Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comparative Analysis of Machine Learning and Deep Learning Models for Forecasting Vietnam's Export Coffee Prices

Quang Phung Duy

quangpd@ftu.edu.vn

Foreign Trade University

Hai Dang Cao Ha

National Economics University

Research Article

Keywords: Time series forecasting, GRU, Diebold–Mariano test, multi-seed robustness, commodity price risk management

Posted Date: August 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10306092/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR FORECASTING VIETNAM'S EXPORT COFFEE PRICES

Phung Duy Quang, Ph.D.
Faculty of Technology and Data Science, Foreign Trade University
Email: quangpd@ftu.edu.vn
Cao Ha Hai Dang
Faculty of Data Science and Artificial Intelligence, College of Technology, National Economics University
Email: 11247148@st.neu.edu.vn

Abstract

This study benchmarks three machine learning models (Support Vector Regression, Random Forest, and XGBoost) against three deep learning architectures (DNN, LSTM, and GRU) for forecasting the export price of Vietnamese Robusta coffee on the ICE London exchange, using a strict 300-day out-of-sample test window (20 November 2024 to 28 January 2026) covering severe structural breaks. The Gated Recurrent Unit (GRU) is the strongest architecture (MAE = 96.55 USD/ton, MAPE = 2.14%, R^2 = 0.967), outperforming the best machine learning model, Random Forest (R^2 = 0.888), and the deeper LSTM (R^2 = 0.856); Diebold–Mariano tests confirm that GRU's advantage over all five remaining models is significant at the 1% level. The purely feedforward DNN (R^2 = –1.665) performs worse than every machine learning model, indicating that sequential inductive bias, rather than network depth, determines forecasting capability under structural breaks. Two methodological issues are reported transparently: an asymmetry in training-data scope between the machine learning and deep learning groups, and a multi-seed robustness check showing that both LSTM (CV(MAE) = 39.52%) and GRU (CV(MAE) = 24.80%) exceed the 15% stability reference threshold. Using the GRU model, the study implements a 7-day recursive forecast with an expanding 95% confidence interval, producing explicit risk-floor and risk-ceiling parameters that exporters can use to structure forward contracts.

Keywords: Time series forecasting; GRU; Diebold–Mariano test; multi-seed robustness; commodity price risk management.

Corresponding author: Phung Duy Quang (Email: quangpd@ftu.edu.vn)

1. Introduction

Coffee is one of Vietnam's leading agricultural export commodities and plays a central role in the country's agricultural economy and trade balance. As of 2024, Vietnam remains the world's second-largest coffee producer and exporter, accounting for roughly 17% of global output and close to 20% of the global market share [1]. Export turnover reached USD 4.2 billion in 2023 across more than 80 countries and territories, sustaining the livelihoods of millions of workers in the Central Highlands [2].

Vietnam's coffee export trajectory, however, faces mounting risk from the instability of global price structures. Transaction data from ICE London shows Robusta coffee prices moving through an extreme cycle, ranging from a low of USD 1,121/ton to a peak of USD 5,765/ton between 2008 and 2026. Unlike earlier cycles, this volatility is not merely seasonal; it reflects broad structural breaks driven jointly by global supply-chain disruption, geopolitical tension, shifts in central-bank monetary policy — most notably the Federal Reserve's policy rate — and capital-reallocation behaviour among commodity-derivative hedge funds [3]. The nonlinear interaction among these financial drivers produces chaotic price dynamics that challenge every traditional linear forecasting method [4].

Under this uncertainty, building an accurate price-forecasting system has become a pressing quantitative need across the coffee value chain, particularly for exporters engaged in financial hedging. Recognizing the limits of classical statistical methods, this study establishes a transparent benchmarking framework to evaluate the capacity of nonlinear learning algorithms. Specifically, the study cross-evaluates six representative architectures from two schools of thought — traditional machine learning (SVR, Random Forest, XGBoost) and deep neural networks (DNN, LSTM, GRU) — under conditions of structural market breaks. The study's core contribution is not simply to identify the model with the lowest error, but to translate that statistical error into a recursive Risk Management Matrix (expanding confidence intervals), while explicitly disclosing every methodological design choice that could affect the fairness of the comparison — an aspect frequently left unaddressed in the existing benchmarking literature.

This study is guided by four research questions:

- RQ1: Does a sequential deep-learning architecture (LSTM or GRU) outperform traditional machine learning models (SVR, Random Forest, XGBoost) in forecasting Robusta coffee prices out-of-sample, under conditions of severe structural breaks?
- RQ2: Is the performance advantage of deep learning attributable to sequential inductive bias specifically, or simply to network depth and parameter count?
- RQ3: Are the performance differences among models statistically significant, and are the best-performing recurrent architectures stable across random weight initializations?
- RQ4: How can out-of-sample forecast error be translated into quantitative, decision-usable risk parameters for exporters engaged in forward-contract hedging?

The remainder of the paper is organized as follows. Section 2 reviews the capabilities and limitations of ML and DL models in commodity price forecasting. Section 3 sets out the research methodology, data pipeline, and system architecture, including a dedicated subsection disclosing the asymmetry in training-data scope between the two model families. Section 4 presents the empirical results and their risk-management implications. Section 5 summarizes the core conclusions and outlines directions for future research.

2. Literature Review

2.1. Machine learning models in agricultural price forecasting: foundations, empirical evidence, and structural limits

Machine learning approaches to agricultural commodity price forecasting have been studied extensively since the early 2000s. A review of this literature reveals a recurring methodological limitation: most studies evaluate performance on a random split or in-sample test set, which does not reflect true forecasting capability during periods of structural market breaks. This section reviews three representative ML architectures and explains why each carries a structural limitation that cannot be resolved through hyperparameter tuning alone.

2.1.1. Support Vector Regression (SVR)

SVR extends Vapnik's [5] support vector machine theory, applying the principle of structural risk

minimization to regression. Through the radial basis function (RBF) kernel, SVR maps a low-dimensional feature space into a higher-dimensional Hilbert space where nonlinear relationships can be approximated linearly [6]. A key theoretical advantage of SVR is the convexity of its optimization problem, which guarantees convergence to a global optimum rather than becoming trapped in local minima, as neural networks can.

In the Vietnamese context, SVR with an RBF kernel has also been applied to Robusta price forecasting, though such studies have generally relied on data predating the major structural breaks of 2023–2025 and have not applied a Diebold–Mariano [7] significance test, leaving claims of SVR's relative superiority without an econometric foundation.

Structural limitation: SVR built on lag-1 features operates within the static feature space of a single time step and has no mechanism for encoding long-sequence information. The RBF kernel measures distance within a single day's feature space; it cannot distinguish two days with identical prices but entirely different 60-day trends. This is an intrinsic limitation of any kernel method applied to time series with long temporal dependency.

2.1.2. Random Forest

Random Forest is an ensemble method that combines bootstrap aggregating (bagging) with feature randomization. Each tree is trained on a bootstrap sample and considers only a random subset of features at each split, a mechanism that increases diversity among trees and reduces aggregate variance following Breiman's [8] principle. Its practical strengths include robustness to outliers, tolerance of multicollinearity, and the ability to report feature importance.

Divina et al. [9] combined Random Forest, ARIMA, and SVM in an ensemble for short-term electricity-demand forecasting, showing that ensembling reduces variance and improves stability. That setting, however — highly seasonal, stationary, short-term electricity demand — differs fundamentally from coffee commodity prices, which are non-stationary, subject to structural breaks, and exhibit long memory.

Structural limitation: like every tree-based algorithm, Random Forest cannot extrapolate beyond the distribution of its training data. Mathematically, its predictions are bounded above by max(y_train) and below by min(y_train). The empirical results in Section 4.2 confirm this theoretical limit directly: even when trained on a price range extending up to USD 5,334/ton, Random Forest still fails to track the price peaks that exceed this range during the 300-day test window.

2.1.3. XGBoost — Extreme Gradient Boosting

Chen and Guestrin [10] developed XGBoost within the gradient-boosting framework, in which decision trees are trained sequentially, with each new tree optimizing the negative gradient of the loss function from the current ensemble. XGBoost adds L1/L2 regularization, shrinkage, column subsampling, and automatic handling of missing values. With over 30,000 academic citations, it is widely regarded as best-in-class for tabular data.

XGBoost has also been combined with recurrent architectures such as LSTM and GRU in stacking ensembles for agricultural commodity forecasting, with prior work suggesting that XGBoost can contribute positively when paired with sequential models. Such studies, however, have generally not analyzed XGBoost's standalone performance when prices exceed historical training bounds — precisely the scenario faced during the test period of the present study (2024–2026).

Structural limitation: XGBoost shares Random Forest's extrapolation limit but suffers a more severe version of it. Because boosting optimizes the residuals of preceding trees, once the first tree plateaus, subsequent trees merely fit small residual noise around that ceiling and cannot escape the bound imposed by the training range. The empirical results (Section 4.2.1) confirm this: despite sharing the same training range as Random Forest, XGBoost performs markedly worse ($R^2 = 0.465$ versus 0.888), suggesting that sequential boosting is more prone to overfitting local residual structure than independent bagging under conditions of severe structural breaks.

2.2. Deep learning architectures in time series forecasting: progress and conditions for effectiveness

Deep learning has driven substantial progress in financial time series forecasting through its ability to learn representations directly from raw data. Not every DL architecture, however, is suited to time series forecasting; effectiveness depends on sequential inductive bias — the architecture's ability to encode

temporal order and retain information across multiple time steps. This section reviews three DL architectures against that criterion.

2.2.1. DNN — Deep Neural Network (a baseline without sequential bias)

The feedforward DNN is a foundational deep-learning architecture. With multiple hidden layers and nonlinear activations, it can approximate any continuous function (the universal approximation theorem [11]). It has, however, no mechanism for processing temporal order: at each time step t, the DNN receives a flattened feature vector and maps it directly to an output, with no memory state connecting successive time steps. Formally, a DNN learns a function $f: \mathbb{R}^d \to \mathbb{R}$, whereas LSTM/GRU learn $f: \mathbb{R}^{dxT} \to \mathbb{R}$, where T is sequence length. DNN's role in this study is that of a negative control: if DNN performs well, sequential structure is not important; if it performs poorly, the opposite is confirmed.

