risks

MDPI

Article

Task-Aware Exchange Rate Forecasting: A Unified Empirical Framework for Level, Return, and Volatility Prediction

Shamaila Butt [1], Muhammad Ali Chohan [2], Mohammad Abrar [3,4], Karima Sayari [2] and Shahid Kamal [3,*]

[1] Faculty of Business, Sohar University, Sohar 311, Oman; sbutt@su.edu.om
[2] Managerial and Financial Sciences Department, Al Zahra College for Women, Muscat 111, Oman; ali.chohan@hotmail.com (M.A.C.); karima@zcw.edu.om (K.S.)
[3] Center for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Parsiaran Multimedia, Cyberjaya 63100, Selangor, Malaysia; abrar.m@aou.edu.om
[4] Faculty of Computer Studies, Arab Open University, Muscat 130, Oman
* Correspondence: shahid.kamal@mmu.edu.my

Abstract

Exchange rate forecasting is a core issue in empirical finance, but in the majority of studies, the impact of target construction, feature representation, and model structure is not disaggregated, and the evaluation is confined to one prediction problem. The three tasks are: (1) prediction of ER_{t+1}; (2) prediction of $r_{t+1} = ER_{t+1} - ER_t$; and (3) prediction of future volatility defined as the sample standard deviation of $r_{t+1}, \ldots, r_{t+10}$, where the horizon is the next 10 recorded observations. The seven model families are evaluated with a strict chronological 70%/15%/15% split; ER is excluded as a direct raw predictor, deep models use seeds 7, 42, and 123, and task-specific financial benchmarks are included. The findings reveal markedly different out-of-sample forecasting behavior across the three tasks. For level prediction, the no-change random walk gives RMSE = 0.0005923 and $R^2 = 0.99815$, while Linear Regression gives RMSE = 0.0005939 and $R^2 = 0.99814$; the difference is not significant (DM = 0.480, $p = 0.631$). For return prediction, Linear Regression gives RMSE = 0.0005920 and $R^2 = -0.00061$ and does not significantly outperform the zero-return benchmark (DM = -0.083, $p = 0.934$). For future volatility, the best average finance-aware learned model is FeatureAttention_Only (RMSE = 0.0004407 ± 0.0000180; $R^2 = -0.027 \pm 0.085$), and its three-seed ensemble gives RMSE = 0.0004298 and $R^2 = 0.024$. FeatureAttention_Only improves squared-error loss relative to EWMA and GARCH in the fixed hold-out, but is not significantly superior to GARCH under QLIKE; rankings also vary across forecast origins.

Keywords: exchange rate prediction; return prediction; volatility prediction; LSTM; GRU; dual-stage attention; finance-aware feature engineering; deep learning; time-series forecasting; ablation analysis

Academic Editors: Difang Huang and Jue Wang

Received: 31 July 2026
Revised: 29 August 2026
Accepted: 3 September 2026
Published: 8 September 2026

1. Introduction

One of the most difficult issues in time-series analysis of finance is exchange rate forecasting. The exchange-rate series are noisy, nonstationary and usually regime-dependent due to the determination of currency values through trading activity, macroeconomic fundamentals, market expectations, institutional policy, and geopolitical events (Frömmel et al. 2022; Rossi 2013). In spite of these challenges, precise forecasting is very vital in monetary planning, trade settlement, corporate treasury operations, portfolio hedging and financial risk management (Alvarez et al. 2007; Poon and Granger 2003). One endemic

issue of the literature is that the entire idea of exchange-rate forecasting is presented as one undifferentiated job. There are three fundamentally different prediction tasks based upon the same underlying exchange-rate series, which have varying statistical characteristics and varying predictability: Exchange Rate Prediction, the next raw exchange rate level, which takes advantage of the strong autocorrelation of the level series; Exchange Rate Return Prediction, prediction of the subsequent first-difference return, eliminating persistence and examining whether signed changes in the future are predictable based on known information (Engel and West 2005; Meese and Rogoff 1983); and Exchange Rate Volatility Prediction, prediction of future volatility of returns, which aims at the magnitude of fluctuations and is linked to risk behavior such as volatility clustering and volatility persistence (Bollerslev 1986; Engle 1982).

These three activities are hardly compared within the same evaluation system. As a result, there is conflicting material in the literature: studies targeted at raw level prediction usually find high model performance due to autocorrelation; studies targeted at returns usually find that models perform poorly; and studies targeted at volatility have mixed results depending on whether informative features have been used. Deep learning models, including LSTM (Hochreiter and Schmidhuber 1997), GRU (Cho et al. 2014), and attention-based architectures (Bahdanau et al. 2015; Vaswani et al. 2017), have been increasingly applied to financial forecasting, with prior studies reporting performance improvements in specific forecasting settings (Fischer and Krauss 2018; Iqbal et al. 2024). However, these results do not establish systematic superiority across different exchange-rate forecasting targets.

The current research addresses this evaluation gap by comparing seven model families across the three targets under the same chronological evaluation, target-appropriate benchmarks, and forecast-loss inference. The novelty, is therefore, the controlled task-aware empirical evaluation framework rather than a new neural architecture.

The objectives of this study are:

- To evaluate the performance of classical machine learning and deep learning models under a common experimental setup.
- To examine how target choice affects predictability in exchange-rate forecasting.
- To test whether finance-informed feature engineering improves volatility forecasting performance.
- To identify which models and input representations are most suitable for different foreign-exchange prediction tasks.

The objectives lead to the following main contributions:

1. A controlled comparison of level, return, and genuinely future-volatility targets under a common chronological design.
2. Task-specific random-walk, zero-return, historical-volatility, EWMA, and GARCH benchmarks.
3. Forecast-loss inference, multiple deep-learning seeds, and expanding-window forecast origins.
4. A causal finance-aware volatility representation with attention ablation and permutation-importance diagnostics.
5. Empirical evidence that conclusions about model performance depend on the target, information set, benchmark, and loss function.

The remainder of this paper is structured as follows. Section 2 presents a literature review of related work on exchange-rate prediction, machine learning, and deep learning techniques. Section 3 explains the dataset, feature construction, forecasting tasks, models, and experimental design. Section 4 presents the empirical findings and performance comparisons among the considered models. Section 5 discusses the major findings and

their implications. Finally, Section 6 summarizes the study and outlines potential avenues for future research.

2. Related Work

2.1. Exchange Rate Predictability

The seminal finding of Meese and Rogoff (1983) that structural economic models can no longer perform better than a random walk out of sample forms the cornerstone of the exchange-rate forecasting literature. This was confirmed by Rossi (2013) in a wide variety of currencies, horizons, and model specifications. The theoretical framework of the efficient market hypothesis (Fama 1970) gives the background: in weak-form efficiency, all information about future returns is inaccessible in the past in the form of exploitable information. In the present study, the lack of significant improvement over a zero-return benchmark is interpreted only as being consistent with weak-form efficiency within this sample, horizon, and information set, rather than as proof of the efficient market hypothesis.

The classical econometric methods are also sources of reference. Autoregressive models, ARCH (Engle 1982), and GARCH (Bollerslev 1986) offer understandable structure and an explicit theoretical basis, especially in Exchange Rate Volatility Prediction, where they model conditional variance as a persistent dynamic mechanism. Frömmel et al. (2022) demonstrated that exchange-rate adjustment could show nonlinearities in terms of purchasing power parity deviations, and this supports the perception that various dimensions of exchange rates could have diverse predictability profiles.

2.2. Recurrent Neural Networks for Financial Forecasting

The most widely used deep learning method for financial time-series prediction is the recurrent neural network. The Long Short-Term Memory network (Hochreiter and Schmidhuber 1997) solves the vanishing-gradient problem with the use of memory cells and gating operations, which makes it applicable in financial sequences when a signal of interest can appear across several lags (LeCun et al. 2015). A more compact recurrent version with a reduced number of parameters is provided by the GRU (Cho et al. 2014), and it usually performs similarly. The early results reported by Fischer and Krauss (2018), showing that LSTM-based models can learn meaningful structure when predicting financial markets, were influential. Hybrid deep learning was shown by Iqbal et al. (2024) to have an advantage over standard exchange-rate settings. Nelson et al. (2017) and Siami-Namini et al. (2018) also confirmed that LSTMs tend to be superior to traditional ARIMA models when it comes to financial time series.

2.3. Attention Mechanisms and Transformer Architectures

Attention mechanisms enable neural networks to focus on the most relevant portions of a sequence of inputs. In neural machine translation, Bahdanau et al. (2015) proposed additive attention, in which one learns to weight the encoded states, and this weighting yields better results and interpretability. Vaswani et al. (2017) later introduced self-attention as the prevailing design with the Transformer architecture, which has been customized for financial time-series prediction by a number of recent studies (Zeng et al. 2023; Zhou et al. 2021). In the case of multivariate time series, in particular, Qin et al. (2017) developed a two-stage attention recurrent network in which input attention can select the most relevant driving series and temporal attention the most informative encoded time steps. The current paper directly evaluates the empirical support of each of the three forecasting tasks based on added complexity using a controlled ablation design.

2.4. Volatility Forecasting and Feature Engineering

Exchange-rate volatility is not methodologically equivalent to mean forecasting since volatility is clustered and persistent (Bollerslev 1986; Engle 1982). An extensive review by Poon and Granger (2003) demonstrated that good volatility forecasting performance requires informative features. Engineered features, returns, log returns, rolling standard deviations, bid–ask spreads, and lagged volatility reveal volatility dynamics not indicated by raw price levels in machine learning environments (Andersen et al. 2001b; Bandi and Russell 2006). Franke et al. (2011) and Andersen et al. (2001a) showed that realized volatility measures based on intraday returns are among the strongest predictors of future volatility. The current research uses causal historical versions of these feature types and evaluates future volatility against historical-volatility, EWMA, and GARCH benchmarks using both squared-error criteria and QLIKE.

