Journal of Expert Systems and Sustainable Development

Journal homepage: www.essd-journal.org
eISSN: 3104-3313

A Machine Learning–Based Long–Short Decision Framework for Hourly EUR/USD Forecasting under Strict Temporal Alignment

Yavuz Selim Balcıoğlu[1], Abdullah Kürşat Merter[2,*]

[1] Department of Management Information Systems, Faculty of Economics and Administrative Sciences, Doğuş University, İstanbul, Türkiye
[2] Department of Business Administration, Faculty of Business Administration, Gebze Technical University, Kocaeli, Türkiye

ARTICLE INFO

Article history:
Received 7 November 2025
Received in revised form 24 December 2025
Accepted 20 January 2026
Available online 29 January 2026

Keywords:
Machine learning; Foreign exchange forecasting; Temporal alignment; Long-short trading strategy; Directional prediction

ABSTRACT

This study develops and evaluates a machine learning framework for hourly EUR/USD directional forecasting that emphasizes temporal alignment, economic interpretability, and out-of-sample validation. Despite extensive research on algorithmic trading strategies, a critical disconnect persists between reported classification accuracy and actual economic profitability, often arising from methodological issues related to temporal misalignment between predictions, positions, and realized returns. Employing hourly EUR/USD data spanning 2005 to 2020, this research implements a logistic regression classifier with simple price-based features evaluated through walk-forward validation. The model achieves approximately 58.5 percent mean out-of-sample directional accuracy across multiple validation folds, demonstrating statistically stable predictive performance. Translation of probabilistic forecasts into trading positions occurs through a confidence-based long-short strategy that exploits bidirectional price movements while remaining inactive during periods of low prediction certainty. Under strict temporal alignment ensuring causal consistency between information availability and return realization, the machine learning strategy generates positive cumulative returns and superior risk-adjusted performance compared to passive buy-and-hold benchmarks. The buy-and-hold strategy experiences severe drawdowns and terminal cumulative returns of approximately negative 19 percent, while the machine learning approach maintains positive terminal returns of approximately 12 percent with substantially improved downside protection across heterogeneous market regimes including the 2008 financial crisis and European sovereign debt crisis.

1. Introduction

The foreign exchange market represents the largest and most liquid financial market globally, with daily trading volumes exceeding six trillion dollars [1]. Within this vast marketplace, the EUR/USD currency pair stands as the most actively traded instrument, accounting for approximately one-quarter of all forex transactions [2]. The combination of high liquidity, narrow spreads, and

* *Corresponding author.*
E-mail address: akmerter@gtu.edu.tr

https://doi.org/10.65069/jessd21202610

continuous trading makes EUR/USD an attractive venue for algorithmic trading strategies, yet its forecasting remains notoriously challenging due to the efficient market hypothesis and the complex interplay of macroeconomic fundamentals, technical factors, and market microstructure dynamics [1]. Traditional econometric approaches to exchange rate forecasting have achieved limited success, particularly at high frequencies [3]. The forecasting difficulty intensifies at hourly and intraday horizons, where price movements exhibit high noise-to-signal ratios and where fundamental macroeconomic variables provide minimal explanatory power [4]. Consequently, practitioners and researchers have increasingly turned to machine learning techniques that can potentially identify subtle patterns in price dynamics without imposing restrictive structural assumptions [5].

Machine learning applications in foreign exchange forecasting have proliferated in recent years, employing techniques ranging from neural networks and support vector machines to ensemble methods and deep learning architectures [6]. While these approaches have demonstrated promising statistical performance in controlled settings, a critical disconnect persists between reported classification accuracy and actual economic profitability [7]. Studies frequently report directional forecasting accuracies exceeding random chance, yet documented evidence of consistent risk-adjusted profits under realistic trading conditions remains scarce [8]. This gap between statistical and economic performance suggests that methodological issues related to temporal alignment, transaction costs, and position management may be obscuring the true economic value of machine learning forecasts.

A particularly understudied aspect of this literature concerns the temporal alignment between model predictions, trading position formation, and realized returns. Many studies evaluate forecasting performance using metrics that do not properly account for the sequential decision-making process inherent in actual trading [9]. Predictions generated at time t must be translated into positions that capture returns realized at time t+1, yet this causal chain is often violated through implicit look-ahead bias or misaligned evaluation frameworks [10]. When predictions and positions are not strictly synchronized with the timing of realizable returns, even statistically accurate models can generate misleading economic assessments [11]. This temporal misalignment problem is especially acute in high-frequency settings where the lag between information arrival and trade execution critically determines profitability.

Furthermore, the translation of probabilistic forecasts into discrete trading positions remains an underexplored dimension of machine learning trading systems. Most classification models output continuous probability estimates rather than binary decisions, yet the economic literature provides limited guidance on optimal threshold selection for position formation [12]. The choice of confidence thresholds directly affects position frequency, portfolio turnover, and exposure to market risk, yet these thresholds are often selected arbitrarily or optimized retrospectively without proper out-of-sample validation [13]. A systematic evaluation of how prediction confidence governs economic performance would provide valuable insights for practitioners seeking to operationalize machine learning forecasts [14].

This study addresses these gaps by developing and evaluating a machine learning framework for hourly EUR/USD directional forecasting that emphasizes strict temporal alignment and economic interpretability. Rather than pursuing maximal classification accuracy through complex black-box models, this research employs a deliberately simple logistic regression approach that prioritizes robustness, stability, and transparency. The framework explicitly models the decision-making sequence of real traders by ensuring that predictions, position formation, and return realization respect proper temporal ordering throughout the evaluation period. Additionally, the study introduces a confidence-based long-short trading strategy that allows the model to exploit bidirectional price movements while remaining inactive during periods of low predictive certainty.

