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.

Does Machine Learning Beat ARIMA in Foreign Exchange? Out-of- Sample Evidence on Trading Activity, Volatility and Prices for EUR/USD, GBP/USD and USD/CAD, 2010–2021

Shadia Saeed Alqahtani

ai.forex2030@gmail.com

King Abdulaziz University

Research Article

Keywords: exchange-rate forecasting, machine learning, random forest, gradient boosting, ARIMA, realised volatility, tick volume, forecast evaluation, Diebold–Mariano test, multiple testing

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11409815/v1

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

Additional Declarations: No competing interests reported.

Article

Does Machine Learning Beat ARIMA in Foreign Exchange? Out-of-Sample Evidence on Trading Activity, Volatility and Prices for EUR/USD, GBP/USD and USD/CAD, 2010–2021

Shadia Saeed Alqahtani

King Abdulaziz University, Jeddah, Saudi Arabia
Correspondence: ai.forex2030@gmail.com
ORCID: https://orcid.org/0009-0009-6009-4671

An earlier version of this paper, posted on SSRN, reported results that the re-analysis presented here does not reproduce; this version supersedes it, and the SSRN posting is being revised accordingly.

Abstract: Machine-learning models are used more and more to predict financial markets, but whether they beat simple time-series models is rarely tested under strict out-of-sample conditions. This study compares a Random Forest and a Gradient Boosting Machine (GBM), each using 26 price-derived indicators, with ARIMA models and simple rules for EUR/USD, GBP/USD and USD/CAD, using daily data from 2010 to 2021 and an 890-day test period from February 2018 to July 2021. For next-day tick-volume activity the comparison depends on information. Without volume history the Random Forest correctly classifies 31.8–41.6% of test days, while ARIMA correctly classifies 61.5–63.8%; with volume history the difference between the two is not significant. The GBM forecasts five-day realised volatility less accurately than ARIMA for EUR/USD and USD/CAD and is not significantly different from it for GBP/USD. In the main design it beats a trailing-volatility baseline only for GBP/USD, where a constant forecast performs just as well. For the price, ARIMA does no better than the random walk. Seven robustness analyses, including GARCH and HAR benchmarks, yearly re-fits, a single data source and other horizons, leave these results unchanged. Of 114 tests, four machine-learning wins remain significant after the Holm correction, and all of them are against naive baselines. The results suggest that a price-based machine-learning forecast should beat a simple time-series model, on the same information and test days, before it is trusted. The code, a pinned data download and file hashes are released for reproduction.

Keywords: exchange-rate forecasting; machine learning; random forest; gradient boosting; ARIMA; realised volatility; tick volume; forecast evaluation; Diebold–Mariano test; multiple testing

JEL Classification: C45; C52; C53; F31; G17

1. Introduction

Foreign-exchange markets are huge: over-the-counter turnover averaged $9.6 trillion a day in April 2025, which is 28% more than in 2022 (Bank for International Settlements, 2025). Forecasts of exchange rates, trading activity and volatility help with hedging, liquidity management and risk

limits. The International Monetary Fund expects the use of artificial intelligence (AI) in capital markets to grow significantly in the near future. The Fund reports that equities and derivatives are the most likely areas to adopt AI, followed by fixed income and foreign exchange, and it warns that AI could change market structure by making algorithmic trading more powerful (Fund, 2024) . Whether machine-learning tools forecast better than the simpler models they replace is therefore an operational and supervisory question, not just an academic one.

Since (Meese & Rogoff, 1983) the random walk has been the benchmark that exchange-rate models find hard to beat out of sample. Several machine-learning studies report progress. Ridge-type learners that combine fundamentals beat the no-change forecast at a one-month horizon (Amat et al., 2018). Elastic-net, Random Forest and support-vector models built on fundamentals beat it at most horizons for many currencies (Colombo & Pelagatti, 2020) ,and machine-learning versions of fundamental models improve on it for the yen–dollar rate (Zhang & Hamori, 2020). Other studies are more cautious. Neural networks and gradient-boosted trees do not convincingly beat the random walk in level forecasts, and their directional advantage disappears without calendar dummies (Pfahler, 2022). A complex Ridge model beats linear regression mainly in short training windows and beats the random walk for only a few currencies (Kılıç, 2025) .These five studies use monthly data with macroeconomic fundamentals, and daily studies such as (Plakandaras et al., 2015) use broad sets of candidate predictors. This study adds a daily comparison in which the inputs are limited to price-derived technical indicators and the targets include trading activity and volatility. Every machine-learning model is compared with a time-series benchmark for the same target on the same test days, and the study declares a correction for multiple testing.

This study asks whether machine-learning models forecast foreign-exchange trading activity and volatility better than ARIMA and naive rules, out of sample, when the models use only price-derived information. It looks at EUR/USD, GBP/USD and USD/CAD with daily data from January 2010 to July 2021. Each pair puts the US dollar next to a different major currency, so the same design is tested on three separate markets; the three pairs are replications, not a sample of market regimes. The sample is split once, in time order: the models are trained on February 2010 to February 2018 and evaluated on February 2018 to July 2021, with hyperparameters tuned by time-series cross-validation inside the training period. The features are moving averages, the Relative Strength Index, MACD, their lags and rolling return volatility. A Random Forest forecasts the next

day's trading-activity class (low, medium or high tick volume, a proxy for activity, relative to its recent level), and a Gradient Boosting Machine (GBM) forecasts realised volatility over the next five trading days. Each machine-learning model is compared with an ARIMA model fitted to the same target and with a naive rule, using McNemar and Diebold–Mariano tests (Diebold & Mariano, 1995; McNemar, 1947) .For the price itself, ARIMA is compared with the random walk; machine-learning forecasts of returns are examined in a robustness analysis.

Against the time-series benchmarks the answer is no; against the naive rules the result is mixed. The Random Forest classifies next-day volume correctly on 31.8–41.6% of test days, compared with 61.5–63.8% for ARIMA and 61.7–64.6% for the rule that tomorrow's class equals today's ("today's class"), and it is significantly worse than ARIMA for every pair. Its accuracy drops by 41.0, 56.8 and 27.5 points from training to test for EUR/USD, GBP/USD and USD/CAD. The two sides also use different information: ARIMA uses past volume, while the Random Forest uses only price-derived features. When volume history is added, a regularised Random Forest reaches ARIMA's level, with no significant difference (Section 4.6), so the shortfall comes mainly from the price-derived inputs. The GBM forecasts volatility less accurately than ARIMA for EUR/USD and USD/CAD and is not significantly different from ARIMA for GBP/USD. At the five-day horizon it beats the naive trailing-volatility forecast only for GBP/USD, where a constant forecast equal to the training-period mean does as well.

For the price, the ARIMA model chosen for EUR/USD is a random walk. For GBP/USD the ARIMA model is worse than the random walk (Theil's U 1.010, $p = 0.0025$ before correction for multiple testing), and for USD/CAD the difference is not significant (U 1.005, $p = 0.23$).

Seven robustness analyses (return forecasts, GARCH and HAR benchmarks, regularised volume models, yearly re-fits, MetaTrader-only data, alternative seeds with bootstrap intervals, and other horizons and class definitions), plus a wider hyperparameter search, leave these conclusions unchanged. Over the full test period, by the Diebold–Mariano and McNemar tests, no machine-learning model significantly beats the random walk on returns, ARIMA or today's class on volume, or ARIMA, GARCH or HAR on volatility. The only significant machine-learning wins are against weak benchmarks: naive volatility forecasts, at five days for GBP/USD in the main design (and for other pairs, without adjustment, in sensitivity runs) and at one day for all three pairs (where early stopping kept only 1–18 boosting rounds). In addition, for GBP/USD returns a machine-

learning model beats an ARIMA model that is itself worse than the random walk. The single win against ARIMA in a subsample, for EUR/USD volatility in the 70-day COVID-19 window ($p = 0.041$), is unadjusted and does not repeat in the other pairs or on high-volatility days. Of 114 tests, four machine-learning wins survive the Holm correction, all against naive baselines.

The study makes three contributions. First, it offers a test that is built to avoid look-ahead, with benchmarks matched on both target and test days, and it looks explicitly at differences in information sets. Second, it extends the comparison to trading activity and volatility, which, unlike returns, persist long enough to be forecast. Third, it is reproducible: the code, a pinned download of the source data and a hash table that lists every input file are archived at Zenodo (Alqahtani, 2026).The core pipeline was rebuilt twice from a fresh download, and the robustness runs use the same verified code; the yearly re-fit matches the stored main-run predictions to within 5×10^{-9}. The results suggest that, when price-derived features are used, the burden of proof falls on the machine-learning model, and that the benchmark to beat is a simple time-series model rather than a weak one.

The evidence has limits: three currency pairs, one chronological split, only price-derived features, tick volume as a proxy for trading activity, and no trading costs, because the study evaluates forecasts, not trading strategies. Section 2 reviews the theory and evidence, Section 3 describes the data and methods, Section 4 reports the results, Section 5 discusses the implications for how institutions validate and govern forecasting models, and Section 6 concludes.

2. Theoretical Framework and Related Literature

2.1. Market Efficiency and Adaptive Markets

The efficient market hypothesis says that prices fully reflect the information available to people in the market (Fama, 1970) . Its weak form concerns the narrowest set of information, the history of prices itself. If markets are weak-form efficient, no rule that uses past prices can earn abnormal returns. Technical indicators such as moving averages, the Relative Strength Index and MACD are fixed functions of that history. Under the weak form they should not give any reliable information about where prices will go next, and the best guess for tomorrow's exchange rate is close to today's. Foreign exchange is a deep and liquid market, mostly run by professionals, so it is a plausible example of such a market.

Technical analysis is still widely used. (Menkhoff & Taylor, 2007) show that it is common among foreign-exchange professionals and look at the reasons why it persists. Reviewing the evidence on whether it makes money, (Park & Irwin, 2007) find that many studies report profitable technical trading rules in foreign exchange and other speculative markets, at least until the early 1990s, but that most of these studies suffer from data snooping, choosing rules after seeing the results, and problems in measuring risk and transaction costs. Whether technical indicators carry forecasting information that survives a careful out-of-sample test is therefore still open.

The adaptive markets hypothesis softens the efficiency view but does not reject it (Lo, 2004) .Efficiency changes over time with the people in the market, how intense competition is and the market conditions. Profitable patterns can show up, but they attract money and then fade as people exploit them. Any advantage a model gets from price patterns is therefore likely to be temporary, and a model fitted on one period should not be expected to keep it over a long out-of-sample period. Both views suggest that a lasting advantage over a forecast of no change is unlikely for price-based forecasts of exchange-rate changes.

Neither view says the same about the size of price changes or the amount of trading. Market efficiency is about the expected change in prices, not about how much they vary or how much trading happens, and knowing how much a price is likely to move says nothing about which way it will go. Volatility and trading activity can therefore be predictable in an efficient market. For this study, the price is the target on which no gain is expected. Volatility and volume are the targets where a real machine-learning advantage would be easiest to find, as long as the models face the strong simple benchmarks that exist for them.

2.2. The Random-Walk Benchmark

(Meese & Rogoff, 1983) set the standard that still frames exchange-rate forecasting. They found that the leading structural models of the time failed to forecast major exchange rates out of sample better than a random walk, even when the models were given the actual future values of their fundamentals. This result made the random walk the hurdle that every forecasting model has to clear: a model is judged on whether it does better than a forecast of no change, not on how well it fits the data in sample.

ARIMA models are the natural time-series form of this benchmark (Box et al., 2015). The ARIMA approach identifies, estimates and checks a model of how the series is related to its own past, and the random walk is the special case of an integrated model with no autoregressive or moving-average terms. An ARIMA model chosen by an information criterion is therefore a data-driven extension of the random walk: it can do better than the random walk only if the changes in the series have a linear pattern that the data support. When fitted to trading volume or realised volatility, an ARIMA model captures how long these quantities persist with few parameters, which makes it a transparent benchmark for each target in this study.

Later research has shown how hard the random walk is to beat. (Cheung et al., 2005) evaluated the exchange-rate models developed during the 1990s and found that no model consistently did better than the random walk across currencies, horizons and accuracy measures. In a comprehensive survey, (Rossi, 2013) concludes that whether exchange rates seem predictable depends on the predictors, the forecast horizon, the sample period, the model and the way forecasts are evaluated, and that the evidence of predictability often changes over time. A claim to beat the random walk therefore needs out-of-sample tests across currencies and periods, with a formal test of the difference in accuracy.