2.5. Feed-Forward Networks and Hybrid Architectures

Multilayer perceptrons, which are universal function approximators, have been proposed by Hornik et al. (1989) and have been used in financial prediction tasks since at least by White (1990). More modern hybrid models combine feed-forward and recurrent terms with an attention mechanism to capture both local interactions between features and long-range temporal dependencies (Iqbal et al. 2024; Qin et al. 2017). Accordingly, the empirical question in this study is not whether one architecture is universally superior, but whether apparent gains remain after target construction, information timing, financial benchmarks, and out-of-sample evaluation are controlled.

Table 1 highlights that the existing literature provides strong evidence separately on exchange-rate persistence, weak return predictability, volatility dynamics, recurrent architectures, attention mechanisms, and finance-aware representations. However, these strands are typically evaluated under different forecasting targets, information sets, benchmarks, and experimental settings. Consequently, reported gains cannot always be attributed uniquely to model architecture. The gap addressed in this study is therefore not the absence of another forecasting architecture, but the absence of a controlled task-aware comparison in which level, return, and genuinely future-volatility forecasting are evaluated under a common chronological design while explicitly controlling information timing, benchmark choice, feature representation, and forecast-loss evaluation.

Table 1. Summary of related literature, research gaps, and positioning of the present study.

Literature Stream	Representative Studies	Main Forecasting Focus	Main Insight	Limitation/Gap Relative to This Study
Exchange-rate predictability and random-walk benchmarks	Meese and Rogoff (1983); Engel and West (2005); Fama (1970)	Exchange-rate levels and returns	Random-walk benchmarks are difficult to outperform out of sample, while short-horizon returns generally contain weak exploitable predictive information.	Level and return forecasting are usually discussed separately rather than compared jointly with volatility under the same information set and evaluation design.
Econometric volatility forecasting	Engle (1982); Bollerslev (1986); Poon and Granger (2003)	Conditional and future volatility	Volatility exhibits clustering and persistence, motivating ARCH/GARCH-type models and volatility-specific evaluation.	Econometric volatility benchmarks are not always evaluated alongside machine-learning and deep-learning models under an identical chronological forecasting framework.

Table 1. *Cont.*

Literature Stream	Representative Studies	Main Forecasting Focus	Main Insight	Limitation/Gap Relative to This Study
Recurrent neural networks	Hochreiter and Schmidhuber (1997); Cho et al. (2014); Fischer and Krauss (2018); Iqbal et al. (2024)	Financial time-series and sequential prediction	LSTM and GRU models can capture nonlinear temporal dependencies that may not be represented by simpler linear models.	Most studies emphasize model architecture for a particular target, making it difficult to distinguish architectural gains from differences caused by target construction or persistence.
Attention-based and hybrid architectures	Bahdanau et al. (2015); Vaswani et al. (2017); Qin et al. (2017)	Feature selection and temporal weighting in sequential models	Attention mechanisms can adaptively weight input variables or historical time steps and may improve representation learning.	Evidence that additional attention complexity consistently improves out-of-sample financial forecasting is limited and highly task dependent; controlled ablation against simpler models remains necessary.
Finance-aware feature engineering	Franke et al. (2011); Andersen et al. (2001a, 2001b); Bandi and Russell (2006)	Volatility and risk-related forecasting	Returns, realized or historical volatility, spreads, and lagged volatility can expose information that is not apparent in raw price levels.	The influence of feature representation is often assessed independently of target choice and model complexity rather than within one controlled multi-task experiment.
Present study	This study	Level, first-difference return, and genuinely future volatility	Compares linear, feed-forward, recurrent, and attention-based models under a common chronological framework with task-specific benchmarks and causal information timing.	Addresses the above gap by jointly controlling the forecasting target, information set, benchmark, feature representation, model complexity, and out-of-sample evaluation procedure.

3. Methodology

This section presents a common empirical framework for assessing exchange-rate forecasting across three prediction tasks: Exchange Rate Prediction, Exchange Rate Return Prediction, and Exchange Rate Volatility Prediction. The methodological design is controlled so that differences in predictive performance can be attributed to the forecasting target, input representation, and model structure, rather than to inconsistent data processing or evaluation procedures. Each experiment, therefore, uses the same chronological train/validation/test design, leakage-safe preprocessing, sequence-construction procedure, and a common core evaluation framework with task-specific metrics and benchmarks.

Figure 1 summarizes the overall methodology. The framework first preserves the temporal ordering of the exchange-rate observations and applies training-only scaling. Three forecasting targets are then constructed from the same underlying exchange-rate series: the next exchange-rate level, the next first-difference return, and genuinely future volatility. Seven model families, ranging from Linear Regression and MLPRegressor to recurrent and attention-based deep architectures, are evaluated under comparable conditions. MAE, RMSE, and R^2 are reported across tasks; MAPE is restricted to the positive level target, non-zero-movement Directional Accuracy is reported for level and return prediction, and QLIKE is additionally used for volatility. Diebold–Mariano tests are employed for key benchmark comparisons.

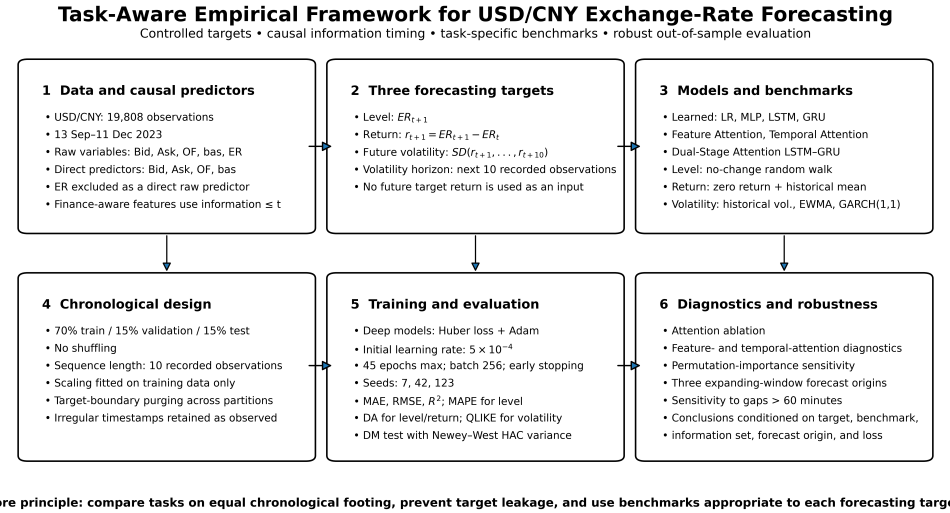

Figure 1. Task-aware forecasting framework showing causal information timing, the genuinely future 10-observation volatility target, and task-specific benchmarks.

3.1. Dataset and Preprocessing

The analysis is based on 19,808 chronologically ordered USD/CNY observations obtained through LSEG/Refinitiv Datastream, with the China Foreign Exchange Trade System (CFETS) as the original market source. The observations span from 13 September 2023 01:35:00 to 11 December 2023 03:36:00.

The observations are irregularly spaced; the median interval is 1 min, and 66.44% of consecutive observations are exactly 1 min apart. There are no weekend observations, and no synthetic non-trading observations are inserted. The data are divided chronologically, without shuffling, into 70% training, 15% validation, and 15% test partitions.

The retained dataset contains no missing values in the required fields; therefore, neither forward filling nor backward filling is performed. MinMax scaling is fitted using training information only and is subsequently frozen for the validation and test data. All engineered variables are constructed causally using information available at or before forecast origin t. Target-boundary purging is applied so that sequences do not share future target observations across the training, validation, and test partitions.

Table 2 summarizes the dataset and preprocessing details.

Table 2. Dataset description and preprocessing details.

Item	Specification
Currency pair	USD/CNY
Data access platform	LSEG/Refinitiv Datastream
Data source	China Foreign Exchange Trade System (CFETS)
Period/N	13 Sep 2023 01:35–11 Dec 2023 03:36/19,808
Timestamp structure	Irregular; median interval = 1 min; 66.44% exactly 1 min
Weekend observations/missing cells	0/0
Raw variables	Bid, Ask, OF, bas, ER
Direct raw predictors	Bid, Ask, OF, bas
Split/sequence length	70/15/15 chronological/10 recorded observations
Bid: mean (SD), min–max	7.247962 (0.070301), 7.1175–7.3198
Ask: mean (SD), min–max	7.247826 (0.070345), 7.1174–7.3198
OF: mean (SD), min–max	0.423920 (5.036542), −65–67
bas: mean (SD), min–max	−0.000135 (0.000288), −0.0028–0.0000
ER: mean (SD), min–max	7.247865 (0.070341), 7.1175–7.3198

Figure 2 shows the exchange-rate series together with the chronological partition boundaries.

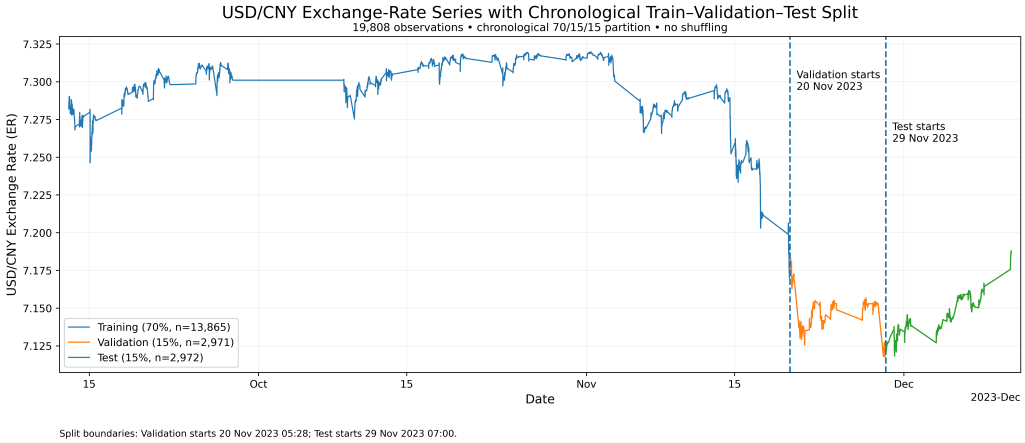

Figure 2. Exchange-rate (ER) series with chronological train/validation/test split boundaries. All 19,808 observations are shown; vertical lines indicate the partition boundaries used across the forecasting tasks.