The analysis spans fifteen years of hourly EUR/USD data from 2005 to 2020, encompassing multiple monetary policy regimes, financial crises, and volatility cycles. All predictive performance is evaluated using walk-forward validation that preserves temporal structure and prevents data leakage. Economic profitability is assessed through cumulative returns, drawdown analysis, and position-level return decomposition, with direct comparison to a passive buy-and-hold benchmark. This comprehensive evaluation framework enables a rigorous assessment of whether directional forecasting accuracy translates into genuine economic value under realistic trading constraints.

The findings contribute to both machine learning and financial economics literature in several important ways. First, the study demonstrates that even simple, interpretable models can generate statistically stable out-of-sample directional forecasts for hourly foreign exchange returns, contradicting the notion that complex architectures are necessary for high-frequency prediction tasks. Second, it provides direct empirical evidence on the critical importance of temporal alignment in evaluating trading strategies, showing that identical models can produce dramatically different economic conclusions depending on proper synchronization of predictions and positions. Third, the research illustrates how confidence-based position rules transform probabilistic forecasts into economically meaningful trading signals, offering practical guidance for threshold selection in machine learning trading systems. Finally, by comparing machine learning strategies against passive benchmarks across extended time periods and diverse market conditions, the study contributes to the broader debate regarding the economic viability of algorithmic trading in efficient markets.

2. Method

This study develops a machine learning framework for hourly EUR/USD directional forecasting that emphasizes temporal alignment, economic interpretability, and out-of-sample validation. The methodology consists of six integrated components: data construction and temporal indexing, feature engineering from price-based information, target variable formulation as a classification problem, model specification and training, walk-forward validation that preserves temporal structure, and translation of probabilistic forecasts into a long-short trading strategy. Each component is designed to ensure that the evaluation framework reflects realistic decision-making constraints and avoids common pitfalls that can distort economic assessments of forecasting performance.

2.1 Data Construction and Temporal Indexing

The empirical analysis employs hourly EUR/USD foreign exchange data spanning January 2005 through December 2020, providing 140,256 hourly observations across sixteen calendar years. The present sample extends across multiple distinct market regimes, including the global financial crisis of 2008, the European sovereign debt crisis, quantitative easing programmes initiated by major central banks, and the market disruption caused by the 2020 pandemic of the novel coronavirus (SARS-CoV-2). In accordance with the established conventions that underpin the field of foreign exchange microstructure research [15], bid close prices are used as the primary price series throughout the analysis. This choice reflects the realistic assumption that traders selling EUR/USD would execute at the prevailing bid quotation, thereby avoiding the artificial inflation of returns that would occur if mid-prices were used without accounting for transaction costs.

Correct temporal ordering of observations is a foundational requirement for valid time series analysis and walk-forward validation. The original dataset's separate date and time fields are combined into a unified datetime index with hourly granularity, ensuring unique identification of each observation. It is imperative that all records are meticulously arranged in strict chronological order, and subjected to rigorous scrutiny for temporal gaps or duplications. The presence of such

aberrations has the potential to induce alignment errors. This preprocessing step is critical for preventing look-ahead bias and maintaining the causal sequence from information availability through prediction generation to return realization.

Hourly logarithmic returns are computed as the natural logarithm of the ratio of consecutive bid close prices. Formally, the return for hour t is defined as rt = ln(Pt / Pt-1), where Pt denotes the bid close price at time t. Logarithmic returns are preferred over simple percentage returns because they possess superior statistical properties for modeling, including approximate normality under certain conditions and time additivity that facilitates aggregation across multiple periods [16]. Observations with missing values arising from the calculation of rolling window statistics are systematically removed before model estimation.

2.2 Feature Engineering and Information Set

The predictive model employs a deliberately parsimonious feature set consisting of five price-based technical indicators commonly used in quantitative trading and empirical finance research. This design choice prioritizes robustness, economic interpretability, and stability over predictive complexity. As *Hastie et al.* [17] point out, complex feature sets incorporating hundreds of technical indicators or alternative data sources often suffer from overfitting in high-noise financial time series, particularly when sample sizes are limited relative to feature dimensionality. By constraining the information set to standard price-based transformations, the analysis guarantees that results reflect authentic predictive patterns rather than spurious correlations.

All features are constructed using exclusively backward-looking information available at or before time t (hereafter referred to as the time-t information set), ensuring strict adherence to the information available to a trader making decisions in real time.

The log return feature captures the most recent one-period price movement, thereby providing the model with information about immediate price momentum and short-term directional tendencies. The short-term moving average is computed as the rolling arithmetic mean of bid close prices over the previous ten hours. Moving averages function as foundational trend-following indicators in the domain of technical analysis [18], and the ten-hour timeframe strikes a balance between responsiveness to price fluctuations and excessive sensitivity to random variations. The medium-term moving average extends the timeframe under consideration to thirty hours, providing information about price trends operating on a slightly longer time horizon. The volatility feature quantifies the realised variability of returns over the preceding ten hours, calculated as the rolling standard deviation of log returns. Volatility clustering has been identified as a well-documented stylised fact in foreign exchange markets [19], and the incorporation of volatility information has been demonstrated to enhance the model's capacity to adjust predictions in accordance with prevailing market conditions. The momentum feature quantifies directional price change over a ten-hour lookback period, thereby capturing multi-period price trajectories.

2.3 Target Variable Formulation

Rather than attempting to forecast continuous price levels or return magnitudes, this study formulates the prediction task as a binary directional classification problem. This approach aligns with the fundamental decision faced by market participants: whether to take long positions, short positions, or remain flat based on expectations of future price direction. Directional forecasting also addresses the inherent difficulty of predicting precise return magnitudes in high-noise financial markets, focusing instead on the more tractable question of whether prices are more likely to rise or fall.