2.3. Machine-Learning Forecasting

The two machine-learning models in this study are ensembles of decision trees. A Random Forest grows many trees on bootstrap samples of the data and considers a random subset of predictors at each split, then averages the trees' predictions; averaging reduces variance while allowing for non-linear relationships and interactions among predictors (Breiman, 2001).A Gradient Boosting Machine instead adds trees one at a time, each fitted to the errors of the current model, and controls complexity through the learning rate and the number of trees (Friedman, 2001) .Both show the trade-off between bias and variance (Hastie et al., 2009) : flexible models reduce bias but increase variance, so how well they fit the training data says little about how accurate they are on new data. When the signal is weak compared with the noise, as in many financial series, variance can dominate, and the usual failure is overfitting: the model fits the noise in the training data and then loses accuracy out of sample. Regularisation, early stopping and validation on unseen data are the standard defences, and the gap between training and test performance is a direct sign of whether they have worked.

Large-scale comparisons give reason for caution. (Makridakis et al., 2018) compared machine-learning and standard statistical methods on a large set of monthly time series and found that the statistical methods were more accurate out of sample on both accuracy measures and at every forecast horizon they examined. They argued that new methods should be compared with simple statistical benchmarks, using properly designed out-of-sample tests, before claims of superiority are made.

Studies of machine learning in foreign exchange report mixed results, under conditions that differ from those examined here. (Amat et al., 2018) combine fundamentals using sequential ridge regression and exponentially weighted averaging, and they beat the no-change forecast at a one-month horizon. (Colombo & Pelagatti, 2020) apply elastic-net regression on splines, Random Forests and support-vector machines to monetary fundamentals for many currencies and beat the random walk at most horizons, with the random walk remaining competitive only at the shortest ones. (Zhang & Hamori, 2020) find that Random Forest, support-vector and neural-network versions of fundamental models outperform the random walk for the yen–dollar rate, although formal tests give mixed results. (Pfahler, 2022) using fundamentals for currency pairs of OECD countries against the US dollar, finds that neither neural networks nor gradient-boosted trees convincingly beat the random walk in level forecasts; their directional forecasts are significant, mainly for the neural networks, but the advantage goes away when calendar dummies are left out. In a Federal Reserve Board discussion paper, revised in 2026, (Kılıç, 2025) finds that a Ridge model with random Fourier features does better than linear regression in short training windows but beats the random walk for only a few of eight currencies, and its out-of-sample R^2 averaged across currencies never goes above zero. These five studies use monthly data and macroeconomic fundamentals. At daily frequency, (Plakandaras et al., 2015) forecast five exchange rates with support-vector regression on decomposed series, choosing predictors from a large candidate set, and report gains over alternative models.(Yıldırım et al., 2021) forecast the daily direction of EUR/USD with long short-term memory networks fed with technical indicators, macroeconomic indicators or both, and they compare these variants with one another rather than with a time-series or no-change benchmark. The case for machine learning in foreign exchange therefore rests mainly on information beyond the price, and it is still debated.

2.4. Forecasting Risk and Trading Activity

Unlike returns, exchange-rate volatility is persistent: large price changes tend to follow large changes, and calm periods tend to follow calm periods. The generalised autoregressive conditional heteroskedasticity (GARCH) model captures this clustering by letting the conditional variance depend on past squared shocks and on its own past value (Bollerslev, 1986) .The simplest form, GARCH(1,1), is very hard to beat: comparing a large number of volatility models out of sample, (Hansen & Lunde, 2005) found no evidence that it was outperformed by more sophisticated models for exchange-rate data. They also showed that volatility forecasts are better evaluated against realised volatility measured from intraday data than against noisy proxies built from daily returns. This study measures realised volatility from daily returns, because it does not use intraday data, so its volatility target is a noisier proxy of the kind they caution against (Section 3).

The heterogeneous autoregressive (HAR) model offers a second simple benchmark (Corsi, 2009) .In its original form it forecasts intraday-based realised volatility from its own past averages over a day, a week and a month, and although it is estimated by ordinary least squares it reproduces the long memory of volatility and forecasts it well. This study uses the model in a daily version, with the absolute daily return and trailing weekly and monthly volatility measured from daily returns (Section 3). Together with ARIMA models fitted to realised volatility, GARCH and HAR set the standard that a machine-learning volatility forecast must meet. A model that beats only a naive forecast has not shown an advantage over methods that practitioners already use.

Machine learning has nonetheless been reported to beat these benchmarks in some settings. (Christensen et al., 2023) find that regularised regressions, regression trees and neural networks forecast the realised variance of the stocks in the Dow Jones Industrial Average better than models in the HAR family, with larger gains at longer horizons. Their setting differs from the present one in two important ways: their realised variance is measured from intraday stock-price data, whereas this study uses daily foreign-exchange data, and their models use the history of realised variance itself rather than technical indicators of the price.

Trading activity raises a measurement problem that is specific to foreign exchange. The spot market is decentralised and traded over the counter, so there is no single record of how much is traded. Trading platforms instead report tick volume, the number of price updates in an interval. OANDA defines the volume of a candlestick as the number of prices created during the period it

covers (OANDA, n.d.),and the MetaTrader platform separates tick volume from trade volume in its price-history records (MetaQuotes, n.d.) .Tick volume is therefore a proxy for activity. Because it counts how often a provider updates its quotes, its level depends on the provider; the two sources used here report systematically different levels over the period in which they overlap (Section 3). Relative measures, which compare activity with its own recent level, reduce this problem. Most research on forecasting trading volume concerns equities at intraday frequency, where dynamic models that capture persistence and intraday patterns forecast better than simpler methods (Brownlees et al., 2011) . As Section 4 shows, trading activity in these pairs is also persistent, so a naive forecast that tomorrow's activity resembles today's is a demanding benchmark.

2.5. Evaluating and Comparing Forecasts

Comparing two forecasts requires a test of whether their difference in accuracy is larger than chance would produce. (Diebold & Mariano, 1995) proposed a test based on the series of differences between the losses of two forecasts, which works with general loss functions and with serial correlation in those differences. Such correlation is unavoidable when multi-day forecast horizons overlap. (Harvey et al., 1997) showed that the original test rejects too often in small samples and proposed a corrected statistic compared with Student's t distribution, which is now the usual form of the test. When two classifiers are evaluated on the same days, their accuracies are paired, and the right comparison looks at the days on which one model is correct and the other is not (McNemar, 1947).

Model selection also needs to respect the order of time. Standard k-fold cross-validation assigns observations to folds at random, so a model can be tuned on data that come after the data used to assess it. (Bergmeir & Benítez, 2012) note that this breaks the assumptions of cross-validation, although they found no practical consequences in their experiments, and they recommend a blocked form of cross-validation as the standard for time series. (Cerqueira et al., 2020) found that cross-validation can be applied to stationary series, but that for real-world series, which are often non-stationary, out-of-sample procedures that keep the time order give the most accurate estimates of forecasting performance. This study tunes hyperparameters on expanding-window (forward-chaining) splits of the training period, in which each validation block comes after the data used to fit the model, and it reports performance on a later hold-out period that plays no part in model selection.

Finally, a study that compares several models across several currency pairs, targets and robustness checks performs many tests, and the chance that at least one of them appears significant by accident grows with their number. (Holm, 1979) sequentially rejective procedure controls the probability of any false rejection within a defined family of tests and is uniformly more powerful than the Bonferroni correction. Section 3 defines the family used here; showing both the nominal and the adjusted results makes clear which conclusions survive the multiplicity of comparisons.

Taken together, this literature points to three gaps that the present study fills. First, the evidence that machine learning beats simple benchmarks in foreign exchange comes mainly from monthly data with macroeconomic fundamentals or from daily models with broad predictor sets; this study limits the inputs to price-derived indicators, so any gain must come from the price history alone. Second, machine-learning gains in volatility forecasting have been reported with intraday data for equities, and volume forecasting has been studied mainly for equities at intraday frequency; this study examines foreign-exchange trading activity and volatility, alongside the price, with daily price-derived inputs, targets that theory suggests, unlike returns, can be forecast. Third, daily studies that use technical indicators, such as (Yıldırım et al., 2021) compare machine-learning variants with one another, whereas a comparison is informative only when it includes the parts this review identifies as necessary: benchmarks that forecast the same target, time-ordered validation, formal tests of differences in accuracy and control for multiple comparisons. The next section describes a design that includes all of them.

3. Data and Methods

3.1. Data

The study uses daily data for EUR/USD, GBP/USD and USD/CAD from 1 January 2010 to 12 July 2021. No single freely available source covers this period with both prices and a volume measure, so two broker sources are joined. From 1 January 2010 to 3 December 2012 the daily bars are built from OANDA one-minute bars, aggregated into days that end at 17:00 New York time, the conventional daily roll of the foreign-exchange market; bars dated on weekends, which contain only a few quotes from outside the trading week, are dropped. From 4 December 2012 the data are MetaTrader daily bars. Each pair therefore has 762 days from OANDA and 2,231 from MetaTrader, and the whole test period (Section 3.2) comes from a single source. Both data sets

were downloaded from public repositories at fixed versions, and every source file is identified by a cryptographic hash (Section 3.6).

The two sources overlap on 1,927 days, from 4 December 2012 to 14 May 2020, which allows the join to be checked. Over the overlap, the median absolute difference between the daily opening prices of the two sources is 1.8, 3.3 and 2.8 pips for EUR/USD, GBP/USD and USD/CAD, the median difference between closing prices is 1.2 to 2.5 pips, and the correlation of daily changes in the closing price is 0.996 to 0.997. No bar in the final series has an open or close outside its own high–low range. As an independent check, the series were compared with the noon buying rates in New York published by the Federal Reserve Bank of St. Louis (FRED), which are used for validation only: the correlation of price levels over 2,883 common days is 0.9997 to 0.9998. The median absolute difference of 12 to 14 pips, and the lower correlation of daily changes (0.79 to 0.88), reflect the different time of day at which the two series are recorded.

Fourteen weekdays are missing from the MetaTrader segment, all of them 25 December or 1 January dates between 2012 and 2021, when the market is effectively closed; the OANDA segment has no missing weekdays. They are filled by linear interpolation between the neighbouring days, which changes 14 values in each of the five price and volume fields per pair, or 0.5% of the sample. Because these days are isolated and carry little trading, interpolation does not alter the dynamics of the series; a sample with many or consecutive gaps would require a different treatment. The final data set has 3,007 daily observations per pair.

Spot foreign exchange is traded over the counter, so there is no consolidated record of traded volume. Both sources report tick volume, the number of price updates in the period, and throughout this paper "volume" means tick volume, a proxy for trading activity rather than a measure of it. Tick-volume levels depend on the provider: over their first 250 common trading days, the median daily OANDA count was 1.76, 1.60 and 1.22 times the MetaTrader count for the three pairs, so the OANDA counts were multiplied by 0.5692, 0.6253 and 0.8189 before the two segments were joined. Levels also drift within a source over time. When the tertiles of raw tick volume in the training period are applied to the test period, only 0.1% to 8.2% of test days fall in the lowest class. For this reason the volume target is defined relative to recent activity (Section 3.2); a version based on raw tick-volume tertiles was estimated as a sensitivity check.

Table 1 summarises the data. Daily returns were less volatile in the test period than in the training period for EUR/USD and USD/CAD, and the five-day volatility target fell accordingly; for GBP/USD the mean volatility is almost unchanged, but the training period contains the 2016 referendum, which gives its returns a skewness of −1.81 and an excess kurtosis of 28.6. The augmented Dickey–Fuller test does not reject a unit root in log prices in the training period ($p = 0.37$ to 0.86) and rejects it for returns ($p < 0.001$). The median daily tick volume of the rescaled OANDA segment is less than a third of the MetaTrader median in the later training years, a drift that the relative volume target is designed to absorb.

Table 1. Descriptive statistics, training (22 February 2010 to 5 February 2018) and test (6 February 2018 to 5 July 2021) periods.

Statistic	EUR/USD training	EUR/USD test	GBP/USD training	GBP/USD test	USD/CAD training	USD/CAD test
Days	2,076	890	2,076	890	2,076	890
Daily log return, mean (%)	−0.005	−0.005	−0.005	−0.001	0.009	−0.002
Daily log return, standard deviation (%)	0.585	0.406	0.561	0.550	0.526	0.425
Skewness	−0.02	−0.29	−1.81	−0.23	0.13	0.20
Excess kurtosis	1.54	1.73	28.61	4.50	1.51	2.22
Minimum / maximum (%)	−2.56 / 3.02	−2.34 / 1.37	−8.71 / 3.03	−4.02 / 2.60	−2.15 / 2.50	−1.94 / 2.12
Five-day volatility target, mean (standard deviation)	0.53 (0.26)	0.37 (0.16)	0.49 (0.29)	0.48 (0.23)	0.47 (0.24)	0.39 (0.18)
Low volume class, share of days (%)	33.3	40.4	33.3	36.9	33.3	38.5
Median daily tick volume, thousands: OANDA segment (rescaled) / MetaTrader segment	25.4 / 83.4	– / 61.3	14.3 / 88.3	– / 83.2	19.6 / 72.6	– / 55.2
ADF p-value, log price / return (training)	0.37 / <0.001	–	0.56 / <0.001	–	0.86 / <0.001	–

Note: returns are daily log returns in percent over the task rows; the volatility target is defined in Section 3.2. The OANDA segment covers 2010–2012 and is rescaled to MetaTrader units (Section 3.1); the test period comes from MetaTrader only. Source: s24_descriptive_stats.csv.