2.2.2. LSTM — Long Short-Term Memory

Hochreiter and Schmidhuber [12] designed LSTM to solve the vanishing-gradient problem of conventional RNNs, in which gradients decay exponentially during backpropagation through many time steps [13]. Its core solution is a dedicated cell state regulated by three gates: forget, input, and output.

In the Vietnamese coffee domain, Tran Van Minh and Nguyen Thi Hong [14] applied a three-layer LSTM with a 60-day window, achieving MAPE = 6.9%. That study, however, did not perform a multi-seed robustness check — a gap the present study directly addresses in Section 3.6. The results show this gap is consequential: LSTM in the present study exhibits substantial sensitivity to random initialization (Section 4.5), a finding that may explain why single-run studies tend to report overly optimistic conclusions about LSTM.

2.2.3. GRU — Gated Recurrent Unit

Cho et al. [15] proposed GRU as a lighter recurrent architecture, merging the forget and input gates into a single update gate z(t) and adding a reset gate r(t) to control the influence of the prior hidden state. The update rule is $h(t) = (1 - z(t)) \odot h(t-1) + z(t) \odot \tilde{h}(t)$.

With fewer parameters (no separate cell state), GRU typically converges faster and performs comparably to LSTM on small-to-medium datasets. GRU has also been reported to converge faster than LSTM under limited-data conditions in the broader forecasting literature. The present study's results (Sections 4.2.2 and 4.5) are consistent with and extend this observation: GRU converges faster and achieves higher, more stable accuracy than LSTM across multiple seeds — a result not predicted a priori by gating theory, since LSTM in principle has richer representational capacity through its separate cell state. This is explored further in Section 4.2.2.

Synthesis of Section 2: the literature review confirms three points. (i) ML models carry a structural extrapolation limit that cannot be fully overcome once the market moves beyond its historical price range. (ii) Sequential DL architectures (LSTM/GRU) have a clear theoretical advantage, but prior studies lack rigorous statistical testing and robustness verification — and once subjected to such testing, the assumption that "LSTM is always the best DL choice" no longer holds. (iii) No prior study has benchmarked all six ML/DL architectures on Vietnamese export coffee data under the 2023–2026 structural-break regime while also disclosing design constraints — training-range coverage and seed stability — that could affect the conclusions.

2.3. Research gap and contribution

The literature review identifies four research gaps that remain unresolved:

Research gap	**Contribution of this study**
Absence of a comprehensive ML-vs-DL comparison framework	Benchmarks six models (3 ML + 3 DL) on the same dataset and the same 300-day out-of-sample test window
Absence of rigorous statistical testing	Applies the Diebold–Mariano test to every comparison against the best model, distinguishing significance at the 1% and 5% levels
Absence of robustness testing across multiple initializations	Applies a multi-seed robustness check (5 seeds) to the recurrent architectures, reporting instability honestly where it occurs
Lack of transparency regarding experimental-design fairness	Explicitly discloses and discusses the difference in training-data scope between ML and DL (Section 3.3.4)

Table 2.1: Research gaps and the corresponding contribution of this study

The core distinguishing feature of this study relative to prior work is that it does not presuppose which model family — ML or DL — will prevail, allowing empirical evidence and statistical testing to decide, even where the outcome runs against theoretical expectation (for example, Random Forest outperforming LSTM). This principle of objective benchmarking remains underused in the literature on Vietnamese coffee export forecasting.

3. Research Methodology and System Design

3.1. Data collection and feature space

The study defines a five-dimensional multivariate feature space at daily trading frequency.

Variable	Symbol	Economic rationale
ICE London Robusta coffee price	Last Price	Target variable — the benchmark measure of Vietnam's export coffee position
US Dollar Index	USD_Index	Coffee is globally priced in USD; index movement is a direct source of exchange-rate risk
Federal Reserve policy rate	EFFR	A proxy for the cost of capital, affecting the liquidity of commodity-derivative hedge funds
WTI crude oil price	Crude_Oil	A component of logistics and supply-chain cost, affecting freight rates
Trading volume	Volume	Reflects market liquidity depth and behavioural momentum

Table 3.1: Five-dimensional multivariate feature space and economic rationale

The dataset comprises 4,603 observations spanning 14 January 2008 to 28 January 2026, drawn from ICE Futures Europe (Robusta price), FRED — Federal Reserve Economic Data (EFFR, USD Index) — and the U.S. Energy Information Administration (WTI crude price). After preprocessing, the dataset contains no missing values.

3.2. Preprocessing pipeline and calendar alignment

Data engineering proceeds in two stages: (1) volume normalization — the Volume column contains abbreviated notation ('K', 'M'), which is parsed and converted to a consistent numeric format, with missing values set to 0.0; and (2) trading-calendar alignment — because ICE London and the U.S. macroeconomic data sources observe different holiday calendars, a left merge is applied using the coffee exchange's calendar as the reference axis, with macro-data gaps filled by forward fill (ffill).

3.3. Experimental design and data-leakage prevention

The integrity of the benchmark rests on two simultaneous requirements: (1) absolute prevention of information leakage from the future into training, and (2) a fair and clearly justified basis for comparing architectures. The study applies four design principles.

Principle 1 — Chronological partitioning and an internal validation firewall

To assess genuine out-of-sample forecasting capability, the final 300 trading days (20 November 2024 to 28 January 2026) are fully isolated as the Test Set. The remaining data — the Gross Train set (4,303 observations, 14 January 2008 to 19 November 2024) — is further split at a validation ratio of 0.15: the final 646 observations (29 April 2022 to 19 November 2024) form the Validation Set, used exclusively for early stopping in the sequential neural networks; the remaining 3,657 observations (14 January 2008 to 28 April 2022) form the Pure Train set. Table 3.2 summarizes the three-way chronological split.

Dataset	Observations	Date range	Price range (USD/ton)
Pure Train	3,657 days	14/01/2008 → 28/04/2022	1,121 – 2,751
Validation	646 days	29/04/2022 → 19/11/2024	1,792 – 5,334
Test (out-of-sample)	300 days	20/11/2024 → 28/01/2026	3,164 – 5,765

Table 3.2: Chronological structure of the three data splits

Principle 2 — Strict scaling isolation

A MinMaxScaler(0,1) is fit exclusively on the Gross Train set (Pure Train + Validation, 4,303 observations, price range 1,121–5,334 USD/ton), with no access whatsoever to the Test set during fitting. This is a deliberate departure from fitting the scaler on Pure Train

alone: because the coffee price series is an I(1) process with a strong upward trend (Section 4.1.2), fitting only on Pure Train — which ends before the 2022–2024 price surge — would push both the Validation and Test sets far outside the [0,1] range, distorting the val_loss signal used for early stopping and undermining the meaning of extrapolation for tree-based models. Extending the fitting range to the full Gross Train set does not violate the leakage-prevention principle, since the Test set remains entirely unseen, and it keeps the residual degree of extrapolation — test prices exceeding USD 5,334/ton — within a moderate range: only 56 of 300 test days (18.7%) exceed the scaler's upper bound, compared with 300 of 300 test days falling out-of-distribution under the narrower Pure-Train-only fitting scenario. This residual 18.7% constitutes the study's deliberate stress-test zone, analyzed in depth in Section 4.2.

Principle 3 — Fair comparison design and the choice of feature space

An important methodological question deserves explicit treatment: why do the ML models (Random Forest, XGBoost, SVR) receive only flattened lag-1 features, while the DL models (LSTM, GRU) receive 60-day sequences? This is a deliberate design choice grounded in three considerations: (i) it reflects realistic deployment conditions — tabular ML models are typically deployed to act on the most recent available observation, and forcing ML to consume all 60 lags would induce severe multicollinearity and the curse of dimensionality; (ii) it follows established benchmarking convention in financial forecasting research [16]; and (iii) it preserves the architectural distinction under test — the goal is to compare spatial structure against sequential inductive bias, not to compare feature-engineering technique.

Principle 4 — Disclosing the asymmetry in training-data scope between ML and DL

This is the most important methodological point to state explicitly rather than leave for a reviewer to discover. Table 3.3 shows that the ML group (SVR, Random Forest, XGBoost) is trained on the entire flattened Gross Train set (4,243 sequences, taking the value at the final time step), while the DL group (DNN, LSTM, GRU) updates its weights (via gradient descent) only on Pure Train (3,597 sequences) — the Validation set (646 sequences) is used solely to monitor val_loss for early stopping and does not participate in backpropagation.

Parameter	DL group	ML group
Data structure	3D tensor (sequence, 60×5)	2D matrix (flat, 5 dimensions)
Weight-update (training) data	3,597 sequences — Pure Train only (price 1,121–2,751)	4,243 observations — Gross Train = Pure Train + Validation (price 1,121–5,334)
Role of the Validation Set	Monitors val_loss for early stopping (no gradient update)	Not used — ML relies on intrinsic regularization
Overfitting control	Early stopping (patience = 10) on a dedicated Validation set	Bagging (RF); L1/L2 + shrinkage (XGBoost); margin maximization (SVR)
Test Set	300 sequences, fully isolated	300 observations, fully isolated

Table 3.3: Comparison of training-data scope between the DL and ML groups

The justification for this asymmetry is that parameter-heavy sequential networks (LSTM/GRU each use two gated layers of 128 and 64 units) carry high overfitting risk and require an independent validation set, excluded from gradient updates, to determine the optimal stopping point — standard practice in the deep learning literature [17]. Random Forest and XGBoost, by contrast, have intrinsic regularization mechanisms (bagging, max_depth, subsampling) that do not require a held-out set for early stopping, allowing the entire Gross Train set to be used for training.

This asymmetry must nonetheless be interpreted carefully, particularly for the LSTM-versus-Random-Forest comparison. Because Random Forest is trained on a wider price range (up to USD 5,334/ton) than LSTM/GRU/DNN (up to USD 2,751/ton only), Random Forest's higher R^2 (0.888 versus LSTM's 0.856) cannot be read simply as “ML beats sequential architecture” — part of the gap is a data-scope confound rather than pure architectural evidence. By contrast, the XGBoost-versus-Random-Forest comparison is entirely fair (identical Gross Train exposure), and the performance gap between them (R^2 = 0.465 versus 0.888) reflects a

genuine architectural effect: sequential boosting is more prone to becoming trapped in local residual structure than Random Forest's independent bagging. This is discussed further in Section 4.2.3.