3.2. Forecasting Tasks

The three forecasting tasks are constructed from the same ER series and evaluated using the same chronological design and model set. The main difference among the tasks is the construction of the prediction target. A common core of evaluation metrics is retained, while task-specific financial benchmarks are used where appropriate. Table 3 summarizes the experimental formulation.

Table 3. Summary of the forecasting tasks and experimental conditions.

Task	Target	Statistical Property	Evaluation Consideration
Exchange Rate Prediction	ER_{t+1}	Strong persistence in levels	Must be interpreted relative to a no-change random-walk benchmark
Return Prediction	r_{t+1}	Weak short-horizon signal after differencing	Compared with zero return; EMH interpretation is sample-specific
Volatility Prediction (raw)	$\mathrm{SD}(r_{t+1}, \ldots, r_{t+10})$	Future volatility over 10 recorded observations	Raw inputs compared with volatility benchmarks
Volatility Prediction (finance-aware)	$\mathrm{SD}(r_{t+1}, \ldots, r_{t+10})$	Same future target with causal historical inputs	Finance-aware models compared with EWMA/GARCH and QLIKE

3.2.1. Exchange Rate Level Target and Benchmark

The first task forecasts the next raw exchange-rate level. Its target is defined as

$$y_t^{(ER)} = ER_{t+1}. \tag{1}$$

Because raw exchange-rate levels exhibit strong persistence, high predictive performance on this task may result primarily from extrapolating the most recent level rather than learning economically meaningful incremental structure (Engel and West 2005; Meese and Rogoff 1983). Accordingly, all learned models are evaluated against the no-change random-walk benchmark

$$\widehat{ER}_{t+1}^{RW} = ER_t. \tag{2}$$

This benchmark is essential because a near-perfect R^2 for the level target can arise from persistence alone.

3.2.2. Exchange Rate Return Target and Benchmark

The second task forecasts the next first-difference return. The first-difference series is defined as

$$r_t = ER_t - ER_{t-1}, \tag{3}$$

and the prediction target is

$$y_t^{(R)} = r_{t+1} = ER_{t+1} - ER_t. \tag{4}$$

Differencing substantially removes the persistence that dominates the level target and directly examines whether the direction and magnitude of the subsequent exchange-rate change can be predicted using information available at time t. Under the weak-form efficient market hypothesis, historical prices should not contain readily exploitable information about future returns (Fama 1970). In this study, however, the interpretation is deliberately narrower: failure to improve significantly over a zero-return benchmark is considered consistent with weak-form efficiency only within the present sample, forecasting horizon, and information set; it is not treated as proof of the efficient market hypothesis.

The primary benchmark for this task is therefore

$$\hat{r}_{t+1}^{0} = 0. \tag{5}$$

3.2.3. Future Volatility Target and Forecast Horizon

The third task predicts genuinely future exchange-rate volatility. At forecast origin t, the target is the sample standard deviation of the next 10 first-difference returns,

$$\sigma_{t,10}^{\text{future}} = \text{SD}(r_{t+1}, r_{t+2}, \ldots, r_{t+10}). \tag{6}$$

The forecasting target is, therefore,

$$y_t^{(V)} = \sigma_{t,10}^{\text{future}}. \tag{7}$$

The horizon represents the next 10 recorded observations, not necessarily the next 10 clock minutes, because the timestamps are irregular. All raw and finance-aware predictors are available at or before forecast origin t. No return from r_{t+1} through r_{t+10} is used as an input when constructing the corresponding target. This temporal separation prevents target-feature overlap and ensures that no information from the future volatility window enters the predictor set.

Volatility clustering and persistence (Bollerslev 1986; Engle 1982) may provide conditional structure for forecasting, although the achievable out-of-sample predictive accuracy remains an empirical question.

3.3. Sequence Construction

For all sequence-based models, the input at forecast origin t is a sliding window containing $w = 10$ consecutive recorded observations. Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the predictor vector at time t. The input sequence can be represented as

$$\mathbf{X}_t = [\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \ldots, \mathbf{x}_t], \qquad w = 10. \tag{8}$$

The same 10-observation sequence length is used for all three tasks. Only the prediction target changes across tasks, while Task 3 additionally compares raw and finance-aware feature representations.

Because timestamps are irregular, $w = 10$ represents 10 successive records rather than a fixed 10-min interval. Sequences and their associated targets are assigned chronologically, with target-boundary purging used to prevent future target observations from crossing the training, validation, and test boundaries.

3.4. Compared Model Architectures

Seven learned model architectures are compared, ranging from a linear baseline to recurrent and dual-stage attention networks. All models receive comparable 10-observation information sets. The direct raw predictors are Bid, Ask, OF, and bas; ER is not included as a direct raw predictor. For volatility forecasting, the raw input representation is additionally compared with causal finance-aware transformations derived from historical ER returns and available market variables.

3.4.1. Linear Regression

Linear Regression provides the simplest learned benchmark and is expressed as

$$\widehat{y}_t = \beta_0 + \sum_{j=1}^{d} \beta_j x_{t,j}, \tag{9}$$

where the coefficients are estimated by minimizing the residual sum of squares. The model provides a reference for assessing whether nonlinear and recurrent architectures yield additional predictive value.

3.4.2. Multilayer Perceptron

The multilayer perceptron (MLP) is a nonlinear feed-forward model (Hornik et al. 1989; White 1990). For hidden layer l,

$$\mathbf{h}^{(l)} = \phi\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{10}$$

where $\phi(\cdot)$ denotes the nonlinear activation function. Although the MLP can learn nonlinear relationships, it does not explicitly model sequential dependence. Its performance is, therefore, evaluated empirically alongside recurrent and attention-based architectures.

3.4.3. Long Short-Term Memory Network

The Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber 1997) models temporal dependencies through input, forget, and output gates:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i), \qquad \mathbf{f}_t = \sigma\left(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f\right), \tag{11}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c), \tag{12}$$

$$\mathbf{h}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o) \odot \tanh(\mathbf{c}_t). \tag{13}$$

3.4.4. Gated Recurrent Unit

The Gated Recurrent Unit (GRU) (Cho et al. 2014) uses a more compact gating mechanism:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z), \qquad \mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r), \tag{14}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tanh(\mathbf{W}_h\mathbf{x}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h). \tag{15}$$

3.4.5. Feature Attention

The feature-attention model learns variable-importance weights following the attention framework of Bahdanau et al. (2015). Feature-attention weights are represented as

$$\boldsymbol{\alpha}_t = \text{softmax}\Big[\mathbf{v}_F^\top \tanh(\mathbf{W}_F \mathbf{x}_t + \mathbf{b}_F)\Big], \tag{16}$$

and the attended input is

$$\widetilde{\mathbf{x}}_t = \boldsymbol{\alpha}_t \odot \mathbf{x}_t. \tag{17}$$

3.4.6. Temporal Attention

Temporal attention learns lag-importance weights over encoded hidden states (Bahdanau et al. 2015; Qin et al. 2017). The context vector is defined as

$$\mathbf{c}_t = \sum_\tau \beta_\tau \mathbf{h}_\tau, \tag{18}$$

where

$$\beta_\tau = \text{softmax}\Big[\mathbf{v}_T^\top \tanh(\mathbf{W}_T \mathbf{h}_\tau + \mathbf{b}_T)\Big]. \tag{19}$$

3.4.7. Dual-Stage Attention LSTM–GRU

The complete architecture combines feature attention, recurrent encoding, and temporal attention, following the dual-stage attention principle of Qin et al. (2017).

In Stage 1, feature attention produces

$$\widetilde{\mathbf{x}}_\tau = \text{softmax}\Big[\mathbf{v}_F^\top \tanh(\mathbf{W}_F \mathbf{x}_\tau + \mathbf{b}_F)\Big] \odot \mathbf{x}_\tau. \tag{20}$$

In Stage 2, the attended sequence is processed by a bidirectional LSTM followed by a bidirectional GRU, producing recurrent hidden representations $\mathbf{h}_\tau$.

In Stage 3, temporal attention aggregates the encoded sequence using the context formulation defined in Equation (18).

The final representation combines the attention context and recurrent representations:

$$\mathbf{z}_t = [\mathbf{c}_t; \mathbf{h}_t; \mathbf{g}_t], \qquad \widehat{y}_t = f(\mathbf{z}_t). \tag{21}$$

For the raw input setting, the complete DualStageAttention_LSTM_GRU architecture contains 73,446 trainable parameters and uses Layer Normalization and Dropout for regularization.

3.4.8. Task-Specific Financial Benchmarks

The learned models are evaluated against benchmarks appropriate to each forecasting target. Task 1 uses the random-walk/no-change forecast in Equation (2). Task 2 uses a zero-return benchmark and a historical-mean return reference. Task 3 is compared with historical volatility, exponentially weighted moving average (EWMA) volatility with $\lambda = 0.94$, and a constrained zero-mean GARCH(1,1) model.

For the GARCH benchmark,

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{22}$$

subject to

$$\alpha \geq 0, \qquad \beta \geq 0, \qquad \alpha + \beta < 1. \tag{23}$$

3.5. Finance-Aware Feature Engineering for Volatility Prediction

The initial volatility experiment uses the same four direct raw predictors as Tasks 1 and 2. The finance-aware experiment constructs a richer representation grounded in the volatility-forecasting literature (Andersen et al. 2001a; Bandi and Russell 2006; Bollerslev 1986; Engle 1982; Poon and Granger 2003). Importantly, all derived variables use historical information available no later than forecast origin t.