3.2. Features, Targets and Sample Split

All variables follow one timing rule: a row dated t uses information available at the close of day t, and every target refers to day t+1 or later. Every model therefore produces a genuine forecast, and no input is computed from data after the forecast origin. The rule would be broken by, for example, a centred moving average, a scaler or class cut-off estimated on the whole sample, or a cross-validation scheme that fits on later data than it validates on; none of these is used.

The machine-learning models use 26 predictors, all computed from the daily closing price. They are the 5-day and 12-day simple moving averages, the 26-day exponential moving average, the 14-day Relative Strength Index with Wilder's smoothing (Wilder, 1978) , the MACD line (the 12-day minus the 26-day exponential moving average) and its 9-day exponential signal line, each of these six indicators lagged by one, two and three days, and the standard deviation of daily returns over the past 5 and 20 days. No volume information enters the predictor set. The longest look-back, the MACD signal line with three lags, removes the first 36 observations, so the first complete row is 22 February 2010. Predictors are standardised with means and standard deviations estimated on the data used for fitting.

Three quantities are forecast. The price target is the next day's closing price, forecast by ARIMA and by the random walk; in the main design no machine-learning model forecasts the price, and machine-learning forecasts of next-day returns are examined in a robustness analysis (R1, Section 3.5). The volume target is a class. Relative volume is defined as tick volume on day t+1 divided by the median tick volume of the 60 days ending on day t, so the normaliser uses data up to t only (at least 20 days are required, which binds only in early 2010). Relative volume is cut into Low, Medium and High classes at its tertiles in the training period, separately for each pair (0.893 and 1.099 for EUR/USD, 0.896 and 1.100 for GBP/USD, 0.880 and 1.116 for USD/CAD). By construction each class holds a third of the training days. Because activity declines over the sample, the test period has more Low days (36.9% to 40.4%) and fewer High days (28.7% to 32.8%), so a constant forecast of the most frequent test class would be right on 37% to 40% of days, not 33%. The risk target is realised volatility over the next five trading days: the sample standard deviation of the daily log returns of days t+1 to t+5, in percent. It is computed from daily returns rather than intraday data and is therefore a noisier measure of volatility than intraday realised volatility. A five-day window gives a target that can be measured from daily data with less noise than a single day's absolute return; R7 repeats the analysis at horizons of one and ten days.

The last five rows lack a complete five-day target and are dropped, which leaves 2,966 rows per pair, the same rows for every task. The sample is split once, in time order, into a training period of 2,076 days (22 February 2010 to 5 February 2018) and a test period of 890 days (6 February 2018 to 5 July 2021), about 70% and 30% of the rows, with the same dates for every pair. All estimation, tuning, scaling and class cut-offs use the training period only. Each model is fitted once, and its

parameters are held fixed over the test period; the ARIMA forecasts are updated with the observed data as each test day passes, but their parameters are not re-estimated. The analysis that re-fits the models each year is R4. The test period contains the COVID-19 shock of 2020. The 2016 referendum on the United Kingdom's membership of the European Union, which produced the largest daily move in the sample (8.3% for GBP/USD on 24 June 2016), falls in the training period.

3.3. Models and Benchmarks

Table 2 shows what each model forecasts, what it is compared with and what information it uses. The information sets differ by design. The Random Forest and the GBM see only the 26 price-derived predictors, while ARIMA and the naive rules use the history of the target series itself, including past volume. The main design therefore asks whether price-derived information is enough to match models built on the target's own history. Two further analyses give the volume classifiers the volume history they otherwise lack: a sensitivity check that adds today's relative volume and its three lags to the Random Forest, and R3 (Section 3.5).

Table 2. Targets, models, benchmarks and information sets in the main design.

Target (horizon)	Machine-learning model and its inputs	Time-series benchmark	Naive benchmark
Next-day closing price (1 day)	None in the main design (returns: R1)	ARIMA on the closing price, in pips	Random walk: today's close
Next-day volume class: Low, Medium, High (1 day)	Random Forest classifier; 26 price-derived predictors	ARIMA on log tick volume, forecast converted to a class	Today's class carried forward
Realised volatility over days t+1 to t+5 (5 days)	GBM; 26 price-derived predictors	ARIMA on trailing 5-day volatility, 5 steps ahead	Trailing 5-day volatility

The ARIMA benchmarks follow the Box–Jenkins approach (Box et al., 2015) and are estimated by exact maximum likelihood in statsmodels. The order of integration is chosen by the augmented Dickey–Fuller test (Dickey & Fuller, 1979) on the training data at the 5% level: a unit root is not rejected for price levels or for log tick volume, so these are modelled in first differences ($d = 1$, no constant), and it is rejected for trailing volatility, which is modelled in levels ($d = 0$, with a constant). The autoregressive and moving-average orders p and q each range from 0 to 5, and the order is chosen by the Akaike information criterion (Akaike, 1974) among the converged models of the 36 combinations. Models that did not converge are excluded even when their criterion value is lower; this occurred for the price models of EUR/USD and USD/CAD, the volume models of

all three pairs and the volatility models of GBP/USD and USD/CAD. The price models are estimated with prices expressed in pips (multiplied by 10,000) for numerical stability, and the criterion values were checked against the unscaled model. For EUR/USD this rule selects ARIMA(0,1,0), which is the random walk, so for that pair the price comparison has no content. The selected orders are given in Table 3. All six selected volume and volatility orders lie on the edge of the grid ($p = 5$ for the three volume models, $q = 5$ for the three volatility models); the three price orders do not. A wider grid was not searched; it would change only the benchmarks, not the machine-learning forecasts, and the point is listed among the limitations.

The volume ARIMA forecasts log tick volume for day t+1 one step ahead; the forecast is converted to relative volume with the same 60-day median as the target and assigned a class with the same training cut-offs. The volatility ARIMA is fitted to trailing five-day volatility and forecasts it five steps ahead. Trailing five-day volatility at day t+5 is, by construction, exactly the target dated t, so this forecast is aligned with the target; the alignment was checked numerically (maximum difference 5×10^{-9}).

The naive benchmarks carry the most recent value forward. For the price this is the random walk, which forecasts tomorrow's close as today's. For volume it is the rule that tomorrow's class equals today's, where today's class is today's tick volume relative to the 60-day median ending the day before, cut at the same training tertiles. For volatility it is the trailing five-day volatility at day t. These rules need no estimation. Comparing with them as well as with ARIMA shows whether a model improves on the latest observation at all; a model that beat only the naive rule would not have shown an advantage over a standard time-series model, and a model that beat only ARIMA might simply be exploiting a weakness of the selected ARIMA.

The Random Forest (Breiman, 2001) is scikit-learn's classifier (Pedregosa et al., 2011) ,with Gini impurity, the square root of the number of predictors considered at each split and bootstrap samples. Its tuning grid covers the number of trees (100, 300, 500), the maximum depth (5, 10, 20, 30) and the minimum number of observations needed to split a node (2, 10, 20), 36 combinations in all. The combination with the highest mean accuracy across the validation folds is refitted on the whole training period. Standardisation is applied inside the same pipeline; it does not affect tree splits but keeps the procedure identical for every model.

The Gradient Boosting Machine GBM;(Friedman, 2001)is scikit-learn's Gradient Boosting Regressor with squared-error loss, trees of depth 3 and no subsampling. Five learning rates are compared (0.005, 0.01, 0.02, 0.05 and 0.1). For each rate, a model with 2,000 trees is fitted in every fold and the validation error is recorded after each tree; the number of trees is set where the error averaged across folds is lowest, which is a form of early stopping. The pair of learning rate and number of trees with the lowest average validation error is then refitted on the whole training period. Because the target dated t uses returns up to t+5, the last four rows of every fitting block overlap the period that follows; they are dropped before fitting, both in the folds and in the final fit, so that no fitting target contains returns from the validation or test period.

Table 3. Specifications selected on the training period.

Model	EUR/USD	GBP/USD	USD/CAD
ARIMA, price (p, d, q)	(0, 1, 0)	(2, 1, 4)	(3, 1, 3)
ARIMA, log tick volume	(5, 1, 2)	(5, 1, 2)	(5, 1, 2)
ARIMA, trailing 5-day volatility	(1, 0, 5)	(5, 0, 5)	(3, 0, 5)
Random Forest: trees, maximum depth, minimum split	500, 10, 10	300, 10, 2	300, 5, 20
GBM: learning rate, number of trees	0.05, 16	0.005, 77	0.1, 3

Note: ARIMA orders chosen by AIC among converged models; machine-learning settings chosen by five-fold expanding-window cross-validation in the training period.

Two tools describe the fitted models without entering the forecast comparison. For the Random Forest, permutation importance measures the fall in test accuracy when each predictor is shuffled (30 repetitions). For the Gradient Boosting Machine, exact TreeSHAP values (Lundberg et al., 2020) attribute each forecast to the predictors.

3.4. Evaluation

Hyperparameters are tuned by five-fold cross-validation inside the training period, with expanding windows. Every fold fits on data from 22 February 2010 up to a cut-off date and validates on the 346 days that follow it; the fitting window grows from 346 to 1,730 days across the folds, and the last validation block ends on 5 February 2018. No observation from the test period is used for tuning, and the folds are never shuffled. The models are then evaluated once on the 890 test days.

Price forecasts are evaluated by the root mean squared error (RMSE), the mean absolute error and Theil's U, the ratio of a model's RMSE to that of the random walk, together with the share of days

on which the direction of the change is forecast correctly. Volume classes are evaluated by accuracy and the weighted F1 score, with the macro-averaged F1 score and the confusion matrices in Table S4. Volatility forecasts are evaluated by RMSE, with mean absolute errors in Table S4. For the machine-learning models, the difference between training and test performance and the spread of performance across the validation folds are reported as indicators of overfitting (Table 9, with fold-by-fold values in Table S4).

Differences in accuracy are tested on the same test days. For the volume classes, the exact McNemar test (McNemar, 1947) compares the number of days on which the model is right and the benchmark wrong with the number on which the reverse holds, using the binomial distribution. For the price and volatility forecasts, the Diebold–Mariano test (Diebold & Mariano, 1995) is applied to squared-error losses, with a Newey–West long-run variance (Newey & West, 1987) using $h - 1$ lags and the small-sample correction of (Harvey et al., 1997) compared with Student's t distribution with $n - 1$ degrees of freedom; $h = 1$ for the price and $h = 5$ for volatility. Because successive five-day windows overlap, volatility forecast errors are correlated from day to day, which this version of the test allows for. Plain paired t-tests on the same losses are reported only as a cross-check, since they ignore this correlation. All tests are two-sided, except the Clark–West test used as an additional check in R1, which is one-sided by construction.

Many tests are performed, so the conclusions are also judged after (Holm, 1979) correction. The family includes 114 tests: the 14 of the main design (for each pair, the Random Forest against ARIMA and against today's class, and the GBM against ARIMA and against trailing volatility; and the price ARIMA against the random walk for GBP/USD and USD/CAD, the EUR/USD forecasts being identical), and the formal tests of the robustness analyses R1 (28: ten Diebold–Mariano tests of the two regressors against the random walk and ARIMA, the two EUR/USD comparisons with ARIMA being left out because that ARIMA is the random walk, and 18 binomial sign tests of the three machine-learning models, against 50% and against the share of up days), R2 (18), R3 (24), R5 (12) and R7 (18); R4 and R6 report no p-values. All 114 tests, with their nominal and adjusted p-values, are listed in Table S1 of the supplementary material. Pooling all of them into one family is a choice. It is conservative, because many of the tests share the same test days and closely related models, and Holm's procedure guards against the worst case regardless of that dependence. Diagnostic tests, such as the absolute-error Diebold–Mariano tests and the additional

tests listed in Section 3.5, are reported but not adjusted (Table S2); the stress-window tests, the tests of the two sensitivity checks and the plain paired t-tests quoted in Section 4 are listed in Table S3.

Performance is also recomputed, without re-estimating the models, on two stress windows within the test period: the COVID-19 shock from 24 February to 29 May 2020 (70 days), and the days on which the 20-day standard deviation of returns is at or above its 90th percentile in the training period (20, 86 and 25 days for EUR/USD, GBP/USD and USD/CAD). The test period without the COVID-19 window is reported as a complement. These windows check whether the ranking of models changes when markets are under strain. Earlier events, such as the 2016 referendum, cannot be used for this because they fall in the training period.