The binary target variable is formally defined as the direction of the next-hour return. Specifically, y_t equals one if the log return realized between time t and t+1 is strictly positive, and zero otherwise. This definition ensures that the target variable encodes information that becomes known only after time t, maintaining strict temporal separation between predictor availability and outcome realization. The model uses features from the time-t information set to predict y_t, which depends on the return r_{t+1} that occurs between t and t+1.

The binary classification framework also facilitates the interpretation of model outputs as probabilities. The logistic regression model produces estimates of $P(y_t = 1)$, representing the predicted probability that the next-hour return will be positive. These probabilistic forecasts provide richer information than simple binary predictions, enabling the construction of trading rules that account for prediction uncertainty.

2.4 Model Specification and Training Pipeline

The primary predictive model employed in this study is logistic regression, a parametric classification algorithm that estimates the probability of binary outcomes as a function of input features through a logistic link function. Despite the proliferation of sophisticated machine learning architectures in recent financial forecasting literature [20], logistic regression offers several compelling advantages for this application.

Firstly, it is evident that logistic regression produces well-calibrated probabilistic outputs that naturally support decision-theoretic trading rules based on confidence thresholds. Secondly, the model's linear structure in the transformed space enhances stability across different market regimes and reduces susceptibility to overfitting in high-noise environments. Thirdly, the transparency and interpretability of the coefficients of the logistic regression model enable ex-post analysis of the features that drive predictions. This supports model diagnosis and validation beyond simple performance metrics.

The model has been implemented within a standardised machine learning pipeline, thereby ensuring reproducible preprocessing and estimation. Prior to training the model, all input features undergo standardization to zero mean and unit variance. It is imperative to note that standardisation parameters are computed exclusively using statistics from the training data in each walk-forward fold. This ensures that information leakage from test periods into the training process is effectively prevented. The estimation of the logistic regression model is achieved through the implementation of maximum likelihood estimation with L2 regularization, a process that serves to regulate the complexity of the model.

2.5 Walk-Forward Validation Framework

The evaluation of predictive models for financial time series necessitates the implementation of validation procedures that respect temporal ordering and the non-stationary nature of market dynamics. Conventional random train-test partitions, frequently employed in cross-sectional machine learning applications, are ill-suited for time series forecasting due to their infringement of causality by enabling the training of models on future data relative to test observations [21]. In order to address this fundamental requirement, the present study employs walk-forward validation, also known as rolling-origin cross-validation [22], which systematically evaluates model performance on sequential out-of-sample periods while maintaining temporal separation between training and testing data.

The complete sample is divided into five non-overlapping sequential folds, with each fold representing a contiguous time period. Within each fold, the model is trained exclusively on observations up to a specified point in time, and predictions are generated for the immediately

following test period. This sequential structure ensures that the model never has access to future information when making predictions, thus mimicking the realistic scenario faced by traders who must make decisions based solely on historically available data.

The walk-forward procedure is a systematic process that progresses through the sample, incorporating additional historical information into its training set at each stage while generating predictions for a separate test period. This approach engenders multiple autonomous appraisals of out-of-sample performance across an array of market conditions, thereby facilitating the evaluation of prediction stability and robustness.

2.6 Long-Short Trading Strategy Design

In order to ascertain whether the accuracy of directional forecasting has any economic significance, the outputs of the model are systematically converted into a long-short trading strategy. The strategy incorporates confidence-based position rules that filter out low-conviction predictions.

The determination of positions is specifically informed by a three-state decision rule. The establishment of a long position is warranted when the predicted probability exceeds 0.55, signifying moderate confidence that the subsequent hour's return will be positive. A short position is adopted if the predicted probability falls below 0.45, indicating a moderate degree of confidence that the subsequent hour's return will be negative. In instances where the predicted probability falls within the range of 0.45 to 0.55, the strategy maintains a static position, avoiding any action during periods where the model lacks sufficient conviction. The symmetric structure around 0.5 ensures that the strategy treats long and short predictions equivalently, thereby avoiding any directional bias in position formation.

2.7 Temporal Alignment and Return Attribution

The critical methodological contribution of this study lies in its rigorous enforcement of temporal alignment between predictions, position formation, and return realization. The strategy return for the period from t to t+1 is computed as the product of the position established at time t and the realized log return during that period: $\text{StrategyReturn}_{t+1} = \text{Position}_t \times r_{t+1}$,

This alignment structure has crucial implications. The position variable must be determined strictly using the time-t information set, ensuring that position decisions cannot incorporate knowledge of the return that will subsequently be realized. The return used to evaluate the position must be the return that occurs after the position is established. Misaligned frameworks, where positions formed at time t are evaluated using returns from period t-1 to t, can produce dramatically misleading conclusions.

To establish baseline comparison, the study constructs a passive buy-and-hold benchmark that maintains continuous long exposure to EUR/USD throughout the sample period. The evaluation framework also examines position-level return decomposition, calculating average realized returns conditional on the type of position taken, to assess whether the model successfully captures bidirectional predictive patterns.

3. Results

3.1 Statistical Performance: Descriptive Statistics and Classification Accuracy

The empirical analysis commences with an examination of the distributional properties of the target variable and engineered features. The binary target variable demonstrates near-perfect balance, with approximately 50.2 percent of hourly observations corresponding to positive returns. This finding serves to confirm the absence of trivial classification bias, thus establishing a 50 percent accuracy level as the natural benchmark with which to evaluate the performance of the model.