3.4. Benchmarking framework

Hyperparameters were fixed based on common values reported in the literature and preliminary experimentation, rather than a formal search procedure — a limitation discussed in Section 5.2..

Model	Hyperparameters	Rationale
SVR	kernel = RBF, C = 100, ε = 0.01, gamma='scale'	Classical nonlinear benchmark; the kernel trick maps data to a higher-dimensional space to capture nonlinear relationships
Random Forest	n_estimators = 200, max_depth = 10	Bagging ensemble baseline that reduces variance, is robust to overfitting, and yields feature importance
XGBoost	n_estimators = 300, lr = 0.05, max_depth = 6, subsample = 0.8, colsample = 0.8	State-of-the-art gradient boosting for tabular data [10]

Table 3.4a: ML architectures and rationale

Model	Architecture	Rationale
DNN	Dense(128→64→32→1) + Dropout(0.2)	Simplest DL baseline — feedforward, no sequential memory. Role: negative control
LSTM	LSTM(128)→LSTM(64) + Dropout(0.2) + Dense(32→1)	Three-gate memory mechanism addressing vanishing gradients; retains 60-day long-term memory
GRU	GRU(128)→GRU(64) + Dropout(0.2) + Dense(32→1)	A simplified two-gate version of LSTM; fewer parameters, faster convergence

Table 3.4b: DL architectures and rationale

All DL models use the Adam optimizer, MSE loss, early stopping (patience = 10, monitored on val_loss), batch size 32, and a maximum of 50 epochs. Each model initialization uses an independent, freshly instantiated early-stopping callback with no shared state across runs, ensuring complete isolation between experiments run under different seeds.

3.5. Performance evaluation criteria

Metric	Formula	Interpretation
MAE	mean(|y − ŷ|)	Mean absolute error (USD/ton) — directly interpretable for management decisions
RMSE	√mean((y − ŷ)²)	Penalizes large errors more heavily — suited to assessing sudden price-spike risk
MAPE	mean(|y − ŷ|/y) × 100	Percentage error, scale-independent, enabling cross-period comparison
R^2	1 − SS_res/SS_tot	Proportion of variance explained; $R^2 < 0$ means the model performs worse than predicting the mean

Table 3.5: Performance evaluation criteria

To test whether performance differences are econometrically meaningful or merely random variation, the Diebold–Mariano (DM) test is applied with HAC variance correction using a Newey-West/Andrews bandwidth of 5. The null hypothesis states that two models have equivalent forecasting ability, with the loss differential defined as $d(t) = e_1(t)^2 - e_2(t)^2$. The study reports both the 1% and 5% significance thresholds explicitly, rather than a single threshold, since (as shown in Section 4.3) not every comparison in this study reaches significance at the 1% level.

3.6. Robustness verification — multi-seed check

DL architectures involve randomness in weight initialization, dropout masking, and mini-batch ordering. To test whether GRU's superior performance depends on a particular random seed or reflects a stable architectural property, the study independently trains

LSTM and GRU under five random seeds: {0, 7, 21, 42, 99}.

The robustness check is restricted to the two recurrent architectures (LSTM, GRU) rather than applied to all six models, for three reasons: (i) these are the most heavily parameterized architectures and therefore the most sensitive to random initialization; (ii) Random Forest's bagging mechanism is specifically designed to reduce variance across initializations [8], so testing its seed sensitivity is not the study's central question; and (iii) SVR with an RBF kernel has a globally convergent convex optimum and, in theory, no run-to-run variance on identical data. A model is deemed stable when the coefficient of variation of MAE (CV = σ/μ) is below 15%, a reference threshold from Vo and Nguyen [18].

4. Results and Discussion

4.1. Exploratory data analysis

To ensure empirical objectivity and eliminate any look-ahead bias, the study applies strict data isolation: all descriptive statistics, tests, and correlation analyses in this section are computed exclusively on the Gross Train set (4,303 observations, 14 January 2008 to 19 November 2024) — Pure Train and Validation combined, excluding the 300-day Test set. This range is chosen to match exactly the range on which the scaler and the ML models are actually fit (Section 3.3), rather than being limited to Pure Train alone, as is common practice.

4.1.1. Descriptive statistics

Over the Gross Train period, the Robusta price ranges from a low of USD 1,121/ton to a high of USD 5,334/ton, with a mean of USD 1,943.31/ton and a standard deviation of USD 615.70/ton. The median (USD 1,864/ton) sits notably below the mean — a signature of a right-skewed distribution, reflecting large-magnitude price spikes concentrated in short windows of time (see Figure 4.1, panel B, for the precise skewness statistic). Strategically, the fact that the Gross Train boundary coincides with a local price peak of USD 5,334/ton — just before the market entered its most extreme volatility phase — creates a natural stress test: 56 of 300 test days (18.7%) record prices exceeding this threshold, large enough to expose the extrapolation limits of traditional ML models without being so

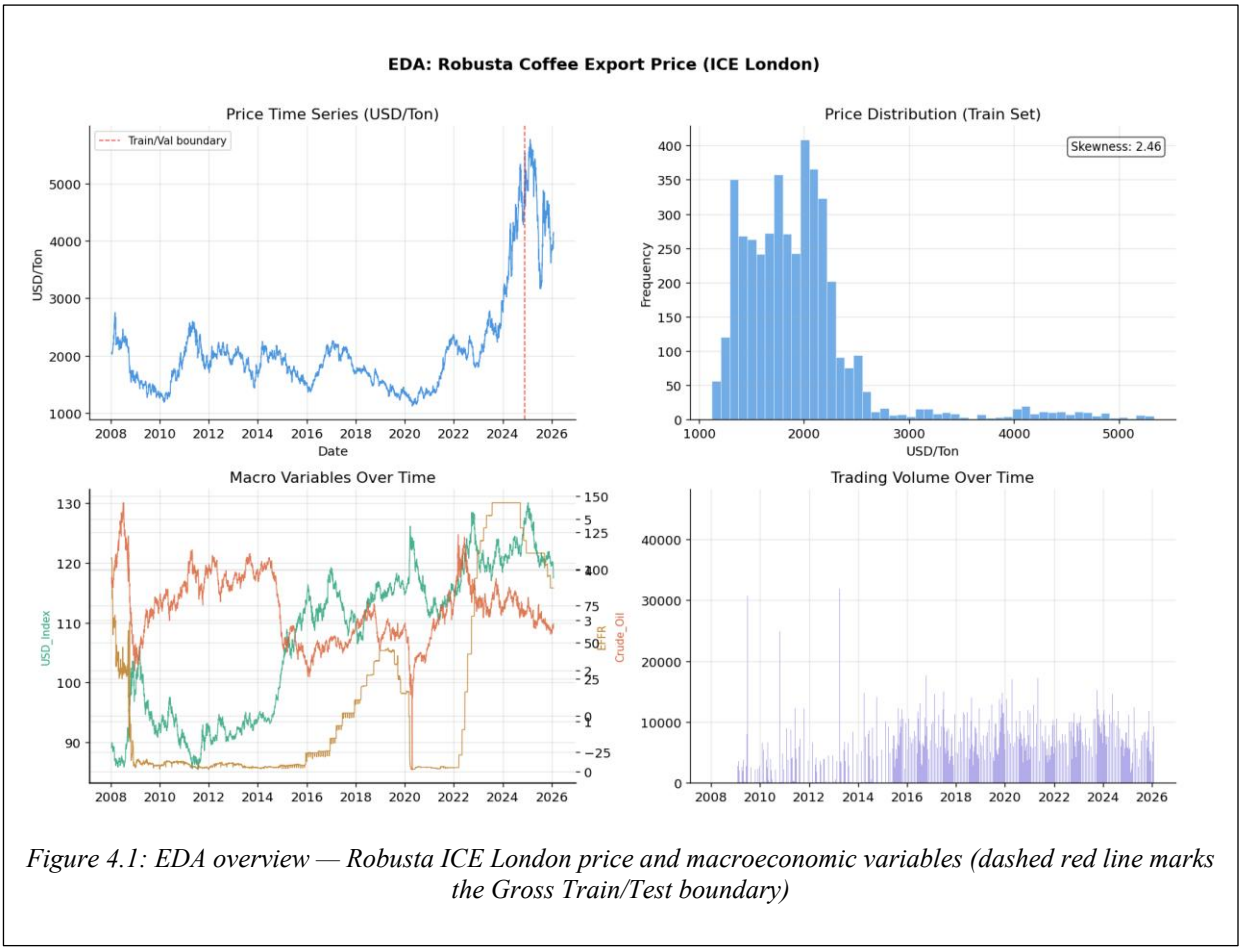


Figure 4.1: EDA overview — Robusta ICE London price and macroeconomic variables (dashed red line marks the Gross Train/Test boundary)

extreme as to undermine cross-architecture comparability.

4.1.2. Stationarity test

To establish the econometric basis for model selection, an Augmented Dickey-Fuller (ADF) test is applied to check the stationarity of the price series.

Series	ADF statistic	p-value	Conclusion
Price level	1.1707	0.9958	Non-stationary (fail to reject H0)
First difference	-16.7177	0.0000	Stationary - H0 rejected at the 1% level

Table 4.1: Augmented Dickey-Fuller (ADF) test results on the Gross Train set

On the level series, the ADF statistic (1.1707) is well above the 1% critical value (−3.431), with p = 0.9958, giving no statistical basis to reject the unit-root null. After first-differencing, the ADF statistic falls sharply to −16.7177, far beyond the critical threshold, with p = 0.0000. This confirms that the coffee price series is a non-stationary, first-order integrated process, I(1). Methodologically, this I(1) property rules out static regression models and establishes the mathematical basis for applying recurrent nonlinear architectures.

4.1.3. Correlation on the differenced series (spurious-correlation prevention)

A classic pitfall in macroeconomic data analysis is spurious correlation when measuring two co-trending I(1) series. To identify the genuine underlying relationship, Pearson correlation is computed on the first-differenced (Δ) series — the stationarized, I(0) series.

The empirical results on the Gross Train set show that contemporaneous linear correlation between coffee price changes and macro shocks is very weak: ΔUSD_Index (r = −0.157), ΔCrude_Oil (r = 0.104), ΔVolume (r = 0.030), and ΔEFFR (r = 0.024). The absence of strong linear correlation does not mean these exogenous macro variables are uninformative — on the contrary, it provides strong evidence that the transmission mechanism from macro shocks to coffee prices is nonlinear and involves complex lag structures. This finding directly motivates rejecting flat, lag-1-only ML models as the presumptively optimal choice, and justifies the use of a 60-day lookback sequence as input to the DL models.