3.5. Robustness Analyses

Seven further analyses, R1 to R7, were specified after the main results were known, to test whether those results depend on particular design choices, and a further analysis, R9, was added in response to review to widen the hyperparameter searches. The label R8 is used for the reproduction check described in Section 3.6. Unless stated otherwise, they keep the 26 predictors, the split, the test dates and the random seed: R5 starts its training period on 23 January 2013, and R7's ten-day horizon ends the test period on 28 June 2021. Any tuning they need uses the training period only. Table 4 summarises them: R5 changes the data; R1 and R7 change the target; R2 and R3 change the models or their information; R4 and R6 change only the way the forecasts are estimated and evaluated. Two sensitivity checks, run with the same code after the main results had been computed in order to interpret them, complement these analyses: the Random Forest with volume history (Section 3.3), and the GBM without the price-level predictors (the moving averages and their lags), which are not stationary.

Table 4. Robustness analyses.

Run	What changes	Models and benchmarks	Tests
R1 Returns	Target: next-day log return, in percent	Random Forest and GBM regressors; Random Forest up/not-up classifier; random walk (zero return), ARIMA (price forecasts converted to returns), training-period mean	Diebold–Mariano, h = 1; exact binomial tests of the sign hit rate against 50% and against the share of up days in the test period

Run	What changes	Models and benchmarks	Tests
R2 Risk benchmarks	Benchmarks: GARCH(1,1) and a daily HAR	GARCH(1,1) with constant mean and normal errors, estimated on training returns, parameters fixed and filtered forward, forecast = square root of the mean 1- to 5-step variance; HAR regression of the target on |r_t| and trailing 5- and 22-day volatility, by OLS, last four training rows dropped; compared with the GBM, ARIMA and trailing volatility	Diebold–Mariano, h = 5
R3 Volume models	Simpler, regularised classifiers; information set	Shallow Random Forest (300 trees, depth 2–4, minimum leaf 20–100) and L2-penalised multinomial logistic regression (C from 0.01 to 10), each with the 26 predictors or with the 26 plus today's relative volume and its three lags	Exact McNemar against ARIMA and today's class
R4 Yearly re-fit	Estimation: models re-fitted on 6 February 2018 and on 1 January 2019, 2020 and 2021 on all earlier data	All main-design and R1 models, and the Random Forest with volume history; tuned settings, ARIMA orders and class cut-offs frozen, parameters re-estimated; results by calendar year	None
R5 MetaTrader only	Data: MetaTrader segment only, from 4 December 2012	Features, targets and class cut-offs rebuilt; training period 23 January 2013 to 5 February 2018 (1,314 days), same 890 test days; Random Forest and GBM re-tuned, ARIMA re-estimated	Exact McNemar; Diebold–Mariano, h = 5
R6 Seeds and intervals	Evaluation: random seeds and sampling uncertainty	Random Forest and GBM re-fitted with seeds 0 to 9; moving-block bootstrap (Künsch, 1989) of the test days (2,000 resamples, percentile 95% intervals), blocks of 10 days for the accuracy difference with ARIMA and 20 days for the RMSE ratio to ARIMA	None (intervals)
R7 Horizon and classes	Targets: next-day absolute return (h = 1), 10-day volatility (h = 10, 885 test days), and two volume classes (High against not High)	GBM re-tuned with h − 1 rows dropped; ARIMA with the stored volatility orders on the matching trailing series; trailing baselines; Random Forest re-tuned for two classes	Diebold–Mariano at the matching horizon; exact McNemar
R9 Tuning breadth	Models: wider hyperparameter searches, added in response to review after all other results were known	GBM: depth 1 to 6, subsampling 0.5 or 1, three learning rates, up to 1,000 trees chosen on the validation curve; Random Forest: 48 settings of depth, minimum leaf size and features per split	Diebold–Mariano, h = 5; exact McNemar

R4 also re-fits the Random Forest with volume history. It builds the volume lags from the full raw series and so keeps the first three training rows that the sensitivity check drops; its single-split accuracy therefore differs slightly from that check (61.0% and 60.8% against 61.7% and 61.3%

for EUR/USD and GBP/USD; identical for USD/CAD); the paper reports R4 results for the main models only.

Some components were added during these analyses beyond their original plan and are labelled as additional throughout: the Clark–West test (Clark & West, 2007) , the Pesaran–Timmermann test (Pesaran & Timmermann, 1992) and absolute-error losses in R1; a GARCH forecast rescaled for the bias of a five-day sample standard deviation in R2; the Random Forest with volume history in R5; GARCH and HAR in R5 and R7; and alternative block lengths (5 and 20 days, and 10 and 40 days) in R6. They are supporting evidence, and no conclusion of the paper depends on them alone.

Two details affect how R7 and the re-estimated ARIMA models should be read. First, in the one-day volatility analysis of R7, early stopping chose a single tree for EUR/USD and GBP/USD and 18 trees for USD/CAD, so the GBM forecasts at that horizon are close to constant. Second, when an ARIMA model of fixed order is re-estimated in R4, R5 and R7 and the default fit does not converge, it is re-fitted with up to 2,000 iterations, and the second fit is kept only if it converges with a log-likelihood at least as high. The retry was needed in 11 of the 36 re-estimations in R4, in 4 of the 6 in R7 and for the volatility models of all three pairs in R5, and all of these converged. One model, the EUR/USD volume ARIMA in R5, did not converge in either fit; both fits reach the same log-likelihood, and its results are reported with this caveat. All reported results use this rule.

3.6. Software and Reproducibility

The analysis is written in Python 3.13 and uses NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, scikit-learn 1.9.1 (Pedregosa et al., 2011) for the machine-learning models, statsmodels 0.14.6 (Seabold & Perktold, 2010) for the ARIMA models and unit-root tests, and arch 7.2.0 (Sheppard, 2024) for the GARCH model and the block bootstrap. statsmodels and arch were built from their source code; arch carries a one-line change to an import that affects only a plotting function, which the analysis does not use. The random seed is 42 throughout, apart from the seed analysis in R6.

Three components were written for the study rather than taken from a library: exact TreeSHAP, the Diebold–Mariano test with the Harvey–Leybourne–Newbold correction, and the Holm adjustment. TreeSHAP values are computed by exact enumeration of the feature subsets used in

each tree, and they add up to each model's forecast to within 5×10^{-16}; the other two are short functions included in the published code.

The core pipeline, from the download of the source data to the main evaluation, was rerun twice in empty folders from a fresh download of the three source repositories at fixed versions, after every source file had been checked against its recorded hash. In both reruns 70 of the 79 output files were byte-identical to the originals, the other 9 differed only in recorded run times, and all 53 headline numbers were identical. The robustness analyses R1 to R7 have not been rerun in this way. They use the same code, and the yearly re-fit (R4) reproduces the stored main-design predictions to within 5×10^{-9}. The code, the download script and the table of hashes are archived at Zenodo (https://doi.org/10.5281/zenodo.23341357). Raw price data are not redistributed, because the terms of the original data providers may restrict it; the scripts rebuild every input from the public sources.

4. Results

All results refer to the 890 test days from 6 February 2018 to 5 July 2021 unless stated otherwise. Tests are two-sided, and p-values are nominal unless they are labelled as Holm-adjusted; adjusted values for every test in the Holm family are listed in Table S1. Volatility is measured in percentage points of daily return, and price errors in pips. All models were fitted once on the training period and their parameters held fixed over the test period, except in the yearly re-fits of R4; ARIMA orders were chosen among converged models (Section 3.3).

4.1. Price

Table 5 reports the one-day-ahead price forecasts. For EUR/USD the selected model is ARIMA(0,1,0), so its forecast is today's close and its errors are identical to those of the random walk. For GBP/USD the selected ARIMA(2,1,4) has a root mean squared error 1.0% higher than the random walk (Theil's U 1.010), and the Diebold–Mariano test rejects equal accuracy in the random walk's favour ($p = 0.0025$); the difference does not survive the Holm correction (adjusted $p = 0.154$). For USD/CAD the ARIMA(3,1,3) is 0.5% less accurate than the random walk, which is not significant ($p = 0.23$). The ARIMA models therefore add nothing to the random walk: in one pair they are the random walk, and in the other two they are slightly worse. The direction of the

next day's change is forecast correctly on 45.6% of days for GBP/USD and 53.1% for USD/CAD; for EUR/USD no direction is forecast, because the forecast change is zero every day.

Table 5. One-day-ahead price forecasts, test period.

Pair	ARIMA order	RMSE, ARIMA (pips)	RMSE, random walk (pips)	Theil's U	MAE, ARIMA / random walk (pips)	Direction correct (%)	Diebold–Mariano p (Holm)
EUR/USD	(0, 1, 0)	46.7	46.7	1.000	36.1 / 36.1	–	– (identical forecasts)
GBP/USD	(2, 1, 4)	71.4	70.7	1.010	53.6 / 53.0	45.6	0.0025 (0.154)
USD/CAD	(3, 1, 3)	56.8	56.5	1.005	41.6 / 41.5	53.1	0.23 (1.00)

Note: Theil's U is the ratio of the ARIMA RMSE to the random-walk RMSE; values above 1 favour the random walk. The EUR/USD model forecasts no change on every day, so it has no directional forecast.

4.2. Trading Activity

Table 6 reports the next-day volume class. The Random Forest classifies 41.6%, 31.8% and 34.2% of test days correctly for EUR/USD, GBP/USD and USD/CAD. ARIMA reaches 61.5% to 63.8% and today's class 61.7% to 64.6%, so the Random Forest is 20 to 31 points less accurate than either benchmark. The exact McNemar tests reject equal accuracy in every pair, in favour of ARIMA ($p \leq 1.6\times10^{-17}$) and of today's class ($p \leq 1.0\times10^{-18}$), and all six differences survive the Holm correction. Low is the most frequent class in the test period (40.4%, 36.9% and 38.5% of days), so a constant forecast of Low would have scored that share; the Random Forest is below it for GBP/USD and USD/CAD and above it, by 1.2 points, for EUR/USD. The training classes are equal thirds by construction, so a training-majority rule is not defined in the main design; in the MetaTrader-only run R5, where the training majority is Low, the Random Forest is again below it for GBP/USD and USD/CAD. For USD/CAD it assigns 597 of the 890 test days (67%) to the Medium class, which holds 31% of them.

Table 6. Next-day volume class (Low, Medium, High), test period.

Pair	Random Forest accuracy (%)	Random Forest weighted F1	ARIMA accuracy (%)	ARIMA weighted F1	Today's class accuracy (%)	Low days in test (%)	McNemar p: RF vs ARIMA	McNemar p: RF vs today's class
EUR/USD	41.6	0.419	61.5	0.624	61.7	40.4	1.6×10^{-17}	1.0×10^{-18}
GBP/USD	31.8	0.318	61.7	0.626	62.7	36.9	1.3×10^{-31}	5.3×10^{-36}
USD/CAD	34.2	0.303	63.8	0.648	64.6	38.5	4.4×10^{-34}	2.8×10^{-33}

Note: Low is the most frequent class in the test period, so a constant forecast of Low would score the share in the seventh column. All McNemar differences favour the benchmark and survive the Holm correction.

ARIMA and today's class are close to each other: their accuracies differ by at most 1.0 point, and their forecasts agree on 78.5% to 80.6% of test days. Trading activity in these pairs is persistent enough that a model of the volume series adds little to its latest value. The Random Forest does not use that series at all. Given the same volume history, the classifiers reach 60.8% to 65.2% and are not significantly different from either benchmark (Section 4.6, Table 10), so most of the gap in Table 6 reflects the information set rather than the algorithm.

4.3. Volatility

Table 7 reports the forecasts of realised volatility over the next five days. The GBM is less accurate than ARIMA for EUR/USD (RMSE 0.194 against 0.158, ratio 1.23, $p = 1.4\times10^{-8}$) and USD/CAD (0.190 against 0.165, ratio 1.15, $p = 4.3\times10^{-7}$), and the two are not significantly different for GBP/USD (0.232 against 0.226, p = 0.19). Against trailing volatility, the GBM is more accurate for GBP/USD (0.232 against 0.284, p = 0.00033), a difference that survives the Holm correction (adjusted p = 0.022); for EUR/USD and USD/CAD the differences from trailing volatility are not significant (p = 0.65 and 0.23). The GBP/USD result is therefore a win over the naive benchmark, not over ARIMA. A constant forecast equal to the training-period mean shows what the win is worth. For GBP/USD the GBM is not significantly different from that constant (RMSE 0.232 against 0.234, p = 0.62), and the constant itself beats trailing volatility (p = 0.001), so the GBP/USD result shows that trailing volatility is a noisy benchmark, not that the GBM extracts structure. For EUR/USD and USD/CAD the GBM does beat the constant ($p < 0.001$), but not ARIMA. The GBM forecast is nearly flat: it takes 27, 30 and 12 distinct values over the 890 test days (Figure 1, Table S3).

Table 7. Five-day realised volatility forecasts, test period (RMSE, percentage points).

