIRA, E. M. & MOSTAFA, B. 2025. Enhancing EUR/USD Exchange Rate Predictions Using AIRS and ESN Models with Stacking. *Lecture Notes on Data Engineering and Communications Technologies.*

NARMANDAKH, B., ZHANG, Y., LI, Z. & ANDERSON, P. 2025. A 2D-CNN-LSTM-Based Deep Learning Model for Forex Price Prediction using Lag Features. *Computational Economics*.

NAYAK, R. K., SODHA, M., MISHRA, N., TRIPATHY, S. K., TRIPATHY, R. & PRADHAN, A. K. Predicting Forex Trends: A Comprehensive Analysis of Supervised learning in Exchange Rate Prediction. Communications in Computer and Information Science, 2025. 59-71.

NJOKI, E., WEKESA, J. S. & NJAGI, D. G. 2025. Ensemble Learning for Foreign Exchange Market Trend Prediction. *Computational Economics*.

NONEJAD, N. 2017. Modeling and forecasting aggregate stock market volatility in unstable environments using mixture innovation regressions. *Journal of Forecasting,* 36**,** 718-740.

NOVY-MARX, R. & VELIKOV, M. 2016. A Taxonomy of Anomalies and Their Trading Costs. *The Review of Financial Studies,* 29**,** 104-147.

NTAKARIS, A., MAGRIS, M., KANNIAINEN, J., GABBOUJ, M. & IOSIFIDIS, A. 2018. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting,* 37**,** 852-866.

NTAKARIS, A., MIRONE, G., KANNIAINEN, J., GABBOUJ, M. & IOSIFIDIS, A. 2019. Feature engineering for mid-price prediction with deep learning. *Ieee Access,* 7**,** 82390-82412.

NURBAEV, V., AU, C. H. & CHOU, C.-Y. 2023. When Stablecoin is No Longer Stable-A Case Study on the Failure of TerraUSD.

OPENJA, M., KHOMH, F., FOUNDJEM, A., JIANG, Z. M., ABIDI, M. & HASSAN, A. E. 2024. An empirical study of testing machine learning in the wild. *ACM transactions on software engineering and methodology,* 34**,** 1-63.

PAPATSIMPAS, M. G. & PARSOPOULOS, K. E. 2025. Hybrid decomposition and deep learning approach for data-driven FOREX forecasting optimization. *Journal of Global Optimization*.

PASULA, P. 2023. Real World Time Series Benchmark Datasets with Distribution Shifts: Global Crude Oil Price and Volatility.

PETROPOULOS, F., APILETTI, D., ASSIMAKOPOULOS, V., BABAI, M. Z., BARROW, D. K., TAIEB, S. B., BERGMEIR, C., BESSA, R. J., BIJAK, J. & BOYLAN, J. E. 2022. Forecasting: theory and practice. *International Journal of forecasting,* 38**,** 705-871.

PIPPAS, N., LUDVIG, E. A. & TURKAY, C. 2025. The Evolution of Reinforcement Learning in Quantitative Finance: A Survey. *ACM Comput. Surv.,* 57**,** Article 295.

POH, D. L. B. Z. S. R. S. 2021. Building Cross-Sectional Systematic Strategies by Learning to Rank. *The Journal of Financial Data Science Spring,* 3**,** 70-86.

POLBENNIKOV, S., DESCLÉE, A. & DUBOIS, M. 2021. Practical applications of implementing value and momentum strategies in credit portfolios. *Practical Applications,* 9**,** 1-6.

POON, S.-H. & GRANGER, C. W. J. 2003. Forecasting volatility in financial markets: A review. *Journal of economic literature,* 41**,** 478-539.

PRATA, M., MASI, G., BERTI, L., ARRIGONI, V., COLETTA, A., CANNISTRACI, I., VYETRENKO, S., VELARDI, P. & BARTOLINI, N. 2024. Lob-based deep learning models for stock price trend prediction: a benchmark study. *Artificial Intelligence Review,* 57**,** 116.

QIU, X., HU, J., ZHOU, L., WU, X., DU, J., ZHANG, B., GUO, C., ZHOU, A., JENSEN, C. S. & SHENG, Z. 2024. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*.

QUTTAINAH, M., MISHRA, V., MADAKAM, S., LURIE, Y. & MARK, S. 2024. Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: narrative review and qualitative study. *Jmir Ai,* 3**,** e51834.

RAFI, M. S., SIDDHARTH, N., HRUSHIKESH RAJU, K. & SIDHWA, H. H. AI-Driven Currency Arbitrage Optimization System and Correlation Risk Using Statistical Analysis. 2025 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence, RAEEUCCI 2025, 2025.