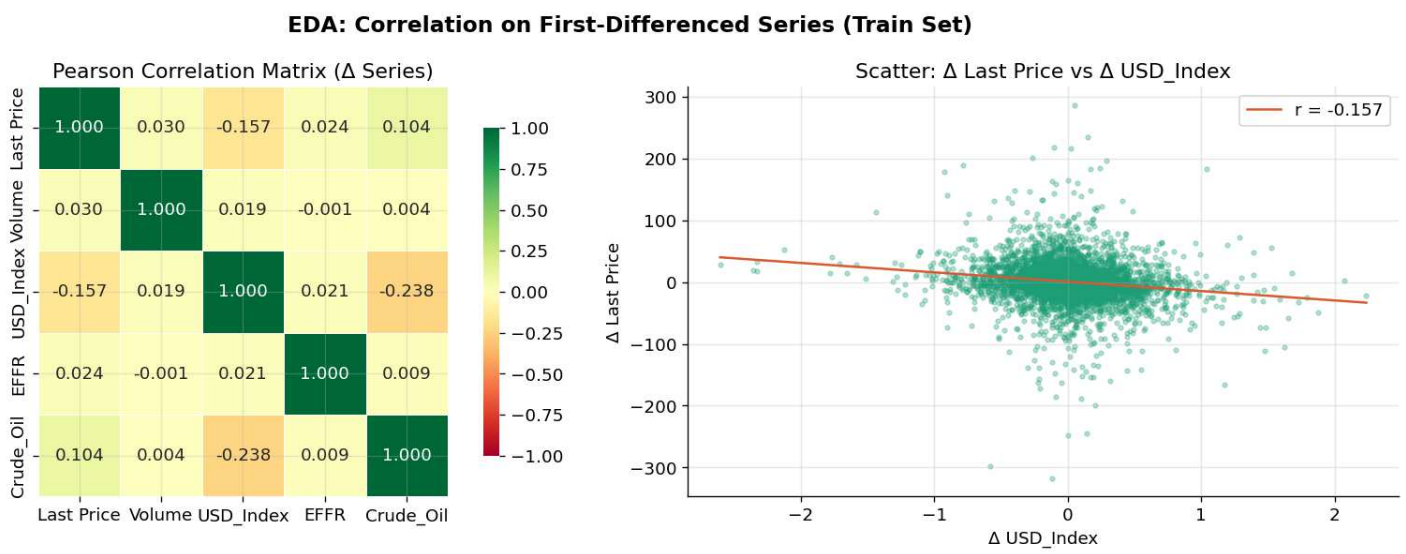


Figure 4.2: Pearson correlation matrix on the differenced series, and a scatter plot of the most strongly correlated pair

4.1.4. Heteroscedasticity and structural-break detection

To assess the market's underlying instability, the study computes 60-day rolling volatility, matching the lookback window used by the DL models.

The results show clear heteroscedasticity: 909 trading days (the top 20% most volatile quintile) record a peak rolling standard deviation of USD 718.1/ton. This sustained and widening volatility confirms that the coffee price series does not follow a stable trajectory but is repeatedly disrupted by structural breaks, particularly concentrated in 2023–2026. This concentration of variance shocks eliminates any feasibility of the homoscedasticity assumption and necessitates architectures capable of dynamically updating and controlling hidden states — a claim directly validated by GRU's superior performance in Section 4.2.2.

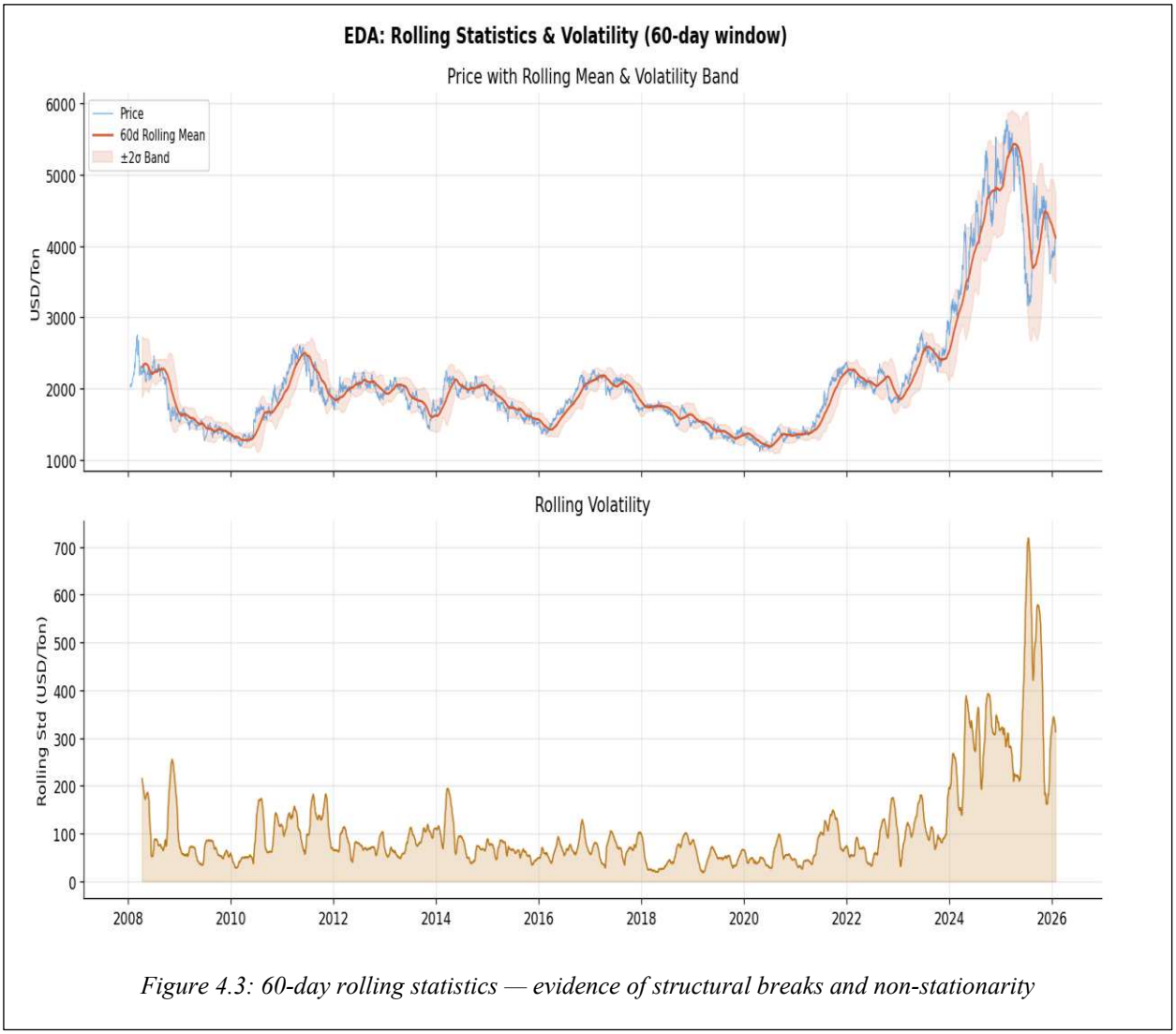


Figure 4.3: 60-day rolling statistics — evidence of structural breaks and non-stationarity

4.2. Out-of-sample model performance

Performance comparison is conducted uniformly on the 300-day test set (20 November 2024 to 28 January 2026), a period during which the coffee market underwent severe structural breaks and reached a historic peak of USD 5,765/ton, creating the most demanding possible evaluation environment.

4.2.1. Machine learning group

Random Forest achieves the best performance among the ML models (MAE = 169.39 USD/ton, R^2 = 0.888), outperforming LSTM. Bagging reduces forecast variance, and averaging across 200 trees produces substantial stability. Because Random Forest is trained on the entire Gross Train set (price range up to USD 5,334/ton — Section 3.3, Principle 4), it has “seen” a substantially

higher price range than the DL models, which update gradients only on Pure Train (maximum price USD 2,751/ton). This result should therefore be read as evidence of bagging's ensemble strength when trained on data that fully covers the current price regime, not as

evidence against the sequential-inductive-bias hypothesis — that hypothesis is fairly tested only in the XGBoost-versus-Random-Forest comparison (identical training range), where Random Forest still clearly wins (R^2 = 0.888 versus 0.465).

SVR with an RBF kernel achieves MAE = 229.28 USD/ton and R^2 = 0.771, consistent with theoretical

Model	Group	MAE (USD/ton)	RMSE (USD/ton)	MAPE (%)	R^2
GRU	DL	96.55 ★	125.91 ★	2.14% ★	0.967 ★
Random Forest	ML	169.39	232.09	3.47%	0.888
LSTM	DL	226.35	262.92	4.72%	0.856
SVR	ML	229.28	331.34	4.63%	0.771
XGBoost	ML	425.77	506.74	8.91%	0.465
DNN	DL	1,081.47	1,131.19	23.23%	-1.665

Table 4.2: Aggregate out-of-sample performance matrix over the 300-day test window (★ = best)