The feature groups include:

1. First-difference return, r_t, as defined in Equation (3), which captures directional price changes (Poon and Granger 2003).
2. Log return
$$\ell_t = \log\left(\frac{ER_t}{ER_{t-1}}\right), \tag{24}$$
which represents proportional price changes (Bandi and Russell 2006).
3. Absolute and squared returns, $|r_t|$ and r_t^2, which provide ARCH-type instantaneous volatility proxies (Engle 1982).
4. Rolling mean of returns captures local return behavior using historical observations.
5. Rolling standard deviation of returns is constructed only from observations ending at or before t, which provides a causal historical volatility proxy and reflects volatility clustering and persistence (Bollerslev 1986; Franke et al. 2011).
6. Bid–ask spread information is based on the observed bid and ask variables, which provides a market-microstructure measure associated with changing market conditions (Andersen et al. 2001b).
7. Lagged historical volatility captures volatility persistence while remaining temporally separated from the future target returns (Andersen et al. 2001a).

3.6. Training Protocol and Reproducibility

All deep-learning models are trained using the Adam optimizer (Kingma and Ba 2015). Early stopping is used to monitor validation performance, while ReduceLROnPlateau decreases the learning rate when validation performance stops improving. All model-selection, tuning, and stopping decisions are based exclusively on the chronological validation partition.

Deep models use Huber loss, an initial Adam learning rate of 5×10^{-4}, a maximum of 45 epochs, batch size 256, dropout of 0.20 where specified, and early-stopping patience of 5 epochs. To assess stochastic variability, the deep-learning models are trained using random seeds 7, 42, and 123. MLPRegressor uses hidden layers of 128 and 64 units and a maximum of 300 iterations.

Table 4 summarizes the forecasting information sets and model-complexity details.

Table 4. Forecasting information sets and reproducibility details.

Experiment	Target/Horizon	Inputs Available at t	Primary Benchmark	Test N
Level	ER_{t+1}; one-step	$10 \times$ [Bid, Ask, OF, bas]	No change	2971
Return	r_{t+1}; one-step	$10 \times$ [Bid, Ask, OF, bas]	Zero return	2971
Volatility	$\mathrm{SD}(r_{t+1}, \ldots, r_{t+10})$	Same raw inputs or causal finance-aware features	Historical volatility/EWMA/GARCH	2962

Table 4. *Cont.*

Model Setting	Specification
Raw-input parameter counts	LR: 41; MLP: 13,569; LSTM: 19,777; GRU: 15,553; Feature Attention: 68,829; Temporal Attention: 68,738; Dual Stage: 73,446
Finance-aware parameter counts	LR: 371; MLP: 55,809; LSTM: 28,225; GRU: 21,889; Feature Attention: 88,233; Temporal Attention: 85,634; Dual Stage: 96,084
Deep-learning seeds	7, 42, and 123
Model selection	Based on chronological validation performance

3.7. Evaluation Metrics and Statistical Comparison

Forecasting performance is evaluated using a common set of error metrics supplemented with task-specific measures. Let y_i and $\widehat{y}_i$ denote the observed and predicted values for test observation i, respectively.

Mean Absolute Error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i|. \tag{25}$$

Root Mean Squared Error (RMSE) is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}. \tag{26}$$

For the positive exchange-rate level target, Mean Absolute Percentage Error (MAPE) is

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y}_i|}{|y_i|}. \tag{27}$$

The coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{28}$$

MAPE is restricted to the positive exchange-rate level because percentage errors become unstable when the target is close to zero.

For level and return prediction, Directional Accuracy is also evaluated. The primary version excludes observations for which the realized movement is exactly zero, preventing a no-change forecast from receiving a mechanical advantage. For non-zero movements,

$$\mathrm{DA} = \frac{1}{N_{\mathrm{nz}}}\sum_{i\in\mathcal{I}_{\mathrm{nz}}}\mathbb{I}[\mathrm{sign}(\Delta\widehat{y}_i) = \mathrm{sign}(\Delta y_i)], \tag{29}$$

where $\mathcal{I}_{\mathrm{nz}}$ denotes observations with non-zero realized movement. The all-observation Directional Accuracy is retained as a supplementary diagnostic.

For volatility forecasting, the QLIKE loss is additionally reported:

$$\mathrm{QLIKE}_i = \log\left(\widehat{\sigma}_i^2\right) + \frac{\sigma_i^2}{\widehat{\sigma}_i^2}. \tag{30}$$

Key differences in forecast loss are assessed using the Diebold–Mariano test. For two competing forecasts A and B, the loss differential is defined as

$$d_t = L_{A,t} - L_{B,t}. \tag{31}$$

The test statistic is

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}}, \tag{32}$$

where the long-run variance of the loss differential is estimated using the Newey-West heteroskedasticity and autocorrelation-consistent estimator (Newey and West 1987), following the Diebold–Mariano predictive-accuracy testing framework (Diebold and Mariano 1995). HAC lag 0 is used for the one-step level and return targets, whereas lag 9 is used for the overlapping 10-observation volatility target. For volatility forecasting, QLIKE is additionally reported because volatility-model rankings can depend on the evaluation loss and the properties of the volatility proxy (Hansen and Lunde 2005; Patton 2011).

4. Experimental Results

The seven learned models are evaluated under the common chronological design described in Section 3. For stochastic deep-learning models, the reported values represent the mean and standard deviation across seeds 7, 42, and 123 unless otherwise stated. The results are organized according to the three forecasting tasks, followed by attention ablation, cross-task comparison, and robustness analysis.

4.1. Task 1: Exchange Rate Level Prediction

Table 5 reports the fixed-holdout results for Exchange Rate Prediction. The no-change random walk achieves an RMSE of 0.0005923 and an R^2 of 0.99815, while Linear Regression produces an RMSE of 0.0005939 and an R^2 of 0.99814. The difference in squared-error loss between Linear Regression and the random walk is not statistically significant according to the Diebold–Mariano test ($DM = 0.480$, $p = 0.631$). Consequently, Linear Regression does not significantly outperform the no-change benchmark.

Although the R^2 values for both methods are extremely high, this performance must be interpreted in the context of the strong persistence of the exchange-rate level. The random walk itself achieves approximately the same fit as Linear Regression. Therefore, the high R^2 primarily reflects level persistence rather than substantial incremental forecasting skill. Linear Regression attains a non-zero-movement Directional Accuracy of 0.6076.

Table 5. Fixed-holdout results for Task 1: Exchange Rate Level Prediction. Deep-learning results are the mean ± standard deviation across seeds 7, 42, and 123.

Model	RMSE	MAE	R^2	DA (Non-Zero)
Random Walk/No Change	0.0005923	0.000284	0.99815	0.0000
Linear Regression	0.0005939	0.000296	0.99814	0.6076
FeatureAttention_Only	0.004805 ± 0.000689	0.003958 ± 0.000762	0.8763 ± 0.0357	0.4605 ± 0.0423
GRU	0.007443 ± 0.005432	0.005802 ± 0.004152	0.6033 ± 0.4520	0.5174 ± 0.0034
TemporalAttention_Only	0.009916 ± 0.007594	0.007432 ± 0.005655	0.2772 ± 0.9680	0.5030 ± 0.0519
LSTM	0.010873 ± 0.006082	0.008399 ± 0.004722	0.2449 ± 0.6674	0.5315 ± 0.0291
DualStageAttention_LSTM_GRU	0.014587 ± 0.009178	0.011558 ± 0.007697	-0.4214 ± 1.2558	0.5222 ± 0.0487
MLPRegressor	0.018486 ± 0.004464	0.016366 ± 0.004887	-0.8762 ± 0.8487	0.5715 ± 0.0102

Figure 3 compares the fixed-holdout RMSE values. The random walk and Linear Regression are nearly indistinguishable, whereas the recurrent and attention-based architectures generally exhibit substantially larger errors.

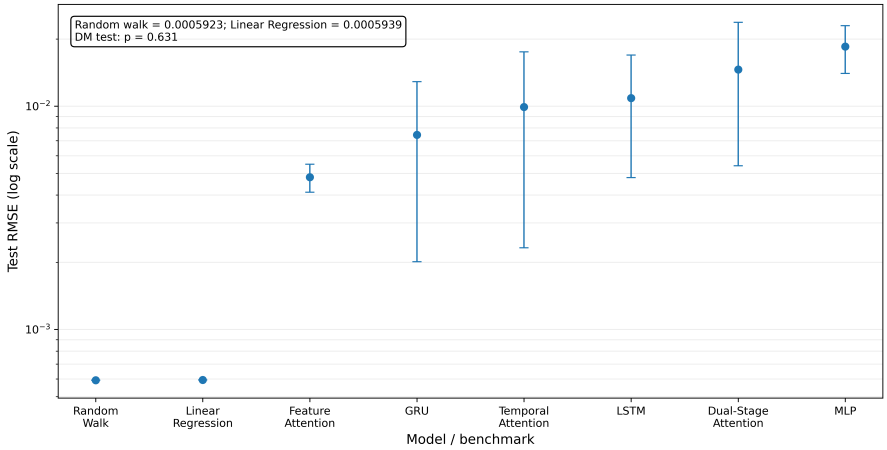

Figure 3. Fixed-holdout RMSE comparison for Task 1. The no-change random walk and Linear Regression are nearly tied, and the Diebold–Mariano test does not reject equal predictive accuracy.

Figure 4 further illustrates the close overlap between the observed USD/CNY level and the forecasts from Linear Regression and the random-walk benchmark.