Pair	GBM	ARIMA	Trailing volatility	Training-period mean	GBM / ARIMA	Diebold–Mariano p: GBM vs ARIMA	Diebold–Mariano p: GBM vs trailing (Holm)	GBM test R^2	GBM learning rate, trees
EUR/USD	0.194	0.158	0.190	0.229	1.23	1.4×10^{-8} (ARIMA better)	0.65	−0.41	0.05, 16
GBP/USD	0.232	0.226	0.284	0.234	1.03	0.19	0.00033 (0.022), GBM better	0.01	0.005, 77
USD/CAD	0.190	0.165	0.204	0.203	1.15	4.3×10^{-7} (ARIMA better)	0.23	−0.07	0.1, 3

Note: Diebold–Mariano tests with h = 5 and the Harvey–Leybourne–Newbold correction. R^2 is computed against the test-period mean of the target.

The GBM forecasts are too high on average. Their mean over the test period is 0.484, 0.512 and 0.462 for EUR/USD, GBP/USD and USD/CAD, against realised means of 0.368, 0.480 and 0.386; the ARIMA means are 0.398, 0.487 and 0.407. Volatility was lower in the test period than in the training period for EUR/USD (mean target 0.37 against 0.53) and USD/CAD (0.39 against 0.47), and the GBM forecasts stay close to the training level, moving little from day to day (Figure 1); for GBP/USD the mean target is almost unchanged (0.49 and 0.48). Its test R^2 is negative for EUR/USD and USD/CAD, so for those pairs it is less accurate than the test-period mean would have been, although it improves on the training-period mean in all three pairs.

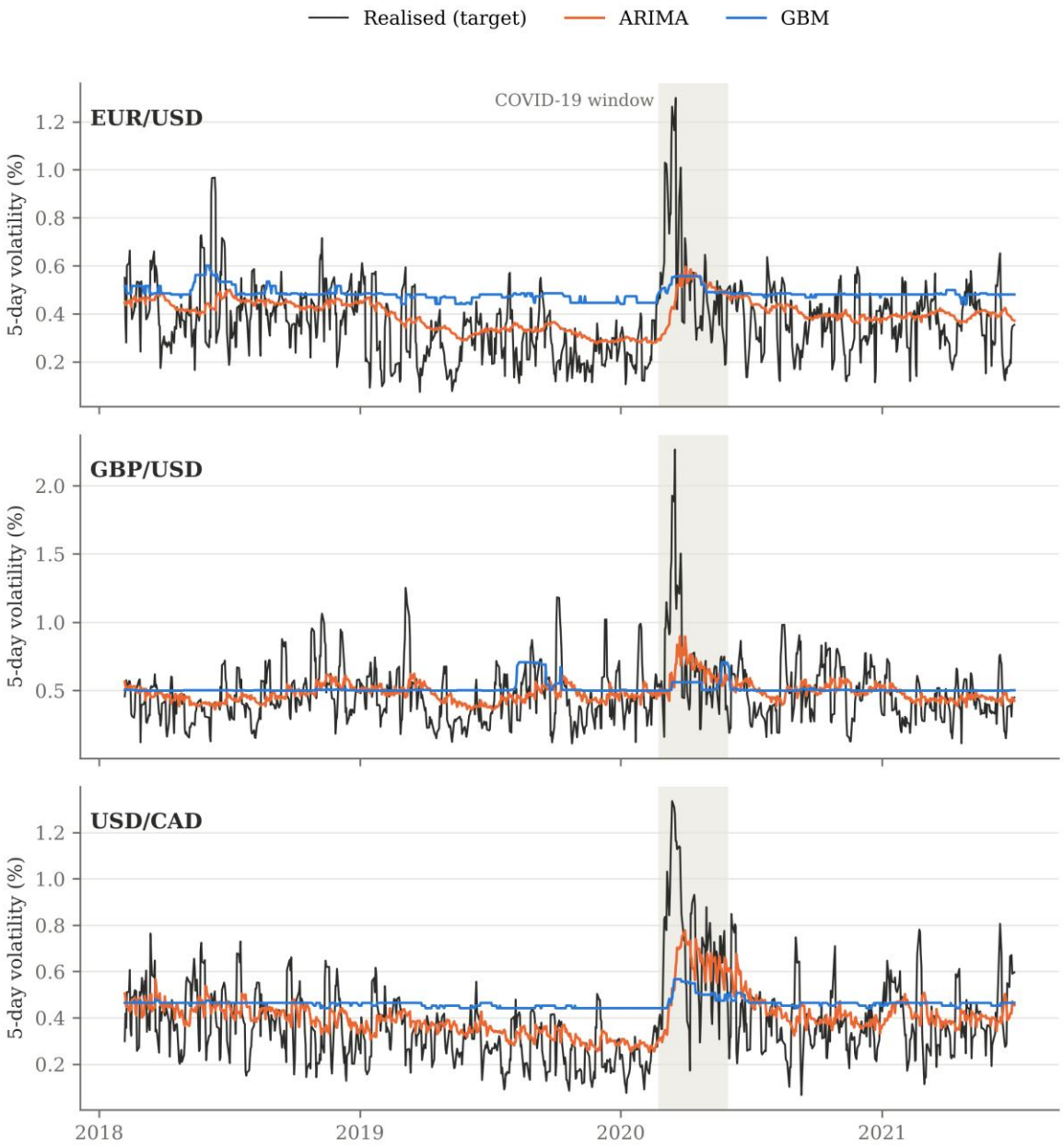


Figure 1. Five-day realised volatility and its ARIMA and GBM forecasts over the test period. The shaded band is the COVID-19 window (24 February to 29 May 2020).

4.4. Stress Windows

Table 8 repeats the main comparisons in the COVID-19 window and on high-volatility days. In the COVID-19 window price errors rise for every pair: the random-walk RMSE increases by 58%, 70% and 92% for EUR/USD, GBP/USD and USD/CAD. Volatility errors also rise for every model, by 39% to 105%. Volume accuracy behaves differently. ARIMA's accuracy rises to 88.6%, 77.1% and 91.4% and today's class to 85.7%, 74.3% and 95.7%, because classes persisted longer

than usual: runs of the same class lasted 6.4, 3.7 and 17.5 days on average in the window, against 2.6 to 2.8 days over the whole test period, so today's class was usually right. The Random Forest stays at 41.4%, 28.6% and 35.7%, close to its full-period level, and remains significantly below ARIMA in all three pairs. Across the six stress comparisons (three pairs, two windows) the Random Forest is below ARIMA every time, significantly in five; the exception is USD/CAD on the 25 high-volatility days ($p = 0.57$).

Table 8. Results in the stress windows.

Pair and window	Days	Price RMSE, random walk (pips)	RF accuracy (%)	ARIMA accuracy (%)	Today's class accuracy (%)	GBM RMSE	ARIMA RMSE	Trailing RMSE	GBM / ARIMA (p)
EUR/USD, full test	890	46.7	41.6	61.5	61.7	0.194	0.158	0.190	1.23 (1.4×10^{-8})
EUR/USD, COVID-19	70	73.8	41.4	88.6	85.7	0.270	0.324	0.272	0.83 (0.041)
EUR/USD, high volatility	20	92.1	60.0	90.0	90.0	0.194	0.197	0.334	0.98 (0.61)
GBP/USD, full test	890	70.7	31.8	61.7	62.7	0.232	0.226	0.284	1.03 (0.19)
GBP/USD, COVID-19	70	120.3	28.6	77.1	74.3	0.460	0.433	0.422	1.06 (0.24)
GBP/USD, high volatility	86	117.9	41.9	66.3	65.1	0.394	0.375	0.420	1.05 (0.41)
USD/CAD, full test	890	56.5	34.2	63.8	64.6	0.190	0.165	0.204	1.15 (4.3×10^{-7})
USD/CAD, COVID-19	70	108.6	35.7	91.4	95.7	0.341	0.332	0.318	1.03 (0.62)
USD/CAD, high volatility	25	129.7	88.0	92.0	92.0	0.345	0.312	0.380	1.11 (0.50)

Note: COVID-19 window = 24 February to 29 May 2020. High volatility = test days on which the 20-day standard deviation of returns is at or above its training-period 90th percentile. Models are not re-estimated for the windows. The last column gives the RMSE ratio and the Diebold–Mariano p-value (h = 5), unadjusted.

In one comparison the GBM beats ARIMA. For EUR/USD in the COVID-19 window, ARIMA's volatility error doubles (RMSE 0.324, up 105%) while the GBM's rises by 39% (0.270), giving a ratio of 0.83 ($p = 0.041$). This is one of six GBM-against-ARIMA comparisons in the stress windows, it is not adjusted for multiple testing, and it does not recur on high-volatility days for EUR/USD ($p = 0.61$) or in either window for the other pairs ($p \geq 0.24$). Without the COVID-19 window, over the remaining 820 test days, the GBM/ARIMA ratio is 1.385, 1.013 and 1.207 ($p = 2.4\times10^{-19}$, 0.52 and 1.8×10^{-8}), and the Random Forest remains 17.6, 28.3 and 27.4 points below

ARIMA. The COVID-19 result is reported as exploratory. Against trailing volatility, the GBM is not significantly different in the COVID-19 window or on high-volatility days in any pair ($p \geq 0.21$); without the COVID-19 window it is more accurate for GBP/USD ($p = 3.4\times10^{-8}$) and USD/CAD ($p = 0.042$), both unadjusted (Table S3). The high-volatility samples hold only 20, 86 and 25 days, too few to rank the models: on the 25 USD/CAD days, for example, the Random Forest scores 88.0%, a figure that rests on three misclassified days and has no counterpart in the other windows.

4.5. Overfitting and What the Models Use

The Random Forest fits the training data far better than the test data (Table 9). Its training accuracy is 82.6%, 88.6% and 61.6%, and its test accuracy is lower by 41.0, 56.8 and 27.5 points for EUR/USD, GBP/USD and USD/CAD. Its accuracy also varies across the five validation folds, by 12.4, 15.3 and 23.4 points from the best fold to the worst, so the tuning step itself was working with an unstable signal. The selected maximum depths are 10, 10 and 5, from a grid of 5 to 30 (Figure 2 shows training, fold and test accuracy), and the shallow forests in R3, with depths of two to four, narrow the train–test gap to 8.3, 21.1 and 13.9 points but leave test accuracy at 31.1% to 37.5%. Restricting the trees reduces the overfitting without improving the forecasts.

Table 9. Training and test performance of the machine-learning models.

Pair	RF training accuracy (%)	RF test accuracy (%)	RF gap (points)	RF validation-fold range (points)	GBM training RMSE	GBM test RMSE	GBM training R^2	GBM test R^2	GBM validation-fold range (RMSE)	Mean target, training / test
EUR/USD	82.6	41.6	41.0	12.4	0.219	0.194	0.28	−0.41	0.108	0.53 / 0.37
GBP/USD	88.6	31.8	56.8	15.3	0.258	0.232	0.19	0.01	0.314	0.49 / 0.48
USD/CAD	61.6	34.2	27.5	23.4	0.224	0.190	0.15	−0.07	0.083	0.47 / 0.39

Note: the GBM training figures exclude the four purged rows. RMSE and the mean target are in percentage points.

The GBM shows no gap of this kind in RMSE: its training RMSE is higher than its test RMSE in all three pairs. For EUR/USD and USD/CAD the test period was calmer, with a target standard deviation of 0.16 and 0.18 against 0.26 and 0.24 in training. For GBP/USD the explanation is the 2016 referendum: excluding June and July 2016 lowers the training RMSE from 0.258 to 0.203, and the largest 1% of training errors account for 42% of the training sum of squared errors. Measured by R^2, the GBM does lose accuracy out of sample (from 0.28, 0.19 and 0.15 in training to −0.41, 0.01 and −0.07 in the test period). Its validation-fold RMSE ranges over 0.108, 0.314 and 0.083 (Table 9); for GBP/USD, the fold from June 2015 to October 2016, which contains the

referendum, has an RMSE of 0.504 against 0.190 to 0.266 for the other folds and contributes 56% of the cross-validated squared error that selected the model. Early stopping chose 16, 77 and 3 trees, and the best number of trees varies widely from fold to fold (Table S4). The GBM's weakness is therefore less overfitting than a weak, unstable signal: few trees, low training R^2 and forecasts that carry the training-period level of volatility into the test period (Section 4.3).

The two importance measures for the Random Forest disagree. By impurity-based importance, which is computed on the training data, the 20-day return volatility is the most important predictor for EUR/USD and GBP/USD. By permutation importance on the test set, it reduces accuracy when it is used: shuffling it raises test accuracy slightly, ranking it 22nd of 26 for EUR/USD and 15th for GBP/USD. Of the 26 predictors, 16, 17 and 7 have zero or negative permutation importance (Figure 3).

The GBM's SHAP values show which inputs drive its forecasts. For EUR/USD and USD/CAD the two return-volatility predictors account for 59% and 71% of the mean absolute SHAP value on the test set, so the model is largely a transformation of recent volatility. For GBP/USD the moving averages of the price level account for 78%, led by the 5-day moving average; MACD and its signal line account for 18%, 4% and 5% across the three pairs, and RSI for 2% to 6%. Price-level moving averages are not stationary, and Section 4.6 shows what happens when they are removed.