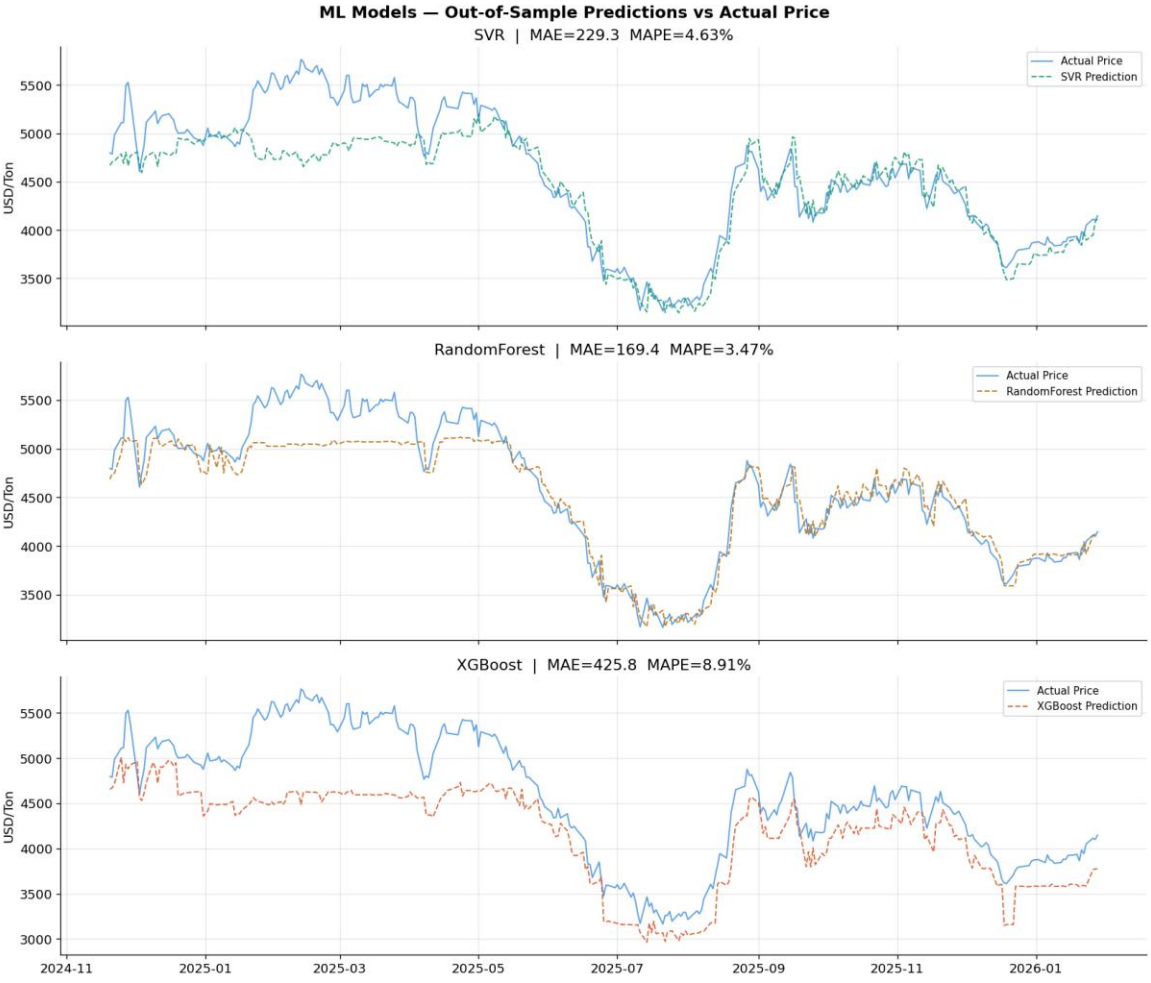


Figure 4.4: Out-of-sample predictions of the ML models — 300-day test window

expectation: the RBF kernel operates on a static (lag-1) feature space and cannot capture longer time-series dependencies, despite being trained on the same Gross Train range as Random Forest.

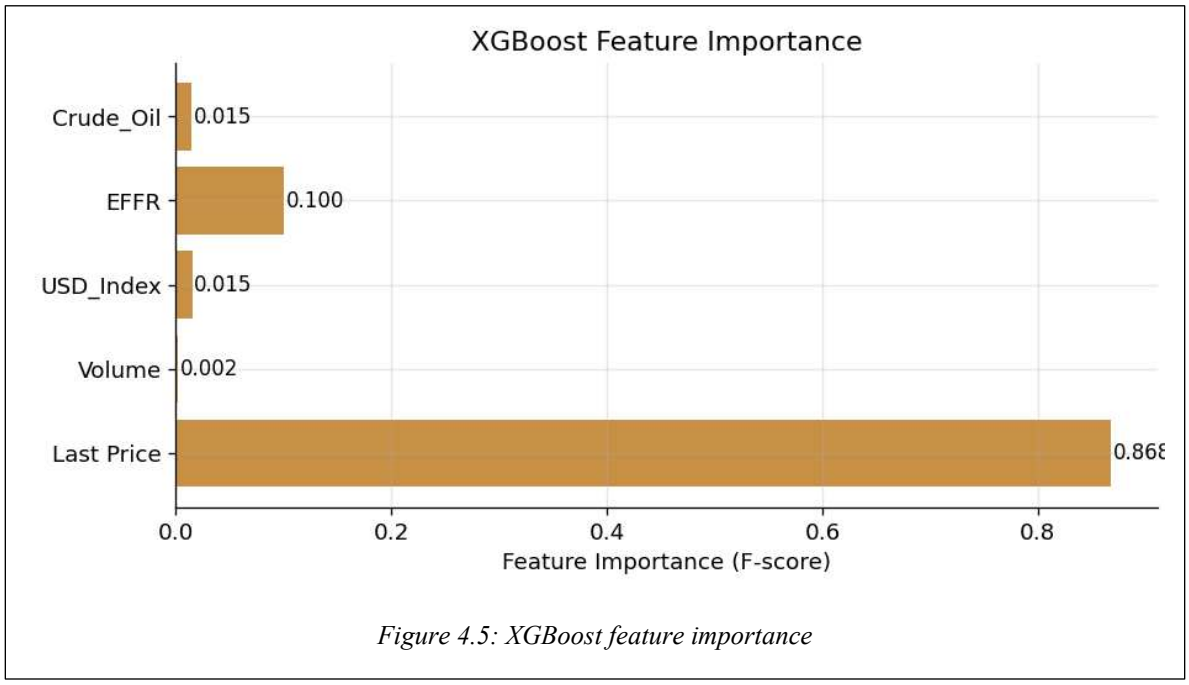


Figure 4.5: XGBoost feature importance

XGBoost records the weakest ML performance (MAE = 425.77, R^2 = 0.465), despite training on the same Gross Train set as Random Forest. This is a notable finding: two tree-based models share identical training-data scope, yet XGBoost's MAE is nearly 2.5 times worse. The core cause lies in the difference between ensembling mechanisms: gradient boosting builds trees sequentially, each optimizing the residual of the current ensemble; once the test-period prices move far outside the training distribution (56 of 300 days), residuals at these extrapolation points become systematically one-directional, and subsequent trees in the boosting chain amplify rather than correct this local bias. Random Forest's independent averaging across 200 uncorrelated trees, by contrast, partially neutralizes this systematic bias. This is empirical evidence that, under structural market breaks, bagging is more stable than boosting — a finding that adds to the literature comparing ensemble methods on non-stationary financial data.

The feature-importance chart (Figure 4.5) confirms that Last Price (lag-1) dominates XGBoost's split decisions, evidencing the strong autoregressive nature of the coffee price series; the macro variables (USD_Index, EFFR, Crude_Oil, Volume) contribute only marginally — consistent with the weak-correlation finding in Section 4.1.3 and reinforcing the conclusion that macro influence on coffee prices is nonlinear and difficult for tree-based models to exploit fully using lag-1 features alone.

4.2.2. Deep learning group

GRU dominates the full evaluation table, with MAE = 96.55 USD/ton, MAPE = 2.14%, and R^2 = 0.967 — not only the best DL model but the best model overall among all six architectures benchmarked, including Random Forest despite the latter's wider training-range advantage. This is compelling evidence for the sequential-inductive-bias hypothesis: although GRU "sees" a narrower price range during gradient updates (up to USD 2,751/ton), its gating mechanism (update gate, reset gate) still allows it to learn the relative dynamics of the series — percentage change, rate of change, local trend — rather than merely memorizing absolute values previously observed, allowing it to generalize better to unseen price regions than tree-based models.

A notable, theoretically unexpected finding is that GRU outperforms LSTM (R^2 = 0.967 versus 0.856), even though LSTM in principle has richer representational capacity through its separate cell state and three-gate mechanism. Two complementary explanations are proposed. First, given the limited Pure Train size (3,597 sequences) and a modest five-dimensional feature space,

LSTM's additional cell-state parameters carry a robustness check in Section 4.4: LSTM shows markedly

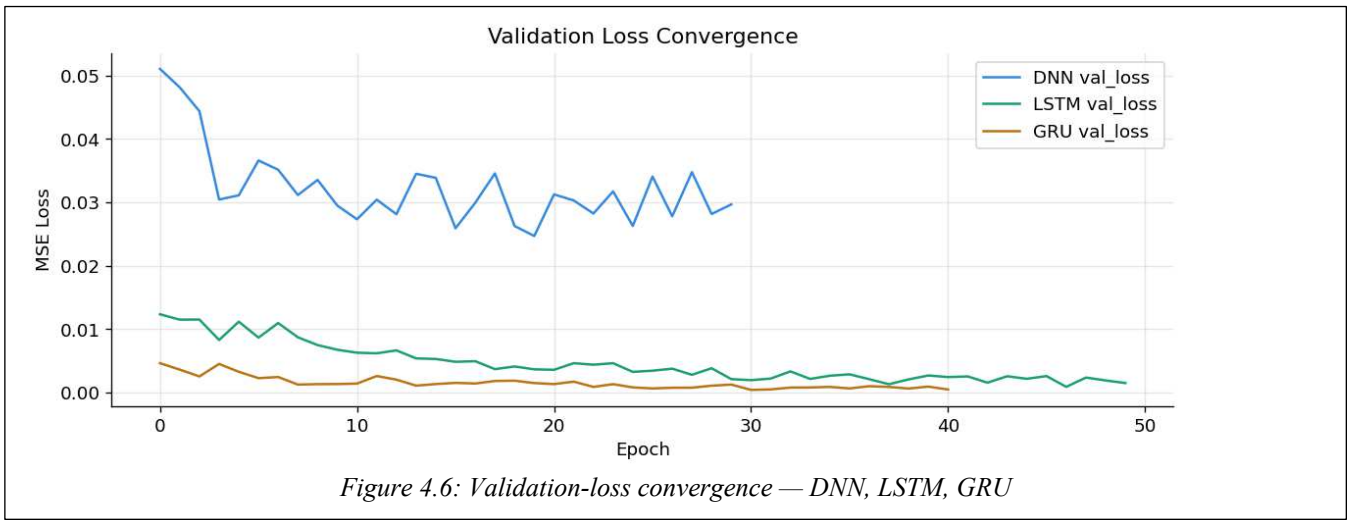


Figure 4.6: Validation-loss convergence — DNN, LSTM, GRU

relatively higher overfitting risk than the representational benefit they provide — GRU's smaller parameter count acts as a better inductive regularizer at this data scale. Second, this result is consistent with the higher sensitivity to random initialization than GRU (CV(MAE) = 39.52% versus 24.80%), suggesting a rougher loss landscape for LSTM on this data, such that