RAHAT, M. Y., GUPTA, R. D., RAHMAN, N. R., PRITOM, S. R., SHAKIR, S. R., SHOWMICK, M. I. H. & HOSSEN, M. J. Advancing Exchange Rate Forecasting: Leveraging Machine Learning and AI for Enhanced Accuracy in Global Financial Markets. 2025 Multimedia University Engineering Conference, MECON 2025, 2025.

SAGHAFI, A., BAGHERIAN, M. & SHOKOOHI, F. 2025. Forecasting Forex EUR/USD Closing Prices Using a Dual-Input Deep Learning Model with Technical and Fundamental Indicators. *Mathematics,* 13.

SAN, T. J., AJIBADE, S. S. M., AJIBADE, T. I., JASSER, M. B., AYODELE, O. E., OYEWOPO, A. O., ADEDIRAN, A. O., MAJEED, A. P. P. A. & LUO, Y. A Machine Learning Model for Foreign Exchange Prediction. Lecture Notes in Networks and Systems, 2025a. 578-593.

SAN, T. J., AJIBADE, S. S. M., YONG, Y. L., AJIBADE, T. I., AYODELE, O. E., OYEWOPO, A. O., MAJEED, A. P. P. A. & LUO, Y. A Hybrid Deep Learning Algorithm for Forecasting of Global Currency Market. Lecture Notes in Networks and Systems, 2025b. 672-693.

SANAEI, B. & DANESHVAR, S. 2025. Enhancing Forex Market Forecasting with ConvLSTM2D: A Comprehensive Analysis of Spatiotemporal Dependencies and Data Preprocessing Techniques. *Computational Economics*.

SAQUR, R., KATO, K., VINDEN, N. & RUDZICZ, F. 2024. Nifty financial news headlines dataset. *arXiv preprint arXiv:2405.09747*.

SASSE, L., NICOLAISEN-SOBESKY, E., DUKART, J., EICKHOFF, S. B., GÖTZ, M., HAMDAN, S., KOMEYER, V., KULKARNI, A., LAHNAKOSKI, J. & LOVE, B. C. 2023. On leakage in machine learning pipelines. *arXiv preprint arXiv:2311.04179*.

SASSE, L., NICOLAISEN-SOBESKY, E., DUKART, J., EICKHOFF, S. B., GÖTZ, M., HAMDAN, S., KOMEYER, V., KULKARNI, A., LAHNAKOSKI, J. & LOVE, B. C. 2025. Overview of leakage scenarios in supervised machine learning. *Journal of Big Data,* 12**,** 135.

SAURABH, P., ALAZZAM, M. B., NAGARAJAN, S., REDDY, D. H., SULAIMAN, S. F. & BALA, B. K. Autoencoder-Lstm Framework for Accurate Foreign Exchange Volatility Prediction. Proceedings of 2025 3rd International Conference on Intelligent Systems, Advanced Computing, and Communication, ISACC 2025, 2025. 485-490.

SHEN, L., WEI, Y. & WANG, Y. 2022. Respecting time series properties makes deep time series forecasting perfect. *arXiv preprint arXiv:2207.10941*.

SINGH, S., WALIA, N., BEKIROS, S., GUPTA, A., KUMAR, J. & MISHRA, A. K. 2022. Risk-managed time-series momentum: an emerging economy experience. *Journal of Economics, Finance and Administrative Science,* 27**,** 328-343.

SONG, Y. & YANG, D. Advanced Forecasting of Financial Data with Neural Networks. Lecture Notes in Networks and Systems, 2025. 156-167.

SYRIANI, E., DAVID, I. & KUMAR, G. 2024. Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages,* 80**,** 101287.

TETLOCK, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance,* 62**,** 1139-1168.

UYGUN, Y. & SEFER, E. 2025. Financial asset price prediction with graph neural network-based temporal deep learning models. *Neural Computing and Applications,* 37**,** 25445-25471.

VERTOGRADOV, V. & SHCHELOKOVA, S. 2025. Comparative Competitive Analysis of Generative AI Chatbots in Russia and Worldwide. *Studies on Russian Economic Development,* 36**,** 643-652.

WANG, A., SINGH, A., MICHAEL, J., HILL, F., LEVY, O. & BOWMAN, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP, 2018. 353-355.