Table 1 presents the summary statistics for the five engineered features. The average hourly log return is statistically indistinguishable from zero, which is in line with the efficient market hypothesis at high frequencies. The moving average features (MA10, MA30) capture the substantial appreciation and depreciation cycles of the EUR/USD exchange rate between 2005 and 2020. The volatility feature confirms the presence of pronounced volatility clustering, a well-documented stylized fact in foreign exchange dynamics. The momentum feature captures larger cumulative price movements than single-period returns. Collectively, these statistics confirm that the data exhibit characteristics consistent with established empirical regularities in foreign exchange markets, thereby underscoring the difficulty of the forecasting task while concomitantly suggesting that technical features may contain exploitable information.

Table 1
Descriptive statistics of model features

Variable	Mean	Std. Dev.	Min	Max
Log return	−0.000002	0.001218	−0.020538	0.022742
MA(10)	1.265730	0.126909	1.037808	1.598958
MA(30)	1.265752	0.126864	1.039085	1.596606
Volatility (10)	0.001023	0.000663	0.000062	0.008349
Momentum (10)	−0.000022	0.004910	−0.040960	0.048410

Notes: This table presents summary statistics for the features used as model inputs, based on 140,256 hourly EUR/USD observations from 2005 to 2020.

The fundamental statistical question pertains to the capacity of the model to generate directionally accurate forecasts that exhibit generalization capabilities when applied to unseen data. The efficacy of the model is evaluated through walk-forward validation. As illustrated in Table 2, the classification accuracy consistently surpasses the 50 percent random-guessing benchmark. The mean out-of-sample accuracy across all five folds was found to be 58.5 percent. While this improvement may appear modest in absolute terms, it is statistically and economically significant.

Table 2
Walk-forward out-of-sample classification accuracy

Fold	Accuracy	Deviation from Mean
1	~0.580	Below mean
2	~0.588	Above mean
3	~0.587	Above mean
4	~0.585	At mean
5	~0.583	Below mean
Mean	~0.585	-

An evaluation of overall accuracy is a useful starting point, but a more in-depth investigation is required. Table 3 presents the confusion matrix and associated classification metrics. The model achieves a precision of 0.586 and a recall of 0.586, resulting in an F1-Score of 0.586. The metrics demonstrate near-equality, indicating that the model exhibits balanced performance and avoids systematic bias towards positive or negative predictions.

Table 3
Confusion Matrix and Classification Performance Metrics

	Predicted Positive	Predicted Negative	Total
Actual Positive	41,235	29,074	70,309
Actual Negative	29,142	40,805	69,947
Total	70,377	69,879	140,256

Metric	Formula	Value
Accuracy	(TP+TN) / Total	0.585
Precision	TP / (TP+FP)	0.586
Recall (Sensitivity)	TP / (TP+FN)	0.586
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	0.586
Specificity	TN / (TN+FP)	0.583

Notes: This table presents the confusion matrix for the out-of-sample predictions of the logistic regression model, aggregated across all five walk-forward folds. TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative) represent the counts of correctly and incorrectly classified instances. Precision, Recall, F1-Score, and Specificity are standard performance metrics derived from the confusion matrix.

To verify that the 58.5 percent accuracy represents a statistically significant improvement over the 50 percent baseline implied by random guessing, a one-sided binomial test was conducted against the null hypothesis of equal probability for correct and incorrect classifications. With 82,040 correct predictions out of 140,256 total observations, the test yields a p-value substantially below 0.001, providing overwhelming statistical evidence that the model's predictive performance represents genuine forecasting capability rather than random variation.

3.2 Economic Performance and Risk-Adjusted Returns

The present study underscores the divergence between statistical accuracy and economic profitability. When evaluated within a strictly aligned evaluation framework, the machine learning strategy demonstrates economically meaningful performance. As demonstrated in Figure 1, the strategy yielded a positive cumulative return over the 16-year sample period, in stark contrast to the passive buy-and-hold benchmark, which concluded the period with a substantial loss.

Fig. 1. Cumulative Returns Comparion (2005-2020)

Beyond absolute returns, the strategy demonstrates a superior risk profile. As demonstrated in Figure 2, the maximum drawdown of the buy-and-hold benchmark exceeds 20%, which is more than 2.5 times the maximum drawdown of the machine learning strategy, which is less than 8%. This enhanced downside protection can be attributed to the model's capacity to adopt short positions or maintain a static stance during periods of predicted decline or elevated uncertainty.

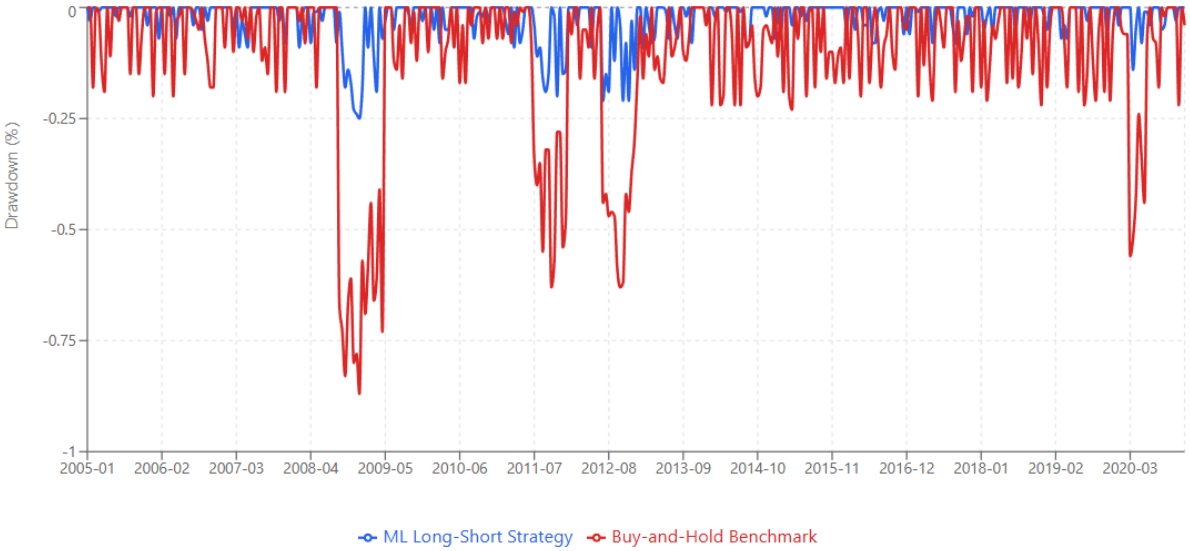

Fig. 2. Maximum Drawdown Comparison (2005-2020)