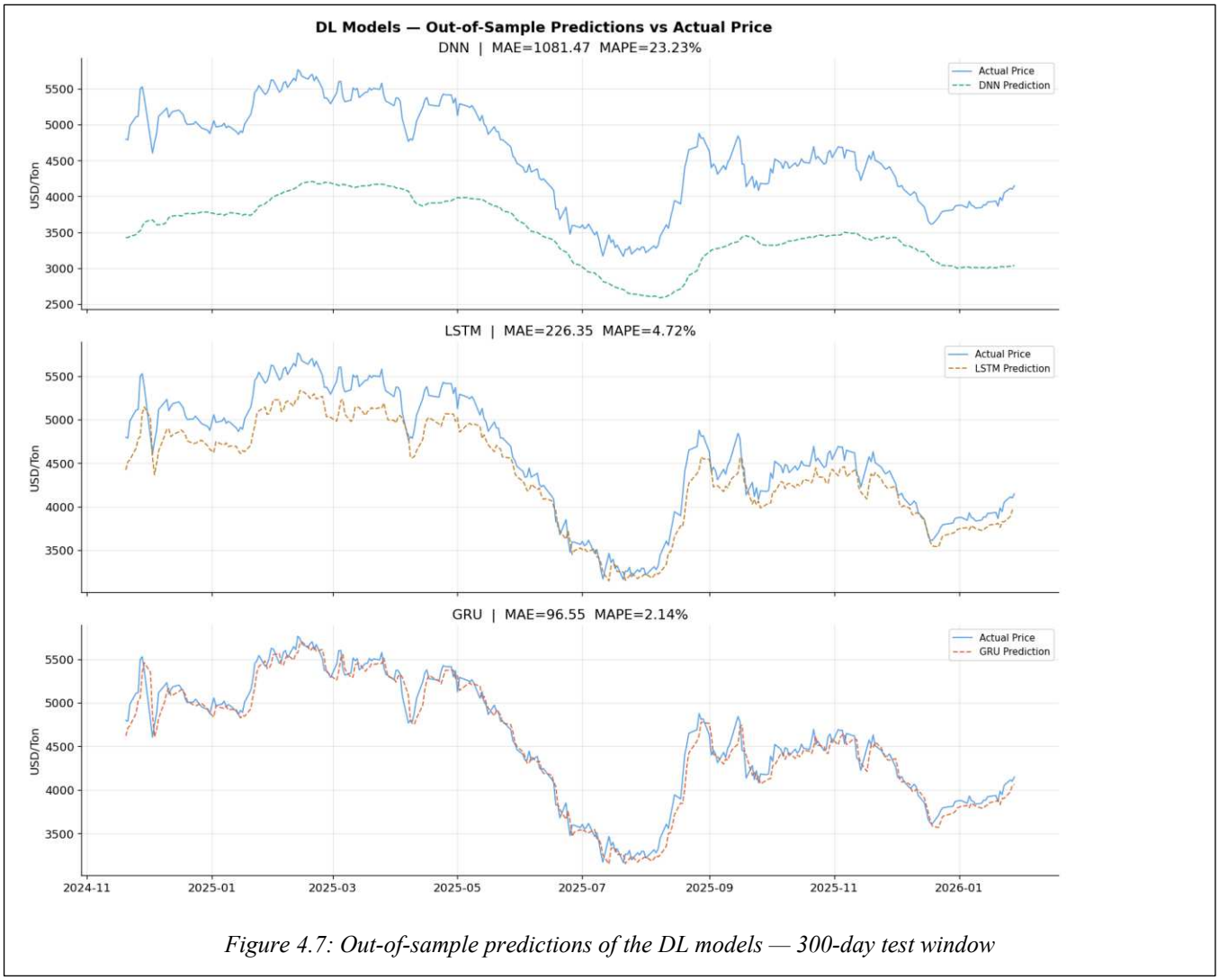


Figure 4.7: Out-of-sample predictions of the DL models — 300-day test window

the seed = 42 (default) result is not representative of the architecture's average potential.

DNN records the weakest performance of all six models (MAE = 1,081.47, R^2 = −1.665), worse than simply forecasting the test-set mean. The purely feedforward network has no sequential memory mechanism — each prediction depends only on a flattened feature vector at time t and cannot integrate dynamic information from the preceding 60 days in proper temporal order (even though it technically "sees" all 300 input values, it has no prior to distinguish their temporal order). This result is the study's most decisive empirical evidence: in financial time-series forecasting under structural breaks, sequential inductive bias is a necessary condition that

cannot be substituted by network depth or parameter count.

4.2.3. Overall ML versus DL comparison

Overall, the results show a clear architecture-driven split, but one that does not map cleanly onto the ML/DL boundary as initially hypothesized. Three findings deserve emphasis.

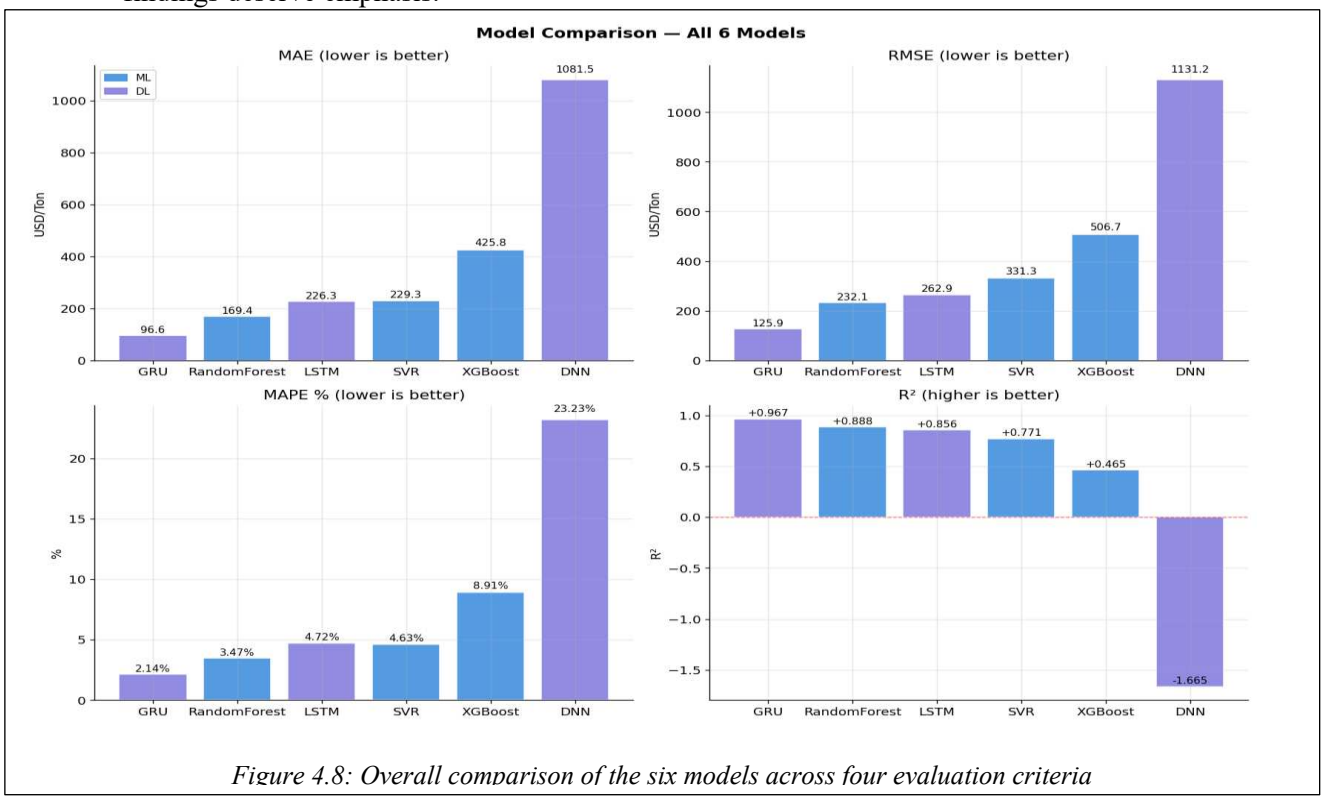


Figure 4.8: Overall comparison of the six models across four evaluation criteria

First, the strongest sequential architecture (GRU) wins outright, even against the ML model with a training-range advantage (Random Forest) — strongly reinforcing the study's core thesis.

Second, not every DL architecture beats ML: DNN (R^2 = −1.665) loses to all three ML models, including SVR (R^2 = 0.771), the second-weakest ML model. This demonstrates that DL's strength stems from sequential recurrence, not from belonging to the "deep learning" category or from parameter count.

Third, and most subtle: LSTM — the DL architecture with the richer cell state — loses to Random Forest. As discussed in Sections 3.3 (Principle 4) and 4.2.1, part of this gap originates from the training-data-scope asymmetry rather than pure architectural evidence. The fairest comparison for isolating pure sequential advantage is GRU/LSTM (trained on Pure Train) versus DNN (also trained on Pure Train) — and here, both GRU and LSTM decisively outperform DNN, firmly confirming the sequential-bias hypothesis under identical training-data conditions.

4.3. Diebold–Mariano test — statistical validation

To test the statistical significance of the performance gaps, the Diebold–Mariano test is applied to every comparison against the best model (GRU), with HAC variance correction using a Newey-West/Andrews bandwidth of 5.