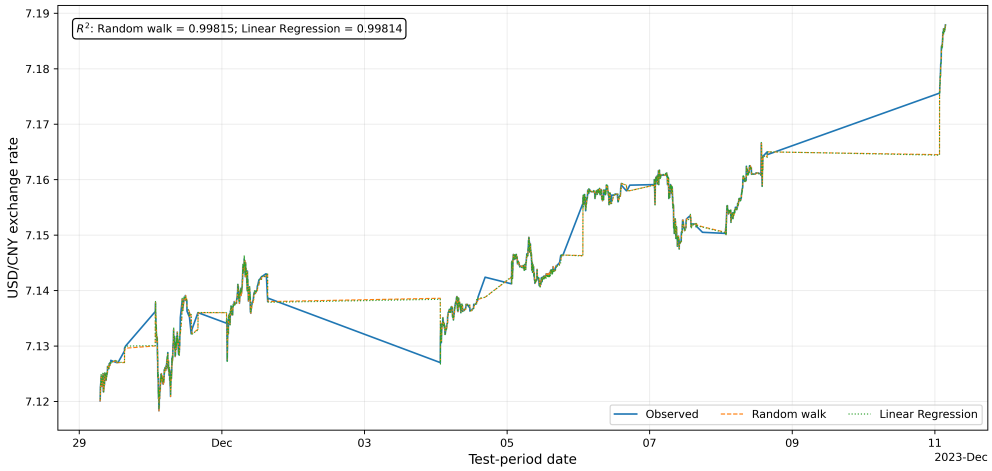

Figure 4. Observed USD/CNY exchange-rate level and fixed-holdout forecasts from Linear Regression and the no-change random walk. Their close overlap demonstrates why high-level R^2 requires interpretation relative to a naive persistence benchmark.

4.2. Task 2: Exchange Rate Return Prediction

The Task 2 results are presented in Table 6. Once the level series is differenced, predictive performance deteriorates substantially. Linear Regression achieves the lowest RMSE among the learned models at 0.0005920, with $R^2 = -0.00061$. However, the zero-return benchmark has an almost identical RMSE of 0.0005923.

The Diebold–Mariano comparison between Linear Regression and the zero-return benchmark is not significant ($DM = -0.083$, $p = 0.934$). Linear Regression therefore provides no statistically meaningful improvement over zero return during the test period. Its non-zero-movement Directional Accuracy is 0.4725.

These results should not be interpreted as proof that USD/CNY returns are random or as verification of the efficient market hypothesis. They indicate only that the selected models and available information set do not provide a significant improvement over zero return within the present sample and forecasting horizon. This finding is consistent with weak-form efficiency in this specific empirical setting.

Table 6. Fixed-holdout results for Task 2: Exchange Rate Return Prediction. Deep-learning results are the mean ± standard deviation across seeds 7, 42, and 123.

Model	RMSE	R^2	DA (Non-Zero)	Interpretation
Linear Regression	0.0005920	−0.00061	0.4725	DM versus zero return: $p = 0.934$
Zero Return	0.0005923	−0.00148	0.0000	Primary financial benchmark
Historical Mean Return	0.0005926	−0.00266	0.4171	Secondary benchmark
FeatureAttention_Only	0.000766 ± 0.000209	−0.7591 ± 0.9861	0.4851 ± 0.0283	Below the zero-return benchmark
LSTM	0.001000 ± 0.000101	−1.8714 ± 0.5925	0.5515 ± 0.0071	Negative out-of-sample R^2
GRU	0.001021 ± 0.000146	−2.0185 ± 0.8655	0.5490 ± 0.0192	Negative out-of-sample R^2
DualStageAttention_LSTM_GRU	0.001261 ± 0.000484	−3.9880 ± 3.7419	0.4736 ± 0.0330	No incremental predictive gain
MLPRegressor	0.001419 ± 0.000729	−5.7557 ± 6.6634	0.4904 ± 0.0803	No incremental predictive gain
TemporalAttention_Only	0.002561 ± 0.001556	−22.3297 ± 18.9757	0.5117 ± 0.0680	Largest forecasting error

Figure 5 emphasizes the similarity between Linear Regression and the zero-return benchmark. The difference is considerably smaller than the errors of the nonlinear and deep-learning alternatives.

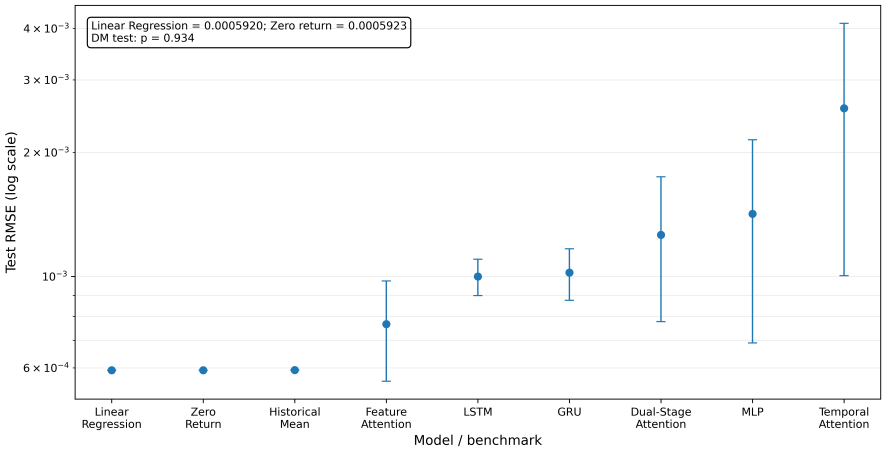

Figure 5. Fixed-holdout RMSE comparison for Task 2. Linear Regression and the zero-return benchmark have almost identical forecasting errors.

4.3. Task 3: Exchange Rate Volatility Prediction

The volatility experiment holds the prediction target fixed across the raw and finance-aware settings. In both cases, the target is the sample standard deviation of the next 10 recorded first-difference returns. Consequently, differences between the two experiments are attributable to the input representation rather than to a change in target construction.

4.3.1. Volatility Prediction Using Raw Features

Table 7 reports the results obtained using the same four direct raw predictors employed in Tasks 1 and 2. Among the learned models, DualStageAttention_LSTM_GRU achieves the lowest mean RMSE of 0.000457 ± 0.000010. However, its mean R^2 remains negative at −0.1018 ± 0.0496. None of the raw-feature learned models, therefore, demonstrates strong out-of-sample volatility predictability.

The financial benchmarks are also included. GARCH(1,1) produces the lowest QLIKE among the reported raw-feature and benchmark models, with QLIKE = 1.2483, despite having a larger RMSE than several learned models. This difference already indicates that model rankings depend on the loss function used for evaluation.

Table 7. Task 3 fixed-holdout results using raw predictors. Deep-learning results are mean ± standard deviation across seeds.

Model	RMSE	R^2	QLIKE
DualStageAttention_LSTM_GRU	0.000457 ± 0.000010	-0.1018 ± 0.0496	1.7235 ± 0.1177
LSTM	0.000462 ± 0.000026	-0.1321 ± 0.1311	4.7850 ± 5.6927
MLPRegressor	0.000468 ± 0.000017	-0.1572 ± 0.0839	87.7221 ± 149.2298
FeatureAttention_Only	0.000474 ± 0.000027	-0.1888 ± 0.1371	1.8771 ± 0.3729
GRU	0.000482 ± 0.000044	-0.2333 ± 0.2267	1.6400 ± 0.2206
TemporalAttention_Only	0.000505 ± 0.000054	-0.3595 ± 0.2798	1.6283 ± 0.2101
EWMA ($\lambda = 0.94$)	0.000506	−0.3544	2.7050
Historical Volatility	0.000512	−0.3854	539.5776
GARCH(1,1)	0.000514	−0.3943	1.2483
Linear Regression	0.000592	−0.8501	1.7296

4.3.2. Volatility Prediction Using Finance-Aware Features

The distribution of first-difference returns and the corresponding historical volatility series show episodic concentration of large movements. This motivates the causal historical volatility, return, spread, and lagged features introduced in Section 3.5.

Figure 6 illustrates the empirical return and volatility dynamics that motivate the finance-aware feature representation. Figure 6a shows that first-difference returns are strongly concentrated around zero, with relatively infrequent large positive and negative movements. Figure 6b shows the corresponding 30-observation rolling historical volatility, which exhibits episodic spikes over the sample period. These patterns support the use of causal return transformations and lagged historical-volatility features for the future-volatility forecasting task.