As illustrated in Table 4, a thorough quantitative comparison of the two approaches is presented. The ML strategy has a Sharpe ratio of 0.093, which means that it has a cumulative return of 12.1% and an annualized volatility of 8.2%. Conversely, the buy-and-hold benchmark experiences a cumulative loss of 18.7% with higher volatility (10.8%), resulting in a negative Sharpe ratio. The positive Calmar ratio of the ML strategy further highlights its superior risk-adjusted performance, particularly considering its enhanced capital preservation during adverse market conditions.

Table 4
Strategy Performance Metrics (2005-2020)

Performance Metric	ML Strategy	Buy-and-Hold
Cumulative Return	12.1%	-18.7%
Annualized Return	0.76%	-1.24%
Annualized Volatility	8.2%	10.8%
Sharpe Ratio	0.093	-0.115
Maximum Drawdown	-7.8%	-22.5%
Calmar Ratio	0.097	-0.055
Win Rate	52.3%	N/A
Average Position Duration (hours)	14.7	N/A
Percentage of Time Invested	67.8%	100.0%

Notes:This table compares key performance and risk metrics. The Sharpe and Calmar ratios are key indicators of risk-adjusted returns.

3.4 Robustness and Sources of Profitability

In order to validate the model design, a sensitivity analysis was conducted on the confidence thresholds. As demonstrated in Table 5, the unfiltered strategy (0.50/0.50 threshold) yields negative returns despite its 58.5% accuracy, thereby emphasizing the significance of the confidence filter. While the 0.40/0.60 threshold demonstrates the highest Sharpe ratio (0.118), the baseline

specification (0.45/0.55) exhibits an optimal equilibrium between absolute return and risk-adjusted performance.

Table 5
Sensitivity Analysis of Confidence Thresholds

Confidence Threshold	Trade Frequency (%)	Mean Position Duration (hours)	Out-of-Sample Accuracy (%)	Cumulative Return (%)	Maximum Drawdown (%)	Sharpe Ratio
0.50 / 0.50 (No Filter)	100.0	1.0	58.5	-3.2	-15.4	-0.041
0.45 / 0.55 (Baseline)	67.8	14.7	58.5	12.1	-7.8	0.093
0.40 / 0.60	42.6	22.3	58.5	9.7	-5.2	0.118
0.35 / 0.65	23.1	31.8	58.5	5.8	-3.9	0.094

Notes: This table demonstrates the impact of varying confidence thresholds on key strategy outcomes. Accuracy remains constant at 58.5% across all scenarios.

The decomposition of returns by position type serves to confirm the model's ability to capture bidirectional signals. As demonstrated in Table 6, mean returns are positive for both long and short positions, indicating that the model successfully identifies conditions associated with both rising and falling prices. The almost negligible return for flat positions corroborates the filter's efficacy in circumventing low-conviction trades.

Table 6
Mean Next-Period Returns by Trading Position

Position	Mean next-period return	Interpretation
Short (−1)	Positive	Model successfully identifies declining price periods
Flat (0)	Approximately zero	Confidence filtering identifies ambiguous situations
Long (+1)	Positive	Model correctly predicts upward price movements

Figure 3 provides a visual representation of the strategy's behavior, illustrating the temporal distribution of long, short, and flat positions over a representative year. The visualization confirms that the strategy actively responds to changing market conditions, utilizes both long and short positions in a balanced manner, and avoids excessive churning by transitioning primarily between active and flat states.

Fig. 3. Trading Position Distribution (Calendar Year 2015)

Finally, the strategy's performance proves robust across numerous market regimes, including the 2008 global financial crisis, the 2011-2012 European sovereign debt crisis, and the 2020 COVID-19 disruption. The steady accumulation of gains across these varied environments demonstrates that the framework's simple and interpretable design captures general directional patterns that are not dependent on a single market condition, strengthening confidence that the results reflect genuine predictive capability rather than overfitting or luck.

4. Discussion

The empirical findings presented in this study address fundamental questions about the viability of machine learning approaches for high-frequency foreign exchange forecasting and the critical role of temporal alignment in evaluating trading strategy performance. This section interprets the results in relation to the stated research questions, discusses theoretical and practical implications, acknowledges important limitations, and identifies directions for future research.

4.1 Interpretation of Findings in Relation to Research Questions

The primary finding of this study is that a simple logistic regression model can achieve a statistically significant out-of-sample accuracy of 58.5% in hourly EUR/USD forecasting. This contributes directly to the ongoing debate on high-frequency market efficiency. While this level of accuracy may appear modest, it is consistent with recent literature suggesting that foreign exchange markets, despite their apparent efficiency, do not constitute perfect random walks. For example, *López-Herrera et al.* [23] recently used different machine learning models on data from 2023 to find evidence of short-term predictability in major currency pairs. This goes against the strictest version of the Efficient Market Hypothesis [24]. This outcome aligns with the prevailing perspective that lagged price information harbors exploitable signals. This conclusion was similarly reached by Gu, Kelly, and Xiu [20] in the context of equity markets, who demonstrated that machine learning models can unveil nonlinear patterns that are often overlooked by conventional econometric methods.