WANG, S., SCELLS, H., ZHUANG, S., POTTHAST, M., KOOPMAN, B. & ZUCCON, G. Zero-shot generative large language models for systematic review screening automation. European Conference on Information Retrieval, 2024a. Springer, 403-420.

WANG, Y., WU, H., DONG, J., LIU, Y., WANG, C., LONG, M. & WANG, J. 2024b. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*.

WANG, Y., XU, J., GAO, T., ZHANG, H., HUANG, S.-L., SUN, D. D. & ZHANG, X.-P. 2025. FinTSBridge: A New Evaluation Suite for Real-world Financial Prediction with Advanced Time Series Models. Ithaca: Cornell University Library, arXiv.org.

WANGCHAILERT, P. & PAIREEKRENG, W. 2025. Enhancing FOREX Market Predictions: A Comparative Study of Candlestick Patterns and the MIDDAM Patterns. *ECTI Transactions on Computer and Information Technology,* 19**,** 121-134.

WANNA, K. & PORNTRAKOON, P. 2025. MULTI-SOURCE FINANCIAL DATA INTEGRATION VIA LSTM-GRU-ATTENTION MODELS FOR ROBUST FOREX PREDICTIONS. *ABAC Journal,* 45**,** 93-113.

WEN, H., YUAN, M., WANG, S., LIANG, L. & FU, X. 2024. A multilevel wavelet decomposition network hybrid model utilizing cyclic patterns for stock price prediction. *Complexity,* 2024**,** 1124822.

XIAO, Y., VENTRE, C., WANG, Y., LI, H., HUAN, Y. & LIU, B. 2025. LiT: limit order book transformer. *Front Artif Intell,* 8**,** 1616485.

XU, Z., WU, T., ZHOU, Y. & LI, D. 2025. Forex forecasting: The critical role of feature selection. *Finance Research Letters,* 86.

XUE, P. & YE, Y. 2026. Attention-enhanced reinforcement learning for dynamic portfolio optimization. *Intelligent Systems with Applications,* 29.

YANG, C., BROWER-SINNING, R. A., LEWIS, G. & KÄSTNER, C. Data leakage in notebooks: Static detection and better processes. Proceedings of the 37th IEEE/ACM international conference on automated software engineering, 2022. 1-12.

YANG, X., LI, J. & JIANG, X. 2024. Research on information leakage in time series prediction based on empirical mode decomposition. *Sci Rep,* 14**,** 28362.

YE, W., DENG, S., ZOU, Q. & GUI, N. 2024. Frequency adaptive normalization for non-stationary time series forecasting. *Advances in Neural Information Processing Systems,* 37**,** 31350-31379.

YU, J., DING, Z., TAN, J., LUO, K., WENG, Z., GONG, C., ZENG, L., CUI, R., HAN, C. & SUN, Q. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. Findings of the Association for Computational Linguistics: EMNLP 2024, 2024. 10164-10184.

ZABOLOTSKYY, M., ZABOLOTSKYY, T. & TSIAPA, O. 2025. Statistical Analysis of the Sharpe Ratio for the Minimum Value-at-Risk Portfolio. *Journal of Mathematical Sciences***,** 1-24.

ZHANG, F., GAO, H. & YUAN, D. 2024. The asymmetric effect of G7 stock market volatility on predicting oil price volatility: Evidence from quantile autoregression model. *Journal of Commodity Markets,* 35**,** 100409.

ZHANG, J., LIU, S., ZHANG, T. & SHI, Y. 2025. A survey on large language model-based alpha mining. *Frontiers of Information Technology & Electronic Engineering,* 26**,** 1809-1821.

ZHANG, Z. & KHUSHI, M. GA-MSSR: Genetic Algorithm Maximizing Sharpe and Sterling Ratio Method for RoboTrading. 2020 International Joint Conference on Neural Networks (IJCNN), 2020. IEEE, 1-8.

ZHANG, Z., ZOHREN, S. & ROBERTS, S. 2019. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing,* 67**,** 3001-3012.

ZHONG, X., WEI, J., LI, S. & XU, Q. 2025. Deep reinforcement learning for dynamic strategy interchange in financial markets: X. Zhong et al. *Applied Intelligence,* 55**,** 30.

ZHU, M., WENG, Y., YANG, L. & ZHANG, Y. Deepreview: Improving llm-based paper review with human-like deep thinking process. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025. 29330-29355.

ZU, Y., MI, J., SONG, L., LU, S. & HE, J. Finformer: A static-dynamic spatiotemporal framework for stock trend prediction. 2023 IEEE International Conference on Big Data (BigData), 2023. IEEE, 1460-1469.