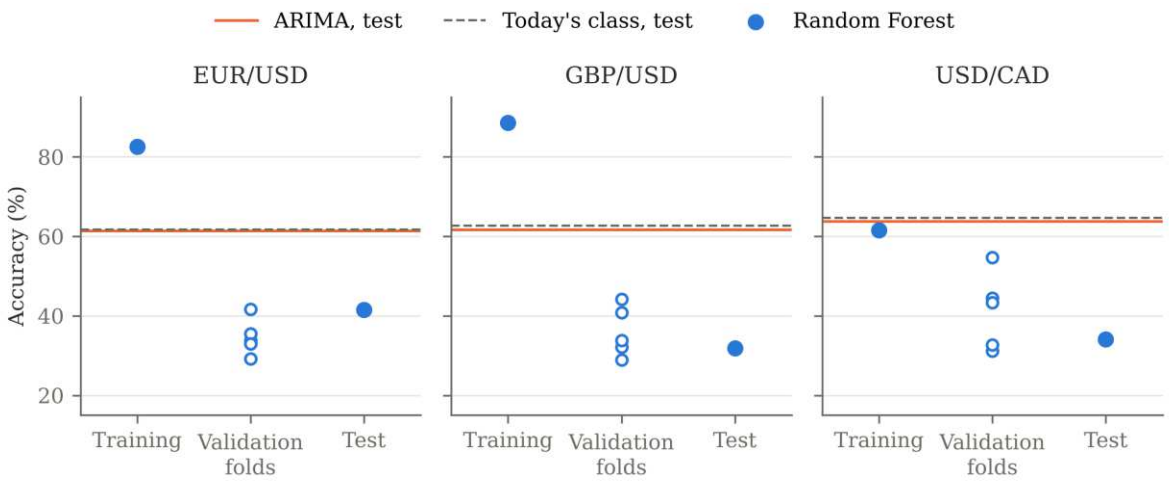


Figure 2. Random Forest accuracy on the training data, in the five validation folds and on the test period, against the test accuracy of ARIMA and today's class.

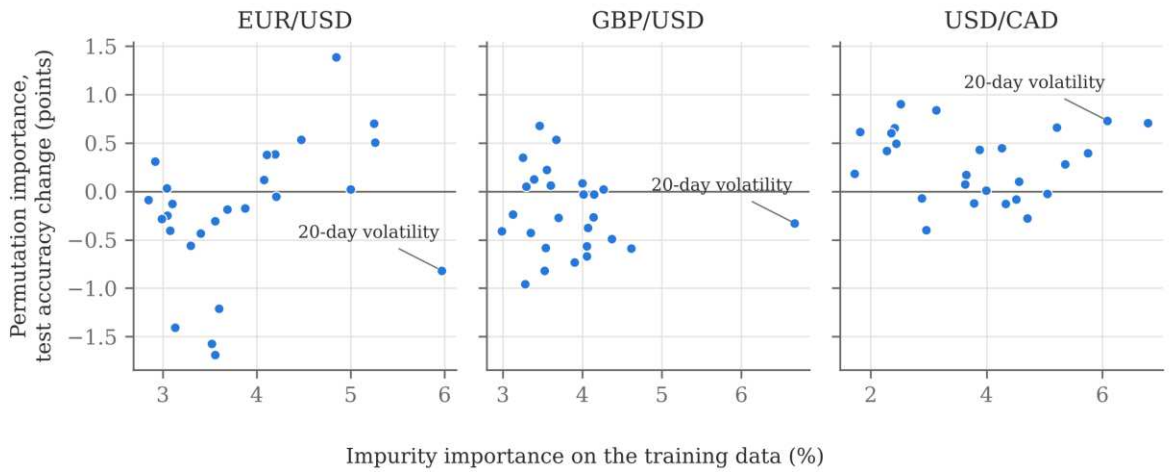


Figure 3. Random Forest predictor importance: impurity-based importance on the training data against permutation importance on the test period. Points below zero are predictors whose shuffling raises test accuracy.

4.6. Information Sets and Sensitivity

The Random Forest's shortfall on volume comes mainly from what it sees. Table 10 compares the main model with versions that are also given today's relative volume and its three lags. The full Random Forest with volume history reaches 61.7%, 61.3% and 64.7%; the shallow Random Forest and the logistic regression of R3 reach 60.8% to 65.2% with volume history, against 31.1% to 40.8% without it. In the main sample, no model with volume history differs significantly from ARIMA or from today's class: the exact McNemar p-values against ARIMA are 0.35 to 1.00 for the R3 models and 0.55 to 0.92 for the full Random Forest, and against today's class 0.20 to 1.00 and 0.39 to 1.00 (Table S3). The one exception is in the MetaTrader-only run R5, where the Random Forest with volume history is significantly less accurate than today's class for GBP/USD (61.0% against 64.3%, McNemar $p = 0.042$; Table S2). The volume comparison in Section 4.2 is therefore mainly an information-set result: once the classifier sees the volume series, it is not significantly different from the time-series benchmark, and it does not beat it.

Table 10. Next-day volume class with and without volume history (test accuracy, %).

Pair	RF, main (26 predictors)	RF + volume history	Shallow RF, 26	Shallow RF + volume history	Logistic, 26	Logistic + volume history	ARIMA	Today's class
EUR/USD	41.6	61.7	37.1	60.9	40.8	61.6	61.5	61.7
GBP/USD	31.8	61.3	31.1	60.8	32.0	61.5	61.7	62.7
USD/CAD	34.2	64.7	37.5	65.2	38.5	63.1	63.8	64.6

Note: "RF + volume history" is the sensitivity check run with the main analysis; the shallow Random Forest and the L2-penalised logistic regression are from R3. Volume history = today's relative volume and its lags of one to three days.

Defining the classes by raw tick volume instead of relative volume makes the task unstable. Because tick volume drifts over time (Section 3.1), the training cut-offs place 0.1% to 8.2% of test days in the Low class; for GBP/USD only one test day is Low. The Random Forest's test accuracy on these classes is 25.4%, 47.2% and 24.2%, and the GBP/USD figure comes from predicting High on 96% of test days, when 45% of them are High, rather than from skill. This variant is not used for any conclusion.

Removing the price-level moving averages from the GBM, which leaves RSI, MACD, their lags and the two volatility predictors, lowers its RMSE for EUR/USD and USD/CAD (0.169 and 0.174, against 0.194 and 0.190) and leaves GBP/USD unchanged (0.233). The reduced model is still less accurate than ARIMA for EUR/USD ($p = 0.033$) and USD/CAD ($p = 0.006$) and not significantly different for GBP/USD ($p = 0.22$). It is more accurate than trailing volatility in all three pairs ($p = 0.028$, 0.0007 and 0.0027). These comparisons lie outside the Holm family and are unadjusted, and this check was specified after the main results had been computed. The reduced model therefore beats trailing volatility in all three pairs, where the main GBM did so only for GBP/USD, but it still does not beat ARIMA.

4.7. Robustness Analyses

Table 11 summarises the main comparison from each robustness analysis. None reverses the main results.

Table 11. Main results of the robustness analyses.

Run	Comparison	EUR/USD	GBP/USD	USD/CAD
R1	RF regressor / random walk, return RMSE (p)	0.999 (0.45)	1.068 (0.004)	0.994 (0.56)
R1	GBM regressor / random walk, return RMSE (p)	1.000 (0.60)	0.999 (0.12)	1.000 (0.66)
R1	GBM regressor / ARIMA, return RMSE (p)	same as random walk	0.989 (0.0017)	0.995 (0.23)
R1	RF classifier, share of signs correct (%)	48.7	51.1	49.8
R2	GARCH / ARIMA, RMSE (p)	1.029 (0.21)	1.087 (0.14)	1.003 (0.89)
R2	HAR / ARIMA, RMSE (p)	1.016 (0.63)	0.989 (0.41)	0.990 (0.56)
R2	HAR / GBM, RMSE (p)	0.826 (6.6×10^{-13})	0.963 (0.19)	0.859 (6.0×10^{-6})
R4	Years in which RF is below ARIMA / GBM is worse than ARIMA	4 of 4 / 3 of 4	4 of 4 / 3 of 4	4 of 4 / 3 of 4
R5	RF accuracy vs ARIMA accuracy (%)	43.8 vs 60.7	35.3 vs 62.6	36.9 vs 64.0
R5	GBM / ARIMA, RMSE (p)	1.262 (2.2×10^{-11})	1.039 (0.080)	1.109 (0.0004)

Run	Comparison	EUR/USD	GBP/USD	USD/CAD
R6	RF minus ARIMA accuracy, 95% interval (points)	−26.6 to −13.3	−37.9 to −22.1	−37.9 to −22.1
R6	GBM / ARIMA RMSE, 95% interval	1.07 to 1.44	0.98 to 1.07	1.07 to 1.28
R7	GBM / ARIMA, next-day |return| (p)	1.124 (2.7×10^{-20})	1.020 (0.21)	1.056 (4.9×10^{-7})
R7	GBM / trailing $\|r_t\|$, next-day |return| (p)	0.817 (1.3×10^{-7})	0.772 (1.7×10^{-6})	0.762 (7.7×10^{-11})
R7	GBM / ARIMA, 10-day volatility (p)	1.103 (0.002)	1.055 (0.27)	1.046 (0.15)
R7	Two classes: RF vs ARIMA vs always "not High" (%)	71.6 vs 83.7 vs 71.3	56.4 vs 79.4 vs 67.2	69.4 vs 81.6 vs 69.4
R9	GBM, wider search / ARIMA, RMSE (p)	1.120 (0.0008)	1.024 (0.55)	1.050 (0.035)
R9	Random Forest, wider search, accuracy vs ARIMA (%)	39.7 vs 61.5	30.8 vs 61.7	36.5 vs 63.8

Note: ratios below 1 favour the model named first. p-values are from the Diebold–Mariano test (h matching the target), unadjusted; Holm-adjusted values are in Table S1. R6 intervals are moving-block bootstrap percentile intervals; each pair is resampled with the same seed, so the pairs share the resampled day blocks, and interval endpoints fall on whole numbers of days out of 890.

Returns (R1). Forecasting next-day returns gives no gain over the random walk. The out-of-sample R^2 of the machine-learning regressors relative to the random walk ranges from −14.1% to 1.2%, and no regressor is significantly more accurate than the random walk. The Random Forest regressor is significantly less accurate for GBP/USD ($p = 0.004$). The GBM regressor beats ARIMA for GBP/USD ($p = 0.0017$), but only because that ARIMA is itself less accurate than the random walk; the result does not survive the Holm correction (adjusted $p = 0.111$). The sign of the next day's return is forecast correctly on 47.9% to 52.2% of days, and no hit rate differs significantly from 50% or from the share of up days. For EUR/USD and USD/CAD the GBM regressor forecasts the same sign on every day. The one-sided Clark–West test, reported as an additional check, gives $p = 0.010$ and 0.047 for the GBP/USD Random Forest and GBM regressors against the random walk. It adjusts for the noise that estimating a nested model adds, so it can favour a model whose squared error is higher, as here for the Random Forest; these p-values are unadjusted and do not change the conclusion.

Volatility benchmarks (R2). GARCH(1,1) and the daily HAR are not significantly different from ARIMA in any pair ($p = 0.14$ to 0.89), more accurate than the GBM for EUR/USD and USD/CAD ($p \leq 4.6\times10^{-5}$, surviving the Holm correction), not significantly different from it for GBP/USD, and more accurate than trailing volatility in every pair ($p \leq 0.003$; five of the six comparisons survive the Holm correction, the exception being GARCH for EUR/USD, adjusted $p = 0.163$). The standard volatility models therefore confirm the ARIMA result rather than change it.

Yearly re-fits (R4). Re-estimating the models at the start of each year leaves the pattern intact. The Random Forest is below ARIMA in all twelve pair-years. The GBM has a higher RMSE than ARIMA in nine of the twelve pair-years and a lower one in three (EUR/USD in 2020, GBP/USD in 2019 and USD/CAD in 2021); R4 reports these comparisons descriptively, without tests. Over the whole test period the re-fitted Random Forest scores 40.9%, 33.7% and 37.5%, against 61.7%, 61.8% and 63.6% for the re-fitted ARIMA.

MetaTrader data only (R5). Using the single later source, with a training period that starts in January 2013, the Random Forest scores 43.8%, 35.3% and 36.9%, again far below ARIMA (McNemar $p \leq 7.4\times10^{-14}$). The GBM is less accurate than ARIMA for EUR/USD and USD/CAD and not significantly different from it for GBP/USD ($p = 0.080$); against trailing volatility it is more accurate for GBP/USD ($p = 0.0014$), which does not survive the Holm correction (adjusted $p = 0.090$). The EUR/USD volume ARIMA in this run did not converge (Section 3.5); its accuracy, 60.7%, is close to that of today's class (61.5%).