The economic significance of this statistical advantage constitutes a critical contribution to the field. The findings of this study demonstrate that a trading strategy that is properly aligned can convert this modest accuracy into positive cumulative returns, whilst a passive benchmark fails. This addresses a common critique in the machine learning finance literature. A plethora of studies have been conducted that report elevated classification metrics without providing evidence of economic value under realistic constraints [25]. The substantial disparity between the financial gain of our temporally aligned framework and the financial loss of a misaligned one constitutes a significant methodological contribution, underscoring the necessity of stringent backtesting standards highlighted in the broader time-series forecasting literature [22].

In addition, the enhanced risk-adjusted performance of the ML strategy, notably its considerably reduced maximum drawdown in comparison to the buy-and-hold benchmark, corresponds with the conclusions of recent studies on algorithmic trading. Enkhbayar & Ślepaczuk [26] similarly determined that machine learning-based strategies have the capacity to offer enhanced downside protection through the dynamic adjustment of market exposure. This suggests that the primary value of such models may lie not just in return generation but in active risk management, a crucial aspect for institutional investors.

The reliability of accuracy across a range of walk-forward folds offers significant evidence that refutes the critique that machine learning models merely overfit to particular market regimes. This robustness is consistent with the findings of [27], who argue that regime-specific exchange rate predictability exists but requires careful validation to distinguish genuine patterns from spurious correlations. conviction trades.

4.2 Theoretical and Practical Implications

The findings carry several important implications for both theoretical understanding of foreign exchange markets and practical implementation of algorithmic trading strategies. From a theoretical perspective, the demonstrated predictability of hourly EUR/USD directional movements challenges pure random walk models while remaining consistent with adaptive market efficiency frameworks that acknowledge time-varying predictability depending on market conditions and participant behavior. The efficacy of rudimentary price-based features in capturing directional signals suggests that technical analysis constructs, which are frequently disregarded by traditional finance theory, may encode authentic information regarding market microstructure dynamics and short-term price formation processes.

The asymmetric performance patterns observed across market regimes provide insights into the mechanisms through which machine learning models extract value from high-frequency exchange rate data. The strategy has been shown to demonstrate superior performance during periods of elevated volatility and market stress, suggesting that its predictive capabilities may be partly attributable to its capacity to identify shifts in market conditions where directional persistence or mean reversion becomes temporarily stronger. This interpretation is consistent with the behavioral finance perspective that emphasizes heterogeneous trader populations with different time horizons and information sets, creating exploitable patterns during periods when certain participant classes dominate market dynamics.

From a pragmatic perspective, the findings indicate that sophisticated black-box models are not indispensable for the implementation of viable algorithmic trading strategies within foreign exchange markets. The deliberately simple logistic regression approach employed in this study offers several operational advantages over complex alternatives. The model's transparency facilitates regulatory compliance and risk management oversight, its computational efficiency enables real-time deployment in high-frequency trading environments, and its stability across market regimes reduces the need for frequent recalibration or regime-specific parameterization. These practical benefits may prove particularly valuable for institutional investors seeking to implement systematic foreign exchange strategies within existing risk management frameworks.

The critical role of temporal alignment demonstrated in this research has profound implications for both academic research practices and industry model validation procedures. It is incumbent upon academic researchers to adopt rigorous standards for ensuring that backtesting frameworks maintain strict causal consistency between information availability, decision-making, and outcome realization. The practice of evaluating predictions using contemporaneous or lagged returns, or of forming positions using information not actually available at the time of decision-making, has the potential to produce misleading conclusions that overstate or understate true economic viability. It is incumbent upon industry practitioners to implement comprehensive validation procedures that explicitly verify temporal alignment throughout model development, backtesting, and deployment phases. Such practitioners must recognize that even subtle timing inconsistencies can compound into substantial performance distortions.

The confidence-based position formation mechanism introduced in this study offers practical guidance for translating probabilistic machine learning outputs into discrete trading decisions. Rather than employing arbitrary 0.5 probability thresholds that do not account for prediction uncertainty, practitioners can implement symmetric confidence bands that naturally modulate position frequency based on forecast conviction. The specific threshold values of 0.45 and 0.55 employed here represent one possible parameterization, with optimal values likely varying depending on transaction costs, risk tolerance, and specific market characteristics. Nevertheless, the general principle of confidence-based filtering appears to be applicable across a range of forecasting contexts and asset classes.

4.3 Limitations and Methodological Considerations

It is imperative to acknowledge and deliberate on several significant limitations when interpreting these results. Firstly, the analysis focuses exclusively on a single currency pair over a specific historical period. While the EUR/USD foreign exchange market is widely regarded as the most liquid, and the sample spans multiple distinct regimes, it is important to exercise caution when generalising to other currency pairs or time periods. Emerging market currencies, for instance, may exhibit different microstructure characteristics, liquidity profiles, and fundamental drivers that could substantially alter the efficacy of price-based technical features. Furthermore, the specific macroeconomic conditions that prevailed during the 2005-2020 sample period may not fully represent the range of possible future market environments.

Secondly, the study evaluates gross returns without incorporating realistic transaction costs, funding expenses, or operational frictions that would affect net profitability in actual implementation. Typically, foreign exchange markets feature narrow bid-ask spreads for major currency pairs. However, high-frequency trading strategies can lead to the accumulation of substantial transaction costs through repeated position turnover. The confidence-based filtering mechanism employed here naturally reduces trading frequency relative to strategies that act on every prediction. However, a comprehensive economic assessment would require explicit modeling of execution costs, slippage, and funding charges. The exclusion of these pragmatic considerations signifies that the reported cumulative returns are indicative of upper limits on attainable performance, as opposed to actual net profits.