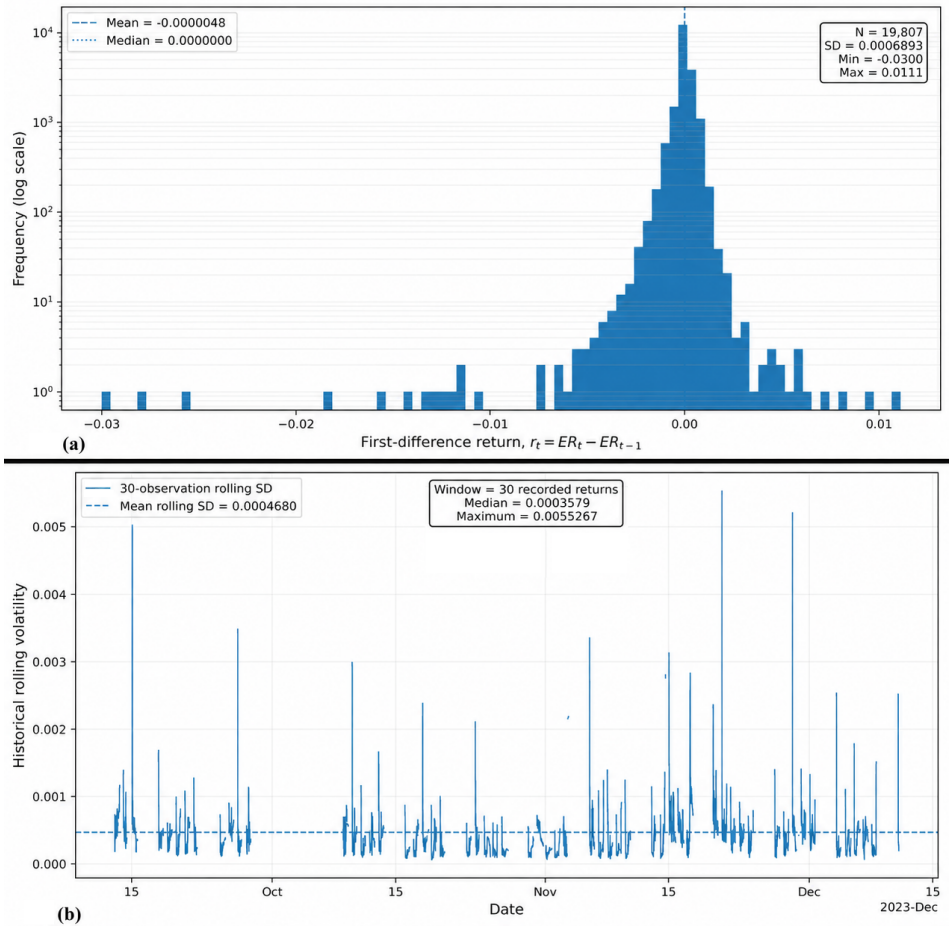

Figure 6. Distribution of (**a**) first-difference returns and (**b**) 30-observation historical volatility for the USD/CNY series.

Table 8 presents the finance-aware future-volatility results. FeatureAttention_Only obtains the best average learned-model RMSE of 0.0004407 ± 0.0000180, with mean $R^2 = -0.027 \pm 0.085$. Its three-seed ensemble further reduces RMSE to 0.0004298 and produces a small positive R^2 of 0.0242.

GRU is also competitive, producing an average RMSE of approximately 0.000448 ± 0.000042 and the lowest average QLIKE among the learned models at approximately 1.2820 ± 0.0064.

The constrained GARCH(1,1) benchmark produces $\alpha = 0.08392$, $\beta = 0.91470$, and persistence $\alpha + \beta = 0.99862$. Its fixed-holdout RMSE is approximately 0.000514, its R^2 is -0.3943, and its QLIKE is 1.2483.

Under squared-error loss, FeatureAttention_Only significantly improves on EWMA ($DM = -2.512$, $p = 0.0120$) and GARCH ($DM = -2.594$, $p = 0.0095$). However, the QLIKE comparison between FeatureAttention_Only and GARCH is not significant ($DM = 0.986$, $p = 0.324$). Thus, the evidence for model superiority is loss-function dependent.

Table 8. Task 3 fixed-holdout results using finance-aware predictors. Deep-learning results are mean ± standard deviation across seeds.

Model	RMSE	R^2	DA	QLIKE
FeatureAttention_Only	0.0004407 ± 0.0000180	-0.0273 ± 0.0848	0.5953 ± 0.0139	1.5029 ± 0.1607
GRU	0.000448 ± 0.000042	-0.0675 ± 0.2025	0.5683 ± 0.0360	1.2820 ± 0.0064
MLPRegressor	0.000459 ± 0.000014	-0.1150 ± 0.0667	0.5580 ± 0.0323	31.2178 ± 37.2404
LSTM	0.000463 ± 0.000049	-0.1400 ± 0.2438	0.5737 ± 0.0482	1.5278 ± 0.3439
DualStageAttention_LSTM_GRU	0.000493 ± 0.000077	-0.3051 ± 0.4180	0.5677 ± 0.0430	1.4429 ± 0.0893
EWMA ($\lambda = 0.94$)	0.000506	−0.3544	0.6161	2.7050
Historical Volatility	0.000512	−0.3854	0.0003	539.5776
GARCH(1,1)	0.000514	−0.3943	0.5469	1.2483
TemporalAttention_Only	0.000540 ± 0.000117	-0.5891 ± 0.7076	0.5683 ± 0.0273	813.2377 ± 1405.9248
Linear Regression	0.000544	−0.5663	0.5631	39.9499

The three-seed FeatureAttention_Only ensemble, which is constructed from the individual seed forecasts, achieves

$$\text{RMSE} = 0.0004298, \qquad R^2 = 0.0242. \tag{33}$$

This is the strongest squared-error result in the fixed-holdout finance-aware experiment. Nevertheless, the GARCH benchmark remains competitive under QLIKE, preventing a general conclusion that the machine-learning model is universally superior.

4.4. Attention Ablation and Diagnostic Analysis

A controlled ablation is used to determine whether feature attention, temporal attention, or their combination adds predictive value on Task 1. Table 9 reports the seed-42 comparison.

The no-attention LSTM–GRU variant achieves RMSE = 0.002085 and $R^2 = 0.9770$. FeatureAttention_Only, TemporalAttention_Only, and the complete dual-stage architecture, all perform worse than this no-attention hybrid and remain substantially below Linear Regression. In this low-dimensional and highly collinear USD/CNY setting, attention mechanisms therefore do not consistently improve out-of-sample performance.

Figure 7 summarizes the average feature-attention weights obtained for the Task 1 diagnostic model. Bid receives the largest average weight (0.4295), followed by bas (0.2682), Ask (0.1824), and OF (0.1199). These weights describe how the fitted model distributes attention across the available inputs and should be interpreted as model-dependent diagnostics rather than as measures of causal or economic importance.

Table 9. Attention ablation for Task 1 at seed 42.

Variant	RMSE	R^2	Interpretation
Linear Regression	0.000594	0.99814	Reference learned baseline; DM versus random walk $p = 0.631$
NoAttention_LSTM_GRU	0.002085	0.9770	Best recurrent ablation; still below Linear Regression
FeatureAttention_Only	0.004181	0.9076	Worse than the no-attention LSTM–GRU variant
DualStageAttention_LSTM_GRU	0.004391	0.8981	Combining both attention stages does not improve performance
TemporalAttention_Only	0.004956	0.8702	Worse than the no-attention LSTM–GRU variant
GRU	0.013051	0.0998	Seed-42 recurrent comparator
LSTM	0.016488	−0.4368	Seed-42 recurrent comparator

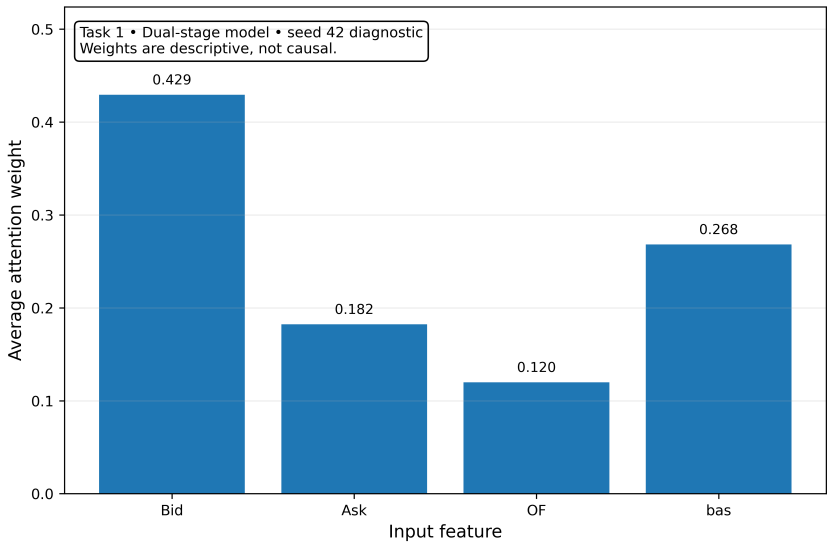

Figure 7. Average feature-attention weights for the Task 1 model. The attention weights are descriptive model diagnostics and should not be interpreted as causal feature importance.

Figure 8 shows the average temporal-attention weights across the 10 recorded lag positions. The weights remain close to the uniform value of 0.10, with only small variation across lag positions. Thus, the fitted model does not identify any single historical lag as consistently dominant.

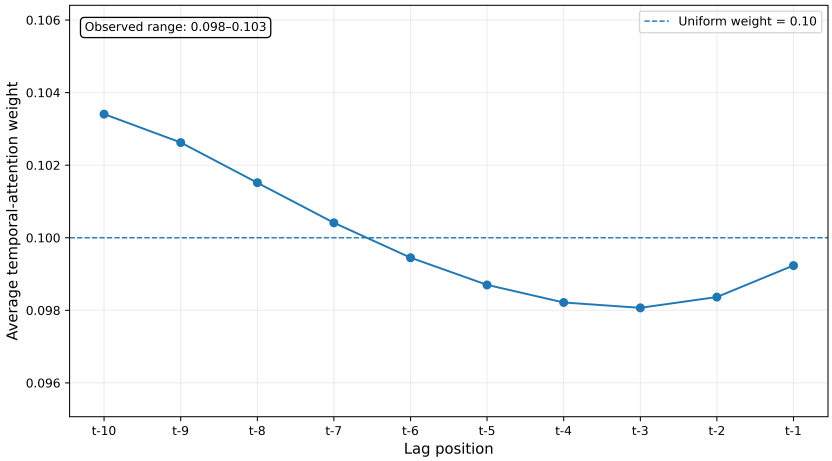

Figure 8. Average temporal-attention weights across the 10 recorded lag positions. The near-uniform distribution indicates that the model does not identify one historical lag as consistently dominant.