Seeds and sampling uncertainty (R6). The headline results use seed 42. Across seeds 0 to 9, the Random Forest's test accuracy averages 42.6%, 32.3% and 33.8% (ranges 41.7% to 43.3%, 31.0% to 33.4% and 32.7% to 34.6%), with standard deviations of 0.50, 0.64 and 0.63 points; the seed-42 figure for EUR/USD, 41.6%, lies just below this range. The GBM's forecasts do not change with the seed. The block-bootstrap intervals for the Random Forest's accuracy deficit against ARIMA exclude zero in every pair, and no bootstrap draw favours the Random Forest. The interval for the GBM's RMSE ratio to ARIMA lies above 1 for EUR/USD and USD/CAD and includes 1 for GBP/USD, where 12.9% of draws favour the GBM.

Horizons and class definitions (R7). At one day ahead, the GBM is less accurate than ARIMA for EUR/USD and USD/CAD and not significantly different from it for GBP/USD, but more accurate than the trailing baseline in all three pairs, with RMSE ratios of 0.76 to 0.82 and p-values that survive the Holm correction. The trailing baseline at this horizon is today's absolute return, a noisy forecast of tomorrow's, and early stopping retained one tree for EUR/USD and GBP/USD and 18 for USD/CAD, so the GBM forecast is close to a constant. ARIMA has the lower RMSE in every pair at this horizon. At ten days, the GBM is less accurate than ARIMA for EUR/USD ($p = 0.002$) and not significantly different from it for the other pairs, and for EUR/USD it is also less accurate than trailing volatility ($p = 0.040$). With two volume classes, High against not High, the Random

Forest scores 71.6%, 56.4% and 69.4%, against 83.7%, 79.4% and 81.6% for ARIMA (McNemar $p \leq 6.4\times10^{-13}$); it is within 0.3 points of the constant forecast of "not High" for EUR/USD and USD/CAD and falls below it for GBP/USD.

Tuning breadth (R9). Widening the hyperparameter searches did not change the comparison. Searching GBM depths of 1 to 6, with and without subsampling, selected trees of depth 1 in all three pairs, with 18, 85 and 35 trees, so the validation data favoured simpler models, not more flexible ones. The resulting RMSE (0.177, 0.232 and 0.173) is lower than the main GBM for EUR/USD and USD/CAD, but the model remains less accurate than ARIMA for EUR/USD ($p = 0.0008$) and USD/CAD ($p = 0.035$) and not significantly different for GBP/USD ($p = 0.55$); against trailing volatility it is more accurate for GBP/USD and USD/CAD ($p = 4.6\times10^{-6}$ and 0.0016) and not for EUR/USD ($p = 0.16$). The wider Random Forest search selected forests of depth 3 and scored 39.7%, 30.8% and 36.5%, again far below ARIMA and today's class (McNemar $p \leq 3.1\times10^{-22}$). These tests are outside the Holm family and are listed in Table S5.

4.8. Multiple Testing

Of the 114 tests in the Holm family, 57 have a nominal p-value below 0.05, and 6 of these favour a machine-learning model. After the Holm correction 49 remain significant, and 4 favour a machine-learning model. Table 12 lists every tested comparison in Tables S1 to S3 and S5 in which a machine-learning model is significantly better than its benchmark at the nominal 5% level by the test used for that comparison. The four that survive the correction are all against naive baselines: the GBM against trailing volatility for GBP/USD in the main design, and the one-day GBM against today's absolute return in all three pairs. None of the 114 tests shows a machine-learning model beating ARIMA, GARCH or HAR over the full test period after correction.

Table 12. Machine-learning wins at the nominal 5% level among the tested comparisons in Tables S1 to S3 and S5.

Source	Pair	Model and target	Benchmark	p (nominal)	Holm p	Status
R7	USD/CAD	GBM, next-day |return|	Today's |return|	7.7×10^{-11}	6.3×10^{-9}	Survives Holm
R7	EUR/USD	GBM, next-day |return|	Today's |return|	1.3×10^{-7}	9.9×10^{-6}	Survives Holm
R7	GBP/USD	GBM, next-day |return|	Today's |return|	1.7×10^{-6}	1.3×10^{-4}	Survives Holm
Main	GBP/USD	GBM, 5-day volatility	Trailing volatility	0.00033	0.022	Survives Holm
R5	GBP/USD	GBM, 5-day volatility (MetaTrader data)	Trailing volatility	0.0014	0.090	Does not survive

Source	Pair	Model and target	Benchmark	p (nominal)	Holm p	Status
R1	GBP/USD	GBM regressor, next-day return	ARIMA (worse than random walk)	0.0017	0.111	Does not survive
Sensitivity	GBP/USD	GBM without price levels, 5-day volatility	Trailing volatility	0.0007	–	Outside family, unadjusted
Sensitivity	USD/CAD	GBM without price levels, 5-day volatility	Trailing volatility	0.0027	–	Outside family, unadjusted
Sensitivity	EUR/USD	GBM without price levels, 5-day volatility	Trailing volatility	0.028	–	Outside family, unadjusted
Stress window	EUR/USD	GBM, 5-day volatility, COVID-19 window	ARIMA	0.041	–	Outside family, unadjusted
Stress window	GBP/USD	GBM, 5-day volatility, test period without COVID-19	Trailing volatility	3.4×10^{-8}	–	Outside family, unadjusted; overlaps the full-period test
Stress window	USD/CAD	GBM, 5-day volatility, test period without COVID-19	Trailing volatility	0.042	–	Outside family, unadjusted; overlaps the full-period test
Constant benchmark	EUR/USD	GBM, 5-day volatility	Training-period mean	4.2×10^{-39}	–	Outside family, unadjusted
Constant benchmark	USD/CAD	GBM, 5-day volatility	Training-period mean	2.2×10^{-7}	–	Outside family, unadjusted
R1 (additional)	GBP/USD	GBM regressor, next-day return, absolute error	ARIMA	0.0042	–	Diagnostic, not adjusted
R1 (additional)	GBP/USD	RF regressor, next-day return, Clark–West (one-sided)	Random walk	0.010	–	Diagnostic, not adjusted
R1 (additional)	GBP/USD	GBM regressor, next-day return, Clark–West (one-sided)	Random walk	0.047	–	Diagnostic, not adjusted
R9	GBP/USD	GBM, wider search, 5-day volatility	Trailing volatility	4.6×10^{-6}	–	Outside family, unadjusted
R9	USD/CAD	GBM, wider search, 5-day volatility	Trailing volatility	0.0016	–	Outside family, unadjusted

Note: the Holm family and its 114 tests are defined in Section 3.4 and listed in Table S1; the diagnostic tests are in Table S2, and the stress-window, sensitivity and paired t-tests in Table S3. Plain paired t-tests ignore the overlap of multi-day windows and are not counted as wins; two of them favour the GBM against trailing volatility where the Diebold–Mariano test does not (USD/CAD in R5, p = 0.0022 against DM p = 0.092; GBP/USD at ten days in R7, p = 0.011 against DM p = 0.22). The two comparisons without the COVID-19 window are subsets of the full-period tests and are not independent evidence. The GBM's comparisons with trailing volatility within the stress windows were added for this revision (Table S3)

Outside the family, ten further comparisons favour a machine-learning model without adjustment: the GBM without price-level predictors against trailing volatility (three pairs), the R9 GBM against trailing volatility (GBP/USD and USD/CAD), the GBM against trailing volatility without the COVID-19 window (GBP/USD and USD/CAD, subsets of the full-period tests), the GBM against

a constant training-mean forecast (EUR/USD and USD/CAD), and the GBM against ARIMA for EUR/USD in the COVID-19 window. Only the last is against a time-series benchmark, and none is against ARIMA, GARCH or HAR over the full test period. Every machine-learning win against a time-series benchmark is either nominal or unadjusted. The GBP/USD return forecasts beat an ARIMA that is itself worse than the random walk (squared-error $p = 0.0017$, Holm 0.111; absolute-error $p = 0.0042$, unadjusted), and the EUR/USD win in the 70-day COVID-19 window ($p = 0.041$, unadjusted) reverses without that window (ratio 1.385). None of these recurs in the other pairs.

5. Implications for Model Validation and Governance

The results concern two models, a Random Forest and a GBM, three currency pairs, price-derived inputs and a test period from February 2018 to July 2021. Within that scope, the findings say more about how forecasting models must be evaluated before anyone relies on them than about whether machine learning works in general.

5.1. A Validation Standard

Three parts of the design changed the conclusions a reader would draw: the benchmark, the information set and the correction for multiple testing. Against a naive benchmark alone, the GBM seems to predict GBP/USD volatility better (Table 7) and to predict next-day absolute returns better in all three pairs (Table 11). Against ARIMA, it does neither. For GBP/USD a constant forecast equal to the training-period mean matches the GBM and also beats trailing volatility, so that win comes from a noisy benchmark, not from real structure the model has captured. The one-day wins come from a model that early stopping reduced to between one and 18 trees, compared with a single day's absolute return. A model that beats only the weakest available benchmark has not shown a real advantage over the method it would replace. This follows the logic of supervisory model-risk guidance, which requires outcomes analysis, benchmarking against alternative models and effective challenge by objective, informed parties (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011) .These results show why the choice of benchmark matters: the same forecasts can pass or fail depending on it.

The information each model can see matters as much as the benchmark. For volume, the Random Forest was 20 to 31 points less accurate than ARIMA (Table 6). Most of that gap comes from the fact that ARIMA forecast from the volume series while the Random Forest did not. When both are

given the volume history, the classifiers are not significantly different from ARIMA or from today's class (Table 10). A comparison that does not check what each model sees may wrongly credit the algorithm with a difference that really comes from the data.

Multiple testing is the third element. Across the 114 formal comparisons, six results favoured a machine-learning model at the 5% level and four survived the Holm correction, all against naive baselines (Table 12). A reader who saw only the GBP/USD return forecast, which beats ARIMA with $p = 0.0017$, or the EUR/USD COVID-19 window, with $p = 0.041$, would see a clear advantage over a time-series model. When the full family of tests is considered, with the qualifications in Section 4.8, neither result holds. Taken together, the study points to a minimum standard for validating a forecasting model: a chronological split with tuning limited to the training period; benchmarks that forecast the same target on the same days, including a simple time-series model and the latest observed value; a clear statement of the information each model uses; tests that respect the structure of the forecasts, such as the Diebold–Mariano and McNemar tests; a declared family of tests with a multiplicity correction; results reported separately for stress periods; and code and data that let others repeat the comparison. None of these ideas is new. The results show what happens to the conclusions when all of them are used together.

5.2. Explanations Also Need Validating

Explanations of a model's forecasts are often offered as a reason to trust the model. In this study the two importance measures for the same Random Forest did not agree. Impurity-based importance, calculated from the training data, put the 20-day return volatility first for EUR/USD and GBP/USD, while permutation importance on the test set put it 22nd and 15th, and between 7 and 17 of the 26 predictors had zero or negative permutation importance (Section 4.5). An explanation drawn from the training data described a model that did not work out of sample. An explanation is therefore a claim about the model that must be checked out of sample, in the same way as the forecast.

SHAP values ((Lundberg et al., 2020; Lundberg & Lee, 2017) attribute each forecast to the inputs of the model; they describe the model, not the market. For EUR/USD and USD/CAD they show that the GBM was largely a function of recent volatility, and for GBP/USD that it relied mostly on moving averages of the price level, which are not stationary. Removing the moving averages left GBP/USD unchanged and improved the GBM for EUR/USD and USD/CAD, but in no pair did it

bring the GBM ahead of ARIMA (Section 4.6). An explanation can thus help to locate a fragile input, but it cannot show that a model is accurate. The European Union's rules on automated decision-making, from which the debate on a right to explanation began, concern decisions about individuals (Goodman & Flaxman, 2017) . An institution's currency forecast is not such a decision, so these rules apply here only by analogy.

5.3. Concentration and Correlated Models

Weak models matter beyond the firms that use them when many firms use similar models. The Bank for International Settlements discusses how the shared use of models and data can produce correlated behaviour or herding, and how reliance on a small number of providers can create concentration (Aldasoro et al., 2024) .The Financial Stability Board lists third-party dependencies and service-provider concentration, market correlations, and model risk, data quality and governance among the vulnerabilities that artificial intelligence may amplify in the financial system (Board, 2024) .This study did not test whether firms' models are correlated and offers no evidence on herding. Its relevance is narrower: if models built on the broker price data available from public repositories have not been shown to beat a simple benchmark, adopting them widely adds the risks these bodies describe without a demonstrated gain in forecast accuracy.

5.4. What the Study Does Not Show

The policy reading has limits. The study measures forecast accuracy, not profits after trading costs, forecast speed or the cost of running a model. It uses only price-derived inputs, so it says nothing about models that use fundamentals, order flow, news or intraday data, nor about other model classes such as neural networks. It covers three currency pairs in one period, so it does not show that the results hold in other markets or periods. Questions of fairness and non-discrimination arise when models make decisions about individuals (Mehrabi et al., 2021; Wachter et al., 2021) ; the forecasts studied here make no such decisions, and the study has no result on them. The argument of this section is therefore a standard of evidence, not a verdict on machine learning in finance.