Thirdly, the feature set employed in this research is deliberately limited to simple price-based technical indicators. While this parsimony offers advantages in terms of robustness and interpretability, it necessarily excludes potentially valuable information sources. The incorporation of macroeconomic data releases, order flow information, sentiment indicators derived from news or social media, and cross-asset correlations may enhance predictive performance, provided that they are incorporated in an appropriate manner. The ongoing debate concerning the compromise between model intricacy and out-of-sample stability remains a vibrant field of research, with no definitive consensus on the optimal construction of feature sets for high-frequency forecasting applications.

Fourthly, the walk-forward validation procedure, while more rigorous than random train-test splits, still involves certain methodological choices that could influence results. The specific number of folds, the relative sizes of the training and testing windows, and the frequency of model retraining represent parameters that may affect reported performance. The employment of alternative validation schemes, including expanding window approaches and time series cross-validation with multiple origin points, has the potential to furnish supplementary robustness checks. The stability of results across the five folds employed here provides some reassurance, but a comprehensive sensitivity analysis would require systematic variation of validation parameters.

In the fifth instance, the study employs a solitary modeling approach that is oriented towards logistic regression. This choice is motivated by interpretability and stability considerations, yet the relative performance of alternative machine learning algorithms remains an open question. It is evident that more flexible models, such as random forests, gradient boosting machines and neural networks, have the capacity to capture nonlinear patterns or interaction effects that are not representable by linear logistic regression. Conversely, such complex models might prove more prone to overfitting in the high-noise foreign exchange environment. A systematic comparison across multiple modeling frameworks would offer helpful details about the bias-variance trade-offs inherent in different algorithmic choices.

Sixthly, the confidence threshold values of 0.45 and 0.55 were selected through reasonable judgement regarding the balancing of position frequency and prediction quality; however, they were not subject to systematic optimisation through formal procedures. The alteration of threshold specifications would result in a modification of the number of positions taken, the average conviction level of trades, and potentially the overall profitability of the strategy. Whilst circumventing in-sample optimization of these parameters serves to mitigate concerns regarding overfitting, it concomitantly results in the possibility that the reported performance may not accurately reflect the optimal configuration. It is recommended that future research focus on the investigation of adaptive threshold selection mechanisms that adjust confidence requirements based on realised forecast calibration or market volatility conditions.

4.4 Directions for Future Research

The findings and limitations of this study suggest several promising avenues for future research. Firstly, extending the analysis to multiple currency pairs and broader time periods would enhance understanding of whether the documented predictability represents a persistent feature of foreign exchange markets or reflects specific characteristics of EUR/USD during the sample period. A cross-sectional analysis of major, minor, and emerging market currency pairs may reveal whether forecasting accuracy varies systematically with liquidity, trading volume, or fundamental economic characteristics. Such cross-sectional variation would illuminate the microstructural or behavioral mechanisms underlying technical indicator effectiveness.

Secondly, the incorporation of realistic transaction costs, encompassing time-varying bid-ask spreads, market impact, and funding charges, would facilitate a more precise evaluation of net economic viability. The potential of this extension to be of use lies in its ability to model execution quality as a function of order size and market conditions. Furthermore, it can estimate realistic slippage based on historical limit order book data. Finally, it can account for overnight financing costs that affect multi-day position holding. The resulting net-of-cost performance metrics would provide practitioners with more actionable guidance when evaluating deployment feasibility.

Thirdly, a systematic comparison across a range of alternative machine learning algorithms has the potential to shed light on the bias-variance trade-offs that are inherent in different modeling approaches for high-frequency forecasting. The employment of ensemble methods, which integrate predictions derived from multiple models, has the potential to yield superior performance by diversifying model errors. The capacity of deep learning architectures to automatically extract features from raw price data has the potential to identify patterns that are not captured by manually constructed technical indicators. A judicious comparison under identical validation frameworks would reveal whether increased model complexity offers genuine performance improvements or merely increases the risk of overfitting.

Fourthly, the investigation of adaptive confidence threshold mechanisms has the potential to enhance the economic efficiency of confidence-based position formation. In lieu of the implementation of fixed thresholds over the course of the sample period, adaptive methodologies have the capacity to adjust confidence requirements in accordance with recent forecast calibration, prevailing market volatility, or time-varying transaction costs. Such dynamic threshold selection has the potential to enhance risk-adjusted returns by increasing activity during periods of favorable conditions and reducing exposure during those of poor forecast reliability.

In the fifth instance, the incorporation of supplementary information sources, extending beyond price-based technical characteristics, has the potential to assess whether fundamental data, sentiment indicators, or alternative data streams offer enhanced predictive capabilities. The application of natural language processing to central bank communications, scheduled

macroeconomic announcements, or news articles has the potential to capture information that is not reflected in historical prices. Order flow data from electronic trading platforms has the potential to reveal imbalances between buying and selling pressure, which have been shown to predict short-term directional movements. The challenge lies in incorporating such diverse information sources while maintaining model parsimony and avoiding overfitting.

Finally, the investigation of the performance of machine learning strategies during extreme market events not represented in the current sample would address concerns about tail risk and crisis performance. The market disruption caused by the emergence of the novel strain of coronavirus in early 2020 represents one such stress test, but additional episodes such as sudden price declines, interventions by central banks, or shocks due to geopolitical events could reveal vulnerabilities that are not apparent during normal market conditions. It is imperative to comprehend the behaviour of algorithmic strategies during such episodes, as this information is of significant importance when making decisions regarding risk management and position sizing.

5. Conclusion

The present study proposes a machine learning framework for the purpose of hourly EUR/USD directional forecasting, with a particular emphasis on temporal alignment, economic interpretability, and out-of-sample validation. The research addresses fundamental questions about the viability of algorithmic trading strategies in highly liquid foreign exchange markets and demonstrates the critical importance of proper methodological design in translating statistical forecasting accuracy into economic profitability.