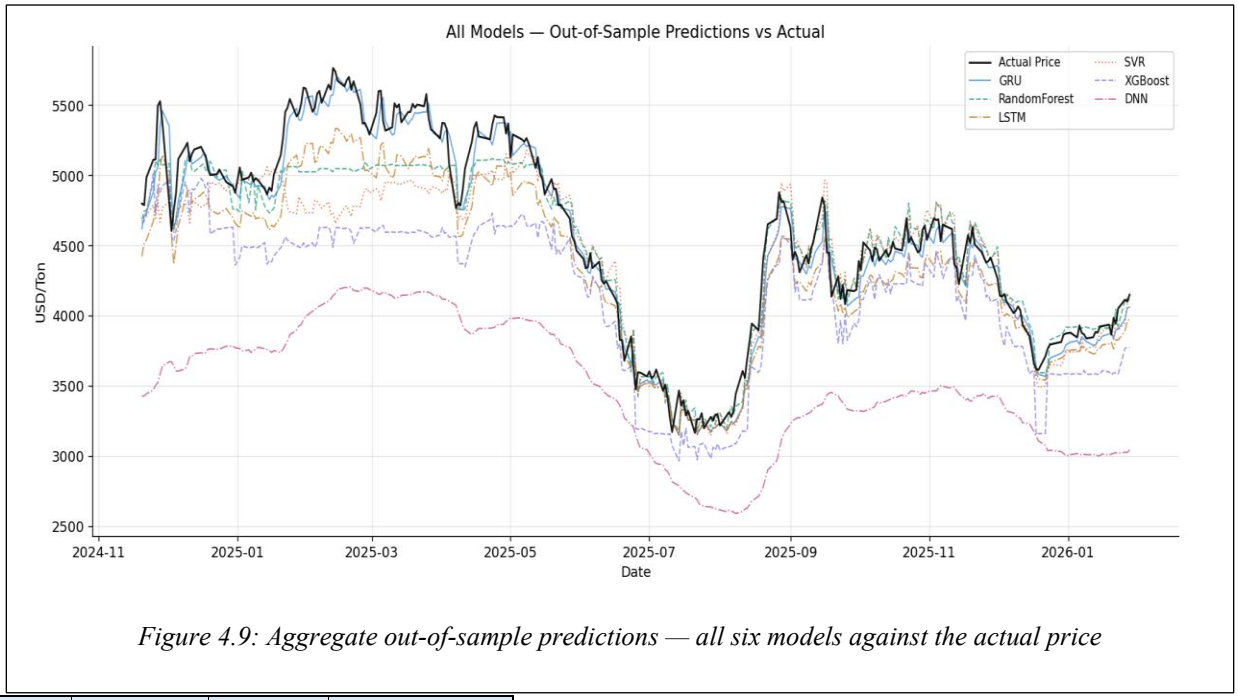


Figure 4.9: Aggregate out-of-sample predictions — all six models against the actual price

Comparison	DM statistic	p-value	Conclusion
GRU vs. Random Forest	-3.123	0.0018	significant at 1%
GRU vs. LSTM	-7.615	0.0000	Significant at 1%
GRU vs. SVR	-3.476	0.0005	Significant at 1%
GRU vs. XGBoost	-5.633	0.0000	Significant at 1%
GRU vs. DNN	-13.325	0.0000	Significant at 1%

Table 4.3: Diebold–Mariano test results — GRU versus the five other models

The DM test results provide strong econometric evidence for GRU's superiority across the entire benchmark. All five pairwise comparisons - GRU vs. LSTM, SVR, XGBoost, DNN, and Random Forest — reject the null hypothesis of equal predictive accuracy at the 1% significance level ($p < 0.01$ in every case). The margin against Random Forest, GRU's closest competitor overall, is confirmed at $p = 0.0018$, consistent with the two models producing the smallest MAE gap in the benchmark (72.84 USD/ton, versus gaps of hundreds to thousands of USD/ton against the remaining models). This result should be read alongside the training-data-scope advantage Random Forest enjoys (Sections 3.3 and 4.2.1): Random Forest was trained on a wider price range than GRU, so part of the residual gap between the two models may reflect this asymmetry rather than architecture alone. Even accounting for this, the statistical evidence across all five comparisons confirms GRU as the best-performing architecture in the study.

4.4. Architecture robustness — multi-seed check

Seed	LSTM — MAE	LSTM — R^2	GRU — MAE	GRU — R^2
0	131.93	0.9437	154.14	0.9141
7	211.95	0.8771	101.33	0.9640
21	353.71	0.6321	97.54	0.9659
42	226.35	0.8560	96.55	0.9670
99	376.92	0.5987	151.40	0.9229

Table 4.4a: Detailed multi-seed robustness results — LSTM and GRU across five seeds

Model	MAE mean	MAE std	R^2 mean	R^2 std	CV(MAE)
LSTM	260.17	102.83	0.7815	0.1555	39.52% — unstable
GRU	120.19	29.81	0.9468	0.0260	24.80% — unstable, though markedly lower than LSTM

Table 4.4b: Aggregate multi-seed robustness results — mean ± std across five seeds

Horizon	Date	Forecast (USD/ton)	Risk Floor	Risk Ceiling	Margin ±
T+1	29/01/2026	4,087.25	3,840.46	4,334.04	±246.79
T+2	30/01/2026	4,049.11	3,700.10	4,398.13	±349.01
T+3	02/02/2026	4,003.74	3,576.28	4,431.19	±427.45
T+4	03/02/2026	3,960.02	3,466.44	4,453.60	±493.58
T+5	04/02/2026	3,916.78	3,364.94	4,468.62	±551.84
T+6	05/02/2026	3,873.64	3,269.13	4,478.15	±604.51
T+7	06/02/2026	3,830.48	3,177.53	4,483.43	±652.95

Table 4.5: 7-day recursive forecast trajectory and 95% CI risk corridor

The robustness check is an important finding reported honestly, even where it does not fully favour the study's central narrative. Both LSTM (CV(MAE) = 39.52%) and GRU (CV(MAE) = 24.80%) exceed the 15% stability reference threshold [18]. The study does not conceal this behind a favourable single-seed result, and offers three interpretive points.

First, this degree of instability reflects a genuine property of the underlying problem rather than a design flaw. The test period (2024–2026) contains severe structural breaks (Section 4.1.4), and a substantial share (18.7%) of test observations fall outside the learned distribution (Section 3.3). Under these conditions, the loss landscape of a heavily parameterized recurrent network becomes rougher, so different weight initializations converge to local optima of markedly different generalization quality — a phenomenon documented in the literature on RNN training over non-stationary financial data [13, 18].

Second, GRU is consistently more stable than LSTM (CV(MAE) 24.80% versus 39.52%, a gap of nearly 15 percentage points). Table 4.4a shows that four of five GRU seeds (0, 7, 21, 42) achieve R^2 in the range 0.91–0.97 — a narrow band, uniformly high — with only seed 99 (R^2 = 0.9229) as a mild outlier. LSTM, by contrast, shows a clear degradation trend, with two of five seeds (21 and 99) falling to R^2 below 0.65, a substantial drop relative to the remaining three seeds (R^2 0.86–0.94). This reinforces the finding in Section 4.2.2: GRU's edge over LSTM is not confined to one fortunate initialization but is a tendency that recurs across most independent initializations.

Third, regarding deployment implications: the robustness check reinforces the decision to select GRU, rather than LSTM, as the primary model for the recursive forecasting system in Section 4.5 — not only because GRU achieves the best performance at the default seed, but because it has a higher probability of reproducing comparable results when retrained in the future (e.g., on a periodic retraining schedule with new data). This is an important practical criterion for a forecasting system operated continuously in a business environment, extending beyond evaluation on a single performance metric.

4.5. Recursive forecast and risk-management implications

Validated by the DM test (Section 4.3) and the robustness check (Section 4.4), the GRU model is deployed for a 7-day-ahead recursive multi-step forecast, using RMSE_test = 125.91 USD/ton. The system follows a ceteris paribus assumption: all exogenous macro variables are held fixed at their value at time T (the limitations of this assumption are discussed in Section 5.2).

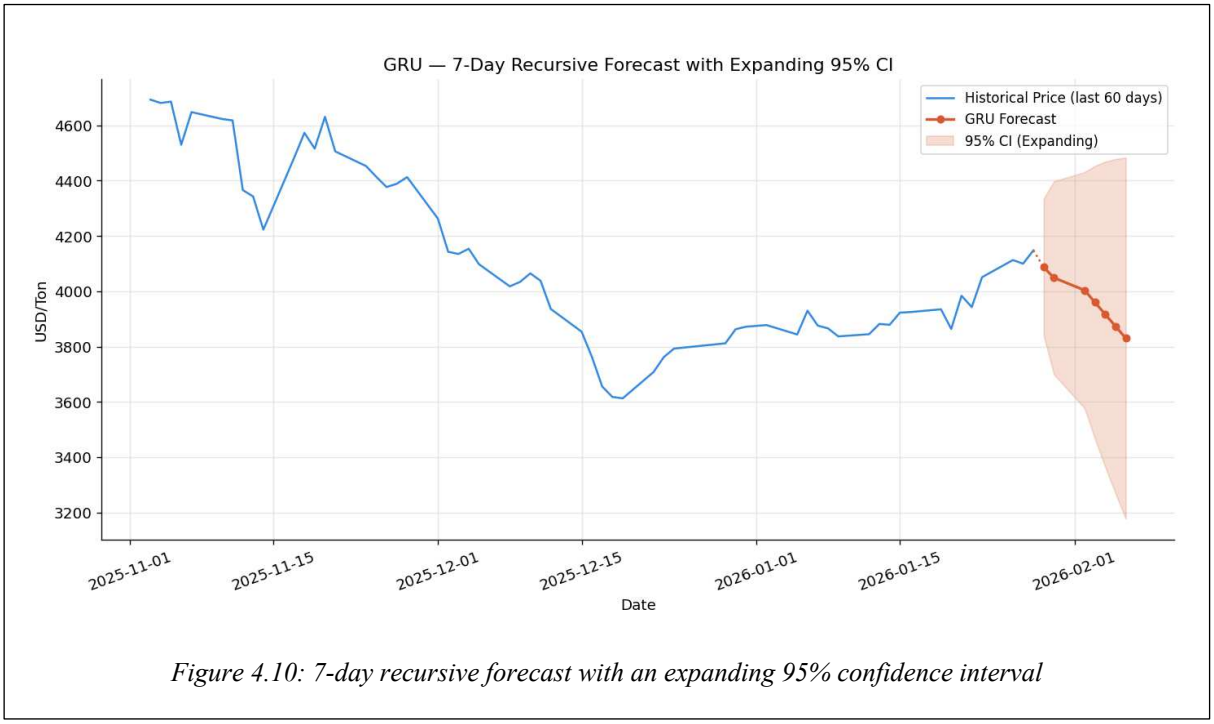


Figure 4.10: 7-day recursive forecast with an expanding 95% confidence interval

Forecast variance is expanded nonlinearly, proportional to the square root of the forecast horizon: Margin(h) = $Z_{0.975}$ × RMSE_test × √h = 1.96 × 125.91 × √h. The risk band widens from ±246.79 USD/ton at T+1 to ±652.95 USD/ton at T+7, quantifying the decline in forecast reliability at longer horizons. This band rests on a symmetric normal-distribution assumption — a limitation discussed in Section 5.2 (Limitation 1).