To complement the attention-weight diagnostics, permutation importance is applied to the Task 1 exchange-rate level forecasting model as an independent measure of predictive sensitivity. Figure 9 reports the increase in test RMSE obtained after independently

permuting each direct input feature while holding the remaining inputs unchanged. Bid produces by far the largest deterioration in predictive accuracy, increasing RMSE by approximately 0.01057, followed by the bid–ask spread variable (bas), with an increase of approximately 0.00163. Permuting Ask results in only a small increase of approximately 0.00015, whereas permutation of order flow (OF) produces essentially no change in the test RMSE. These results indicate that the fitted Dual-Stage Attention LSTM–GRU model relies predominantly on bid information for the level-forecasting task, with substantially smaller incremental sensitivity to the remaining raw predictors. This finding is consistent with the strong persistence and collinearity of the price-related variables observed in the USD/CNY level series. Nevertheless, permutation importance measures the predictive sensitivity of the fitted model rather than causal or economic importance; consequently, the values in Figure 9 are interpreted only as complementary model diagnostics.

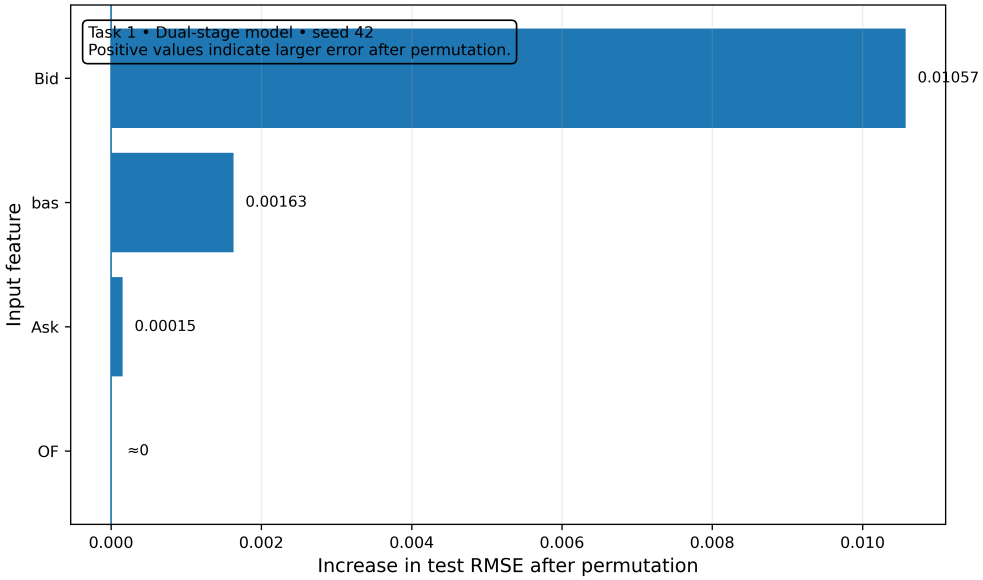

Figure 9. Permutation importance for Task 1 using the Dual-Stage Attention LSTM–GRU model at seed 42.

4.5. Cross-Task Comparison

Table 10 summarizes the central empirical results across the three forecasting targets. The findings differ substantially according to target construction and benchmark choice.

Table 10. Cross-task summary of the fixed-holdout results.

Forecasting Task	Leading Model/Benchmark	RMSE	R^2	Primary Finding
Exchange-rate level	Random Walk/Linear Regression	0.0005923/0.0005939	0.99815/0.99814	High fit mainly reflects persistence; DM $p = 0.631$
First-difference return	Linear Regression/Zero Return	0.0005920/0.0005923	−0.00061/−0.00148	No significant improvement over zero return; DM $p = 0.934$
Future volatility, raw predictors	DualStageAttention_LSTM_GRU	0.000457 ± 0.000010	-0.1018 ± 0.0496	Raw inputs provide only modest out-of-sample information
Future volatility, finance-aware predictors	FeatureAttention ensemble	0.0004298	0.0242	Best squared-error point forecast; GARCH remains competitive under QLIKE

For level forecasting, both the random walk and Linear Regression achieve R^2 values of approximately 0.998, and the difference in squared-error loss is not statistically significant.

For return forecasting, Linear Regression does not significantly improve on the zero-return benchmark. Future volatility shows somewhat greater conditional structure, particularly when causal finance-aware predictors are used, but the magnitude of out-of-sample predictability remains modest. Moreover, the comparison between FeatureAttention_Only and GARCH depends on whether squared-error loss or QLIKE is used.

4.6. Robustness Across Forecast Origins and Timestamp Gaps

To assess whether conclusions from a single fixed hold-out are stable, three expanding-window forecast origins are evaluated. The rankings change across these origins. Table 11 summarizes the resulting robustness comparison. For finance-aware volatility, DualStageAttention_LSTM_GRU has a mean RMSE of approximately 0.000646, FeatureAttention_Only approximately 0.000758, EWMA approximately 0.000784, and GARCH approximately 0.000797. In contrast, GARCH records the lowest mean QLIKE among these models at 1.4999.

Table 11. Expanding-window robustness results for selected finance-aware volatility models.

Model	Mean RMSE	Mean R^2	Mean QLIKE
DualStageAttention_LSTM_GRU	0.000646 ± 0.000208	-0.0379	2.0321
FeatureAttention_Only	0.000758 ± 0.000294	-0.4648	2.1379
EWMA	0.000784 ± 0.000251	-0.5296	2.3975
GARCH(1,1)	0.000797 ± 0.000248	-0.5847	1.4999
GRU	0.000799 ± 0.000336	-0.5595	3.0644
Historical Volatility	0.000812 ± 0.000274	-0.6295	639.7049
Linear Regression	0.000816 ± 0.000302	-0.6304	255.3444

The irregular timestamps also motivate a sensitivity analysis excluding forecast sequences whose input-to-target span crosses a gap exceeding 60 min. Table 12 compares RMSE using all eligible test observations with RMSE after excluding these long-gap cases.

Table 12. Sensitivity of RMSE to sequences crossing timestamp gaps greater than 60 min.

Task/Model	All Test Observations	Excluding > 60-min Gaps
Level: Random Walk	0.000592	0.000425
Level: Linear Regression	0.000594	0.000426
Return: Zero Return	0.000592	0.000425
Return: Linear Regression	0.000592	0.000422
Volatility: FeatureAttention Ensemble	0.000430	0.000267
Volatility: EWMA	0.000506	0.000342
Volatility: GARCH(1,1)	0.000514	0.000353

The decline in RMSE after removing long-gap cases indicates that irregular observation spacing contributes to forecasting difficulty. Importantly, however, the qualitative conclusions remain unchanged: level performance is dominated by persistence, return forecasting does not materially improve over zero return, and volatility-model rankings depend on both the evaluation period and the chosen loss function. These robustness results therefore caution against interpreting the fixed-holdout ranking as evidence of a universally superior forecasting architecture.

5. Discussion

The level-forecasting results show that both the no-change random walk and Linear Regression attain R^2 values of approximately 0.998, while their squared-error difference is not statistically significant. The high-level fit should therefore be interpreted primarily as a consequence of persistence rather than as evidence of substantial incremental forecasting

skill (Engel and West 2005; Meese and Rogoff 1983). In the fixed hold-out, the recurrent and attention-based models do not improve on the random-walk or Linear Regression benchmarks. However, this result is specific to the present low-dimensional USD/CNY information set and should not be generalized to other currencies, information sets, or forecasting horizons.

This result has an important methodological implication. High goodness-of-fit statistics for exchange-rate level prediction should not by themselves be interpreted as economically meaningful forecasting performance. Because the no-change forecast already achieves an R^2 close to one, a forecasting model should be evaluated against an appropriate persistence benchmark before its predictive value is assessed. A model may track the exchange rate level closely while providing little incremental information beyond the most recently observed exchange rate. Reporting results for a differenced return target separately from the level target is therefore important when evaluating whether a model captures predictive information beyond level persistence.

The return-prediction results reinforce this distinction. Linear Regression does not significantly outperform the zero-return benchmark in the present sample, with a Diebold–Mariano p-value of 0.934. This lack of incremental predictability is consistent with weak-form efficiency within the present sample, horizon, and information set, but it does not prove that USD/CNY returns are random or establish the efficient market hypothesis (Engel and West 2005; Fama 1970; Meese and Rogoff 1983). The result only shows that the models and predictors considered in this study do not extract statistically significant short-horizon return information beyond the zero-return benchmark. Consequently, the apparent success of a model on the level target should be distinguished from its ability to forecast subsequent signed exchange-rate movements.

The future-volatility experiment provides a different perspective. With the genuinely future target defined as the standard deviation of the next 10 recorded first-difference returns, mean out-of-sample R^2 values are modest and are generally close to or below zero. The strongest finance-aware ensemble reaches an R^2 of only 0.0242. Nevertheless, finance-aware inputs improve squared-error point forecasts for selected learned models compared with the corresponding financial benchmarks in the fixed hold-out. FeatureAttention_Only, for example, significantly improves squared-error loss relative to EWMA and GARCH. This advantage does not persist under every evaluation criterion: GARCH remains highly competitive under QLIKE, and the QLIKE difference between FeatureAttention_Only and GARCH is not statistically significant. Hence, the evidence does not support a universal ranking of the models. Instead, the preferred forecasting approach depends partly on the loss function used to evaluate volatility predictions.

These findings are consistent with the broader financial-econometrics literature emphasizing volatility persistence, clustering, benchmark choice, and appropriate volatility-loss evaluation (Andersen et al. 2001a; Bandi and Russell 2006; Bollerslev 1986; Engle 1982; Poon and Granger 2003). In particular, the comparison between the raw and finance-aware experiments indicates that input representation can materially affect forecasting performance. Historical return transformations, lagged volatility information, and other causally constructed finance-aware variables provide information that is not represented directly by the raw market variables. However, the magnitude of the resulting future-volatility predictability remains modest. The results therefore support the narrower conclusion that representation, benchmark choice, and loss function can be at least as important as architectural complexity.