6. Conclusions

6.1. Conclusions

In this study I examined whether a Random Forest and a GBM that use 26 price-derived indicators could forecast foreign-exchange trading activity and volatility more accurately than ARIMA

models and naive rules. I ran out-of-sample tests on EUR/USD, GBP/USD and USD/CAD from 2018 to 2021. When compared with ARIMA, GARCH and HAR, the Random Forest and the GBM did not outperform those benchmarks.

The Random Forest classified next-day volume correctly on only 31.8% to 41.6% of the test days, whereas ARIMA achieved 61.5% to 63.8%. When the Random Forest was given the volume history that ARIMA already uses, it was not significantly different from ARIMA, so the gap is mainly one of information. The GBM produced five-day volatility forecasts that were less accurate than ARIMA for EUR/USD and USD/CAD. For GBP/USD the GBM was not significantly different from ARIMA, with an RMSE ratio of 1.03 and a bootstrap interval from 0.98 to 1.07. For the price, ARIMA did not improve on the random walk.

I ran seven robustness analyses, including GARCH and HAR benchmarks, yearly re-fits, a single data source and other forecast horizons, and all of them left these results unchanged. A wider hyperparameter search, labelled R9, selected simpler models but did not alter the comparison with ARIMA.

When I compared the models with naive baselines, the results were mixed. Out of 114 tests, four machine-learning wins survived the Holm correction, and all of them were against naive baselines. One was the GBM beating trailing volatility for GBP/USD at the five-day horizon, a result that a constant training-mean forecast matches. The others were the one-day GBM, which is almost identical to a constant forecast, beating today's absolute return in all three currency pairs. A GBM without the price-level predictors, specified after the main results, also beat trailing volatility in all three pairs, without adjustment. Every win against a time-series benchmark was either nominal or unadjusted. The GBP/USD return forecasts were compared with an ARIMA benchmark that is itself worse than the random walk, and the EUR/USD volatility forecast in the 70-day COVID-19 window reverses when that window is excluded. The Random Forest showed a large gap between training and test accuracy, and its two importance measures disagreed. The GBM showed a weak and unstable signal rather than overfitting, with only a few trees selected.

These results are consistent with weak-form efficiency for the price. However, a null result does not prove efficiency. It shows only that, with these inputs and in this period, the models found no predictive structure beyond what a simple time-series model could capture.

6.2. Limitations

Table 13 lists the limits of the study, the evidence that bears on each and what each does and does not imply. One further limitation concerns the history of this paper: the analysis presented here is a complete re-run, and the original thesis analysis on which an earlier version relied no longer exists, so the two analyses could not be compared.

Table 13. Limitations of the study.

Limitation	Evidence in the study	What it does and does not imply
Three currency pairs and one chronological split (890 test days)	R4 re-fits the models each year; the Random Forest is below ARIMA in all twelve pair-years, and the GBM has a higher RMSE than ARIMA in nine	The pattern is stable across years in these pairs; it does not show that it holds for other pairs, assets or periods
The three pairs share the US dollar	All three are quoted against the dollar, so their errors share a common factor	The pairs are not independent replications; agreement across them is weaker evidence than three unrelated markets would give
Price-derived features only	Given volume history, the classifiers are not significantly different from ARIMA (Table 10)	The volume result depends on the information set; the study says nothing about fundamentals, order flow or news
Model classes and tuning	Two tree ensembles; GBM depth fixed at 3 and 16, 77 and 3 trees selected; R3 and the wider search in R9 selected simpler models (depth-1 GBM trees, depth-3 forests) and did not change the comparison with ARIMA	Other model classes (for example neural networks) and much larger searches were not tried; within the searches run, wider tuning did not change the comparison with ARIMA
ARIMA grid	All six selected volume and volatility orders lie on the edge of the p, q ≤ 5 grid; the three price orders do not	A wider grid would change only the benchmarks; it cannot affect the machine-learning forecasts
Tick volume is a proxy whose level depends on the provider	OANDA counts were 1.76, 1.60 and 1.22 times MetaTrader counts; R5 uses MetaTrader only	Volume results could differ with traded volume; the price and volatility results do not use volume
Two data sources, taken from community repositories of broker data, and 14 interpolated holidays	R5 repeats the analysis on MetaTrader data alone with the same conclusions; the FRED comparison checks the price levels	The join does not drive the results; the broker data are not an official record of the market, and their provenance rests on the repositories
Daily data only	The volatility target is a noisier proxy than intraday realised volatility (Section 3.2)	Comparisons with intraday studies that report machine-learning gains (Christensen et al., 2023) are not like for like
Non-converged ARIMA fits	Excluded from selection; one R5 model did not converge and is flagged (Section 3.5)	Affects only the benchmark's specification; the ARIMA benchmarks still outperform the machine-learning models on volume and volatility, although the price ARIMA does not beat the random walk
Analyses specified after the main results	R1 to R7, the two sensitivity checks and R9 were defined after the main results were known (Section 3.5)	They test robustness; they were not part of a pre-registered design, and their results are reported in full rather than selected
Short stress windows	70 days; 20, 86 and 25 high-volatility days	Too few days to rank models; the COVID-19 result is exploratory
A pooled Holm family	114 tests, defined in Section 3.4 and listed in Table S1	Conservative by design; the unadjusted and diagnostic results are reported in Tables S2 and S3

Limitation	Evidence in the study	What it does and does not imply
No trading costs or economic-value test	Not examined	Forecast accuracy is not profitability; a small accuracy gain could still be unprofitable, and the reverse
Only part of the study reproduced from a fresh download	The core pipeline was reproduced twice; R1 to R7 use the same code, and R4 reproduces the stored predictions to 5×10^{-9}	Further confirmation would come from re-running R1 to R9 from a fresh download

A null result in three pairs should not be read as showing that machine learning cannot forecast exchange-rate activity or volatility. It shows that these models, with these inputs, did not beat simple benchmarks in this period.

6.3. Future Research

Three extensions follow from the limits. First, the same design could be applied with richer inputs, such as macroeconomic fundamentals, order flow or news, to test whether the information set, rather than the algorithm, determines the outcome; the volume-history result suggests that it may. Second, intraday data would allow volatility to be measured as intraday realised volatility and activity to be measured more directly, which would bring the comparison closer to the studies that report machine-learning gains. Third, an evaluation of economic value, with transaction costs and risk limits, across more currency pairs and with rolling forecast origins, would show whether any accuracy difference, in either direction, matters in practice.

Supplementary Materials: Table S1 lists the 114 tests in the Holm family with their nominal and Holm-adjusted p-values; Table S2 lists the 41 diagnostic tests that are reported but not adjusted; Table S3 lists the stress-window, sensitivity-check, constant-benchmark and plain paired t-tests quoted in Section 4; Table S4 gives confusion matrices, macro-F1, volatility MAE, validation-fold values and seed spread; Table S5 gives the wider tuning check R9 (file Supplementary_Tables_S1-S5.xlsx). The code archive (Supplementary_Code_Archive.zip; https://doi.org/10.5281/zenodo.23341357) holds the analysis scripts, the pinned download script, the file hashes and the result tables.

Author Contributions: Shadia Saeed Alqahtani is the sole author and is responsible for the whole project: conceptualization, methodology, validation (re-running the analysis), writing and project administration. The software, formal analysis and data curation of the re-analysis were carried out with an AI system, as described under Use of Artificial Intelligence.

Funding: This research received no external funding.

Data Availability Statement: All source data are public and were downloaded at fixed commits. The OANDA one-minute bars (github.com/FutureSharks/financial-data) cover 1 January 2010 to 3 December 2012, and the MetaTrader daily bars (github.com/ejtraderLabs/historical-data) cover 4 December 2012 to 12 July 2021. The FRED noon buying rates (github.com/datasets/exchange-rates) were used for validation only. The code, the download script, a table of SHA-256 hashes for the 415 source files and 12 derived daily files (427 entries in total) and all result tables are archived at Zenodo (https://doi.org/10.5281/zenodo.23341357) and are also provided as supplementary material (Supplementary_Code_Archive.zip). No price or volume data are redistributed; the scripts rebuild every input from the public sources.

Use of Artificial Intelligence: The author used Claude (Anthropic), a generative AI system, to write and run the analysis code, set up the software environment and carry out the reproduction checks. The same system was used to draft and revise the manuscript text, including Sections 5 and 6, the abstract and the declarations, and to check the text against the output files and the cited sources. The author re-ran all analysis scripts and confirmed that her outputs match the reported results, reviewed and edited the full text, and takes full responsibility for the content of the paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). Intelligent financial system: How AI is transforming finance (BIS Working Papers No. 1194). Bank for International Settlements. https://www.bis.org/publ/work1194.htm

Alqahtani, S. S. (2026). Machine learning versus ARIMA forecasts for EUR/USD, GBP/USD and USD/CAD, 2010–2021: Code and outputs (Version 1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.23341357

Amat, C., Michalski, T., & Stoltz, G. (2018). Fundamentals and exchange rate forecastability with simple machine learning methods. Journal of International Money and Finance, 88, 1–24.

Bank for International Settlements. (2025). OTC foreign exchange turnover in April 2025. Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets in 2025. https://www.bis.org/statistics/rpfx25_fx.htm

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213.

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7). 4 April 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

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

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

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

Brownlees, C. T., Cipollini, F., & Gallo, G. M. (2011). Intra-daily volume modeling and prediction for algorithmic trading. Journal of Financial Econometrics, 9(3), 489–518.

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109(11), 1997–2028.

Cheung, Y.-W., Chinn, M. D., & Pascual, A. G. (2005). Empirical exchange rate models of the nineties: Are any fit to survive? Journal of International Money and Finance, 24(7), 1150–1175.

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

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

Colombo, E., & Pelagatti, M. (2020). Statistical learning and exchange rate forecasting. International Journal of Forecasting, 36(4), 1260–1289.

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

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427–431.

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

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

Financial Stability Board. (2024). The financial stability implications of artificial intelligence. 14 November 2024. https://www.fsb.org/uploads/R14112024.pdf

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

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics, 20(7), 873–889.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

International Monetary Fund. (2024). Advances in artificial intelligence: Implications for capital market activities. In Global Financial Stability Report, October 2024: Steadying the course—Uncertainty, artificial intelligence, and financial stability (Chapter 3). International Monetary Fund.

Kılıç, R. (2025). Virtue or mirage? Complexity in exchange rate prediction (Finance and Economics Discussion Series 2025-089r1, revised version, 2026). Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2025.089r1

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. Annals of Statistics, 17(3), 1217–1241.

Lo, A. W. (2004). The adaptive markets hypothesis. Journal of Portfolio Management, 30(5), 15–29.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE, 13(3), e0194889.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157.

Meese, R. A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of International Economics, 14(1–2), 3–24.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115.

Menkhoff, L., & Taylor, M. P. (2007). The obstinate passion of foreign exchange professionals: Technical analysis. Journal of Economic Literature, 45(4), 936–972.

MetaQuotes. (n.d.). History data structure (MqlRates). MQL5 reference. https://www.mql5.com/en/docs/constants/structures/mqlrates

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

OANDA. (n.d.). Instrument: Candlestick definitions. OANDA v20 REST API documentation. https://developer.oanda.com/rest-live-v20/instrument-df/

Park, C.-H., & Irwin, S. H. (2007). What do we know about the profitability of technical analysis? Journal of Economic Surveys, 21(4), 786–826.

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

Pesaran, M. H., & Timmermann, A. (1992). A simple nonparametric test of predictive performance. Journal of Business & Economic Statistics, 10(4), 461–465.

Pfahler, J. F. (2022). Exchange rate forecasting with advanced machine learning methods. Journal of Risk and Financial Management, 15(1), 2.

Plakandaras, V., Papadimitriou, T., & Gogas, P. (2015). Forecasting daily and monthly exchange rates with machine learning techniques. Journal of Forecasting, 34(7), 560–573.

Rossi, B. (2013). Exchange rate predictability. Journal of Economic Literature, 51(4), 1063–1119.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (pp. 92–96).

Sheppard, K. (2024). arch: Autoregressive conditional heteroskedasticity models in Python (Version 7.2.0) (computer software). https://github.com/bashtage/arch

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41, 105567.

Wilder, J. W. (1978). New concepts in technical trading systems. Trend Research.

Yıldırım, D. C., Toroslu, I. H., & Fiore, U. (2021). Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. Financial Innovation, 7, 1. https://doi.org/10.1186/s40854-020-00220-2

Zhang, Y., & Hamori, S. (2020). The predictability of the exchange rate when combining machine learning and fundamental models. Journal of Risk and Financial Management, 13(3), 48.