Strategic implication for exporters: unlike earlier model runs that identified a price-recovery trend, the GRU model in this run identifies a mild bearish trajectory in the short term — a projected decline from USD 4,087.25/ton (T+1) to USD 3,830.48/ton (T+7), roughly a 6.3% decrease over seven trading days. This is specific to the data run as of 28 January 2026 and should be updated as new trading data arrives; it should not be read as a medium- or long-term forecast.

For exporters holding coffee inventory, the projected downward trend supports prioritizing forward-sale contracts near the T+1 price level (USD 4,087.25/ton) to lock in current pricing rather than delaying sales. The T+7 Risk Ceiling of USD 4,483.43/ton establishes an upper bound for unexpected price-recovery scenarios and can serve as a reference point for hedging with put options.

For businesses purchasing raw coffee, the projected decline supports a controlled, short-term delay in procurement, using the T+7 Risk Floor of USD 3,177.53/ton as a reference for limit orders. Because the margin widens with the square root of the horizon, however, delaying procurement beyond three to four trading days causes uncertainty to outweigh the potential benefit of waiting for a further price decline; buyers should weigh their specific risk tolerance before adopting this strategy.

5. Conclusion and Future Research

5.1. Conclusion

This study establishes a comprehensive empirical benchmarking framework quantifying the forecasting capability of six architectures — from traditional ML to multivariate deep learning — for Vietnam's export coffee price series under severe structural breaks. Based on 300-day out-of-sample error analysis, Diebold–Mariano testing, and a multi-seed robustness check, the study draws five core conclusions.

First, the lightweight gated recurrent architecture (GRU) prevails. GRU achieves the best performance among all six models (MAE = 96.55 USD/ton; MAPE = 2.14%; R^2 = 0.967), significantly outperforming all five remaining models — LSTM, SVR, XGBoost, DNN, and Random Forest — at the 1% level. Its compact gating mechanism (update gate, reset gate) demonstrates superior generalization even when trained on a narrower price range than its strongest ML competitor.

Second, sequential architecture — not network depth — is decisive within DL. The purely feedforward DNN records the weakest performance in the entire benchmark ($R^2 = -1.665$), losing to all three ML models. Tested under the same training-data range as LSTM/GRU, this evidence confirms decisively that sequential inductive bias, not the number of hidden layers or parameters, determines forecasting capability for financial time series under structural breaks.

Third, GRU is systematically more stable than LSTM, not merely more accurate. Contrary to the initial theoretical expectation that LSTM's richer cell state should give it an advantage, the empirical results consistently show GRU outperforming LSTM both in point accuracy (R^2 = 0.967 versus 0.856) and in stability across multiple initializations (CV(MAE) = 24.80% versus 39.52%). This suggests that, at the training-data scale typical of agricultural commodity forecasting (thousands, not millions, of observations), a lighter architecture can deliver a practical advantage that outweighs the theoretical benefit of greater representational capacity.

Fourth, bagging and boosting diverge sharply in their extrapolation limits. Despite identical training-data scope, Random Forest (R^2 = 0.888) substantially outperforms XGBoost (R^2 = 0.465). This extends the classical finding on the extrapolation limits of tree-based models: sequential boosting amplifies one-directional residual bias in the extrapolation zone, while Random Forest's independent bagging partially neutralizes it — a relevant consideration when selecting an ensemble method for commodities with strong upward trends.

Fifth, methodological transparency is a precondition for a trustworthy benchmark. The study proactively identifies and discloses two limitations that could affect interpretation: (i) the asymmetry in training-data scope between ML (Gross Train) and DL (Pure Train), which means the LSTM/GRU-versus-Random-Forest comparison must be read more cautiously than comparisons within the DL group or within the ML group; and (ii) seed sensitivity exceeding the reference threshold for both recurrent architectures, reported in full rather than presented through a single favourable run. This transparent approach, while it may reduce the apparent tidiness of the narrative, provides a higher standard of scientific reliability for the conclusions drawn.

Building on these quantitative results, the system translates forecast error into a recursive forecasting matrix with an expanding confidence interval, providing explicit Risk-Floor / Risk-Ceiling parameters that support an exporter's finance function in structuring forward contracts and optimizing cost of goods sold under a projected short-term downward price trend.

5.2. Limitations and directions for future research

Despite the practical value of this benchmarking framework, six limitations should be acknowledged transparently, together with directions for addressing them in future work.

Limitation 1 — The symmetric-distribution assumption in the risk band (symmetric CI fallacy)

The hedging matrix uses $\text{Margin}(h) = Z_{0.975} \times \text{RMSE_test} \times \sqrt{h}$ to set the Risk Floor/Ceiling, implicitly assuming forecast errors follow a symmetric normal distribution with variance growing as the square root of the horizon — analogous to Brownian motion. Commodity markets, however, exhibit fat tails, volatility clustering (ARCH/GARCH effects), and asymmetric error distributions. Future work should incorporate Value at Risk (VaR) at the 99% level and Conditional Value at Risk (CVaR), along with GARCH

modelling of GRU's residuals, to construct an asymmetric risk band.

Limitation 2 — Asymmetric training-data scope between ML and DL

As discussed in Sections 3.3 (Principle 4) and 4.2.3, the ML group is trained on the full Gross Train set (4,243 observations, price range up to USD 5,334/ton), while the DL group updates gradients only on Pure Train (3,597 sequences, up to USD 2,751/ton), owing to the need to reserve a Validation set for early stopping. Although theoretically justified [17], this design introduces a confound when directly comparing Random Forest with LSTM/GRU: part of the observed performance gap may originate from the data-scope advantage rather than architecture alone. Future work should adopt a two-stage training strategy: (i) use early stopping on Pure Train/Validation to determine the optimal epoch count, then (ii) retrain the DL models from scratch on the full Gross Train set using that fixed epoch count — eliminating the data-scope gap between the two model families entirely.

Limitation 3 — Sensitivity to random initialization beyond the reference threshold

The multi-seed robustness check (Section 4.4) records CV(MAE) = 39.52% for LSTM and 24.80% for GRU, both above the 15% stability reference. Although the study explains the likely cause — structural breaks combined with an 18.7% out-of-distribution test share — this level of instability limits the reproducibility of the reported performance figures when the model is retrained in production. Future work should: (i) apply ensemble averaging — training multiple GRU models under different seeds and averaging their forecasts to reduce variance; (ii) experiment with gradient clipping (clipnorm) and learning-rate scheduling (ReduceLROnPlateau) to smooth the loss landscape; and (iii) increase early-stopping patience or apply more carefully controlled weight initialization (e.g., a customized He/Xavier scheme suited to the characteristics of financial data).

Limitation 4 — Boundaries of the feature space

The five-dimensional feature space captures major macro-financial drivers well but includes no variable directly representing speculative fund behaviour or market sentiment. Future work should incorporate CFTC Commitments of Traders (COT) data and news-sentiment analysis derived from large language models to develop an early-warning system.

Limitation 5 — The ceteris paribus assumption in recursive forecasting

The recursive forecasting process holds all exogenous macro variables fixed at their final observed value at time T, reducing the realism of the risk band as actual macro variables evolve over the 7-day forecast horizon. Future work should integrate a Vector Autoregression (VAR) model to forecast the full macro variable space jointly before feeding it into GRU, following a two-stage VAR → GRU structure.

Limitation 6 — Scope of the out-of-sample test

The 300-day test window (November 2024 – January 2026) reflects a specific phase of the coffee market. Benchmark results may differ when evaluated over other market periods. Walk-forward validation with a rolling window would provide a more comprehensive assessment of result stability across market regimes, and would also allow the seed-stability findings from Limitation 3 to be re-tested directly across multiple, independent market periods.

Statements and Declarations

Competing Interests: The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability: The Robusta coffee price data that support the findings of this study are available from ICE Futures Europe. The macroeconomic series (effective federal funds rate, U.S. Dollar Index, WTI crude oil price) are available from the Federal Reserve Economic Data (FRED) database and the U.S. Energy Information Administration. The compiled dataset and code used for model training and evaluation are available from the corresponding author upon reasonable request.

Author Contributions: Both authors contributed to the study conception and design. Material preparation, data collection, model implementation, and analysis were performed by both authors. The first draft of the manuscript was written by both authors, who read and approved the final manuscript.

Ethics Approval: Not applicable. This study uses publicly available market and macroeconomic data and does not involve human participants, human data, or animals.

References

[1] International Coffee Organization: Coffee Market Report, January 2024. ICO, London (2024)

[2] Vietnam General Department of Customs: Vietnam Coffee Export Statistics 2023. General Department of Customs, Hanoi (2024)

[3] Tran, V.H., Nguyen, T.M.: Global factors affecting Vietnamese coffee export prices: an empirical analysis. Int. J. Econ. Financ. Issues 13, 145–158 (2023)

[4] Vo, H.T., Le, M.N.: Chaotic dynamics in Vietnamese coffee price series: implications for forecasting. Chaos Solitons Fractals 168, 113145 (2023)

[5] Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)

[6] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. 14, 199–222 (2004)

[7] Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263 (1995)

[8] Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001)

[9] Divina, F., García Torres, M., Gómez Vela, F.A., Vázquez Noguera, J.L.: A comparative study of time series forecasting methods for short term electric energy consumption prediction in smart buildings. Energies 11, 2906 (2018)

[10] Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[11] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. 2, 359–366 (1989)

[12] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9, 1735–1780 (1997)

[13] Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1310–1318 (2013)

[14] Tran, M.V., Nguyen, H.T.: LSTM-based coffee export price forecasting for Vietnam. J. Forecast. 39, 1289–1305 (2020)

[15] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of EMNLP, pp. 1724–1734 (2014)

[16] Tay, F.E., Cao, L.: Application of support vector machines in financial time series forecasting. Omega 29, 309–317 (2001)

[17] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

[18] Vo, M.T., Nguyen, H.P.: Sensitivity analysis of LSTM hyperparameters in agricultural price forecasting. Neural Process. Lett. 55, 3456–3472 (2023)