The attention-ablation results provide further evidence against interpreting architectural complexity as a default source of improvement. In the Task 1 seed-42 experiment, the no-attention LSTM–GRU variant achieves a lower RMSE than FeatureAttention_Only,

TemporalAttention_Only, and the complete dual-stage attention architecture, while all of these models remain inferior to Linear Regression. Thus, in this low-dimensional and highly collinear USD/CNY setting, attention mechanisms did not consistently improve out-of-sample performance.

The learned attention weights should also be interpreted cautiously. The feature-attention weights shown in Figure 7 and the nearly uniform temporal-attention weights shown in Figure 8 are model-dependent diagnostics rather than causal evidence of feature or lag importance. Permutation importance is therefore reported as an additional sensitivity diagnostic. Even with this additional analysis, no general claim is made regarding the effectiveness of attention mechanisms outside the present dataset and experimental setting.

The robustness analysis further limits any claim of universal model superiority. Across the three expanding-window forecast origins, volatility-model rankings change noticeably. Although some learned models produce lower mean squared-error measures at particular origins, GARCH produces the lowest mean QLIKE among the principal volatility models considered. Similarly, excluding observations affected by timestamp gaps greater than 60 min reduces forecasting errors, but does not alter the main qualitative conclusions. These findings indicate that forecast origin and irregular observation spacing can influence measured performance and reinforce the importance of evaluating financial forecasting models across more than one test period.

Taken together, Table 10 summarizes task-dependent rather than hierarchical evidence. Level forecasting is dominated by persistence; return forecasting shows no statistically significant improvement over zero return; and future-volatility performance depends on the input representation, forecast origin, benchmark, and loss function. The main implication for model selection is, therefore, that the forecasting target, information timing, benchmark, and evaluation criterion should be specified before claims of model superiority are made (Fischer and Krauss 2018; Iqbal et al. 2024; Poon and Granger 2003; Rossi 2013). Architectural sophistication alone is insufficient evidence that a model provides superior out-of-sample forecasts.

From a practical risk-management perspective, the results suggest that volatility forecasting may offer a more relevant setting for evaluating conditional risk information than raw exchange-rate level prediction. However, the present study evaluates statistical forecasting accuracy rather than realized economic value. No transaction-cost model, trading strategy, hedge-ratio optimization, utility-based evaluation, Value-at-Risk strategy, or ex ante economic-value test is implemented. Consequently, improvements in RMSE or other statistical loss measures should not be interpreted as demonstrated profitability, hedging effectiveness, or economic gains.

6. Conclusions

This study provides a unified empirical framework for exchange-rate forecasting across three distinct prediction tasks: exchange-rate level prediction, return prediction, and future-volatility prediction. Seven learned model families were evaluated together with task-specific financial benchmarks under a strict chronological configuration. The results show that conclusions about forecasting performance depend strongly on the target definition, information set, benchmark, forecast origin, and evaluation criterion.

For the exchange-rate level target, the no-change random walk and Linear Regression both achieve R^2 values of approximately 0.998 and are statistically indistinguishable under squared-error loss. This indicates that the high goodness of fit is driven primarily by persistence in the exchange-rate level rather than by substantial incremental forecasting skill. More complex recurrent and attention-based architectures do not provide a consistent advantage for this task within the present experimental setting.

For the return target, Linear Regression does not significantly outperform the zero-return benchmark. This result is consistent with weak-form efficiency only within the present USD/CNY sample, forecasting horizon, and information set, and should not be interpreted as proof that exchange-rate returns are inherently random or that the efficient market hypothesis is universally valid.

For genuinely future volatility, finance-aware inputs improve squared-error forecasts for selected learned models in the fixed hold-out. The FeatureAttention_Only ensemble provides the strongest squared-error result in that experiment, but the overall level of out-of-sample predictability remains modest. Moreover, GARCH remains competitive under QLIKE, demonstrating that apparent model superiority depends on the loss function used for evaluation. These results indicate that feature representation, benchmark choice, and evaluation criterion can be at least as important as architectural complexity.

The attention-ablation analysis further shows that attention mechanisms do not consistently improve out-of-sample performance in this low-dimensional and highly collinear USD/CNY setting. Attention weights are therefore interpreted only as model-dependent diagnostics rather than causal measures of feature or lag importance. More broadly, the results support a task-aware interpretation rather than a universal hierarchy of model performance: claims of superiority should be conditioned on the forecasting target, information structure, financial benchmark, forecast origin, and loss function.

The study has several limitations. It is restricted to a single USD/CNY currency pair, an approximately three-month sample, irregular timestamps, a 10-recorded-observation input window, and a 10-recorded-observation volatility horizon. These horizons are observation-based rather than fixed clock-time horizons. Future research should therefore evaluate whether the present task-dependent findings generalize to additional currency pairs, longer samples, alternative sampling frequencies, and different market regimes. More information-rich predictors, including macroeconomic fundamentals, interest-rate differentials, additional order-flow variables, and market-sentiment indicators, may also be investigated to determine whether return forecasting can be improved. Future studies may further compare hybrid deep-learning–econometric models, transformer-based time-series models, and regime-switching approaches for volatility forecasting, while explicitly testing robustness to structural breaks and crisis periods. Finally, economic-value evaluations involving trading, hedging, transaction costs, and risk-management applications would provide an important extension beyond the statistical forecast-accuracy analysis conducted in this study.

Author Contributions: Conceptualization, M.A.C. and K.S.; Methodology, M.A.; Software, M.A. and S.K.; Validation, S.K.; Formal analysis, M.A.; Investigation, S.B. and M.A.; Resources, S.K.; Data curation, M.A.C.; Writing—original draft, S.B., M.A.C. and K.S.; Writing—review and editing, S.B.; Supervision, S.K.; Funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Multimedia University (MMU) through its Article Page Sponsorship Scheme.

Data Availability Statement: The authors have uploaded the analysis-ready USD/CNY dataset and the code required to reproduce the preprocessing, forecasting experiments, statistical tests, tables, and figures reported in this study to the public repository and are publicly available at https://doi.org/10.5281/zenodo.22166169 (accessed on 29 August 2026). The exchange-rate observations were accessed through LSEG/Refinitiv Datastream, with the China Foreign Exchange Trade System (CFETS) identified as the original market source. The repository also contains the code and supporting result files required to reproduce the reported empirical findings.

Conflicts of Interest: The authors declare no conflict of interest.

References

Alvarez, Fernando, Andrew Atkeson, and Patrick J. Kehoe. 2007. If exchange rates are random walks, then almost everything we say about monetary policy is wrong. *American Economic Review* 97: 339–45. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001a. The distribution of realized stock return volatility. *Journal of Financial Economics* 61: 43–76. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2001b. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96: 42–55. [CrossRef]

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. Paper presented at International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7–9.

Bandi, Federico M., and Jeffrey R. Russell. 2006. Separating microstructure noise from volatility. *Journal of Financial Economics* 79: 655–92. [CrossRef]

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. Paper presented at the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29. pp. 1724–34.

Diebold, Francis X., and Roberto S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13: 253–63. [CrossRef]

Engel, Charles, and Kenneth D. West. 2005. Exchange rates and fundamentals. *Journal of Political Economy* 113: 485–517. [CrossRef]

Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica* 50: 987–1007. [CrossRef]

Fama, Eugene F. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25: 383–417. [CrossRef]

Fischer, Thomas, and Christopher Krauss. 2018. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* 270: 654–69. [CrossRef]

Franke, Jürgen, Wolfgang Karl Härdle, and Christian M. Hafner. 2011. *Statistics of Financial Markets: An Introduction*. Berlin/Heidelberg: Springer

Frömmel, Michael, Darko B. Vukovic, and Jinyuan Wu. 2022. The dollar exchange rate, adjustment to the purchasing power parity, and the interest rate differential. *Mathematics* 10: 4504. [CrossRef]

Hansen, Peter R., and Asger Lunde. 2005. A forecast comparison of volatility models: Does anything beat a garch(1,1)? *Journal of Applied Econometrics* 20: 873–89. [CrossRef]

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9: 1735–80. [CrossRef]

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks 2*: 359–66. [CrossRef]

Iqbal, Farhat, Dimitrios Koutmos, Eman A. Ahmed, and Lulwah M. Al-Essa. 2024. A novel hybrid deep learning method for accurate exchange rate prediction. *Risks* 12: 139. [CrossRef]

Kingma, Diederik P., and Jimmy Ba. 2015. Adam: A method for stochastic optimization. Paper presented at International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7–9.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521: 436–44. [CrossRef]

Meese, Richard A., and Kenneth Rogoff. 1983. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14: 3–24.

Nelson, David M. Q., Adriano C. M. Pereira, and Renato A. de Oliveira. 2017. Stock market's price movement prediction with lstm neural networks. Paper presented at 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, May 14–19. pp. 1419–26.

Newey, Whitney K., and Kenneth D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–8. [CrossRef]

Patton, Andrew J. 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160: 246–56. [CrossRef]

Poon, Ser-Huang, and Clive W. J. Granger. 2003. Forecasting volatility in financial markets: A review. *Journal of Economic Literature* 41: 478–539. [CrossRef]

Qin, Yao, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W. Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. Paper presented at the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, August 19–25. pp. 2627–33. [CrossRef]

Rossi, Barbara. 2013. Exchange rate predictability. *Journal of Economic Literature* 51: 1063–119. [CrossRef]

Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. 2018. A comparison of arima and lstm in forecasting time series. Paper presented at 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, December 17–20. pp. 1394–401.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Paper presented at Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, December 4–9. pp. 5998–6008.

White, Halbert. 1990. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks* 3: 535–49. [CrossRef]

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? Paper presented at the AAAI Conference on Artificial Intelligence, Washington, DC, USA, February 7–14. pp. 11121–28.

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. Paper presented at the AAAI Conference on Artificial Intelligence, Virtual, February 2–9. pp. 11106–15.