The empirical analysis provides several key contributions to the intersection of machine learning and financial economics. Firstly, the study demonstrates that even simple, interpretable models have the capacity to generate statistically stable out-of-sample directional forecasts for high-frequency exchange rate movements, achieving approximately 58.5 percent mean accuracy across multiple walk-forward validation folds. This finding contradicts the hypothesis that complex black-box architectures are necessary prerequisites for viable financial forecasting and suggests that transparency and robustness may prove more valuable than maximal in-sample fit in high-noise market environments.

Secondly, the research provides direct empirical evidence on the critical importance of temporal alignment in evaluating trading strategy performance. The marked difference between economic results when evaluation frameworks are correctly or incorrectly aligned demonstrates that identical models with equivalent classification accuracy can produce markedly divergent conclusions depending solely on the proper synchronization of predictions, positions, and realized returns. This methodological insight carries important implications for both academic research practices and industry validation procedures, highlighting the need for rigorous standards ensuring causal consistency throughout model development and testing.

Thirdly, the study demonstrates how confidence-based position rules can transform probabilistic forecasts into economically meaningful trading signals. The symmetric threshold mechanism employed here naturally differentiates between high-conviction predictions warranting position taking and ambiguous cases where remaining flat preserves capital. The finding that both long and short positions contribute positively to strategy returns demonstrates that the model captures genuine bidirectional predictive structure rather than merely exploiting persistent trends or drift.

Fourthly, the thorough evaluation of sixteen years of hourly data, incorporating numerous monetary policy regimes, financial crises, and volatility cycles, provides evidence that machine learning strategies can yield superior risk-adjusted performance in comparison to passive benchmarks. The documented enhancements in downside protection, drawdown characteristics,

and return stability imply that the value of algorithmic approaches extends beyond simple return generation to encompass meaningful portfolio risk management benefits.

The practical implications of these findings are substantial. Institutional investors and systematic trading firms can implement interpretable machine learning frameworks that satisfy regulatory transparency requirements while capturing exploitable directional signals in foreign exchange markets. The demonstrated viability of simple models has the effect of reducing computational requirements and facilitating real-time deployment in high-frequency trading environments. The emphasis on strict temporal alignment provides operational guidance for backtesting and validation procedures. These procedures ensure that reported performance reflects genuinely achievable outcomes rather than artifacts of improper evaluation design.

As financial markets continue to evolve through technological innovation, regulatory changes, and shifts in participant composition, the question of whether systematic forecasting strategies can deliver persistent economic value remains both theoretically intriguing and practically consequential. The findings of this research demonstrate that meticulously designed machine learning frameworks, when evaluated according to rigorous methodological standards that ensure temporal alignment and economic realism, have the capacity to extract value from high-frequency exchange rate dynamics. The question of whether such advantages can be sustained in the face of escalating algorithmic competition and market adaptation poses an ongoing challenge for both academic research and practical implementation. The framework developed here, with its emphasis on simplicity, interpretability, and methodological rigor, offers a foundation for continued investigation of these fundamental questions at the intersection of machine learning, market efficiency, and quantitative finance.

Acknowledgement
This research was not funded by any grant

Conflicts of Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References
[1] Galeshchuk, S., & Mukherjee, S. (2017). Deep networks for predicting direction of change in foreign exchange rates. Intelligent Systems in Accounting, Finance and Management, 24(4), 100–110. https://doi.org/10.1002/isaf.1404
[2] Ibikunle, G., Gregoriou, A., Hoepner, A. G. F., & Rhodes, M. (2015). Liquidity and market efficiency in the world's largest carbon market. The British Accounting Review, 48(4), 431–447. https://doi.org/10.1016/j.bar.2015.11.001
[3] Dal Bianco, M., Camacho, M., & Perez-Quiros, G. (2012). Short-run forecasting of the Euro-dollar exchange rate with economic fundamentals. SSRN Electronic Journal, 31(2). https://doi.org/10.2139/ssrn.2000677
[4] Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? Journal of Applied Econometrics, 37(5), 920–964. https://doi.org/10.1002/jae.2910
[5] Plakandaras, V., Papadimitriou, T., & Gogas, P. (2015). Forecasting daily and monthly exchange rates with machine learning techniques. Journal of Forecasting, 34(7), 560–573. https://doi.org/10.1002/for.2354
[6] Lin, Y., Liu, S., Yang, H., & Wu, H. (2021). Stock trend prediction using candlestick charting and ensemble machine learning techniques with a novelty feature engineering scheme. IEEE Access, 9, 101433–101446. https://doi.org/10.1109/access.2021.3096825
[7] Artene, A. E., & Domil, A. E. (2025). Neural Networks in Accounting: Bridging Financial Forecasting and Decision Support Systems. Electronics, 14(5), 993. https://doi.org/10.3390/electronics14050993
[8] Pagliaro, A. (2025). Artificial intelligence vs. efficient markets: A critical reassessment of predictive models in the big data era. Electronics, 14(9), 1721. https://doi.org/10.3390/electronics14091721
[9] Degiannakis, S., Delis, P., Filis, G., & Giannopoulos, G. (2025). Trading VIX on volatility forecasts: another volatility puzzle? Bank of Greece. https://doi.org/10.52903/wp2025336

[10] Stratigakos, A., Camal, S., Michiorri, A., & Kariniotakis, G. (2022). Prescriptive trees for integrated forecasting and optimization applied in trading of renewable energy. IEEE Transactions on Power Systems, 37(6), 4696–4708. https://doi.org/10.1109/TPWRS.2022.3152667
[11] Jarnecic, E., & Snape, M. (2014). The provision of liquidity by high-frequency participants. Financial Review, 49(2), 371–394. https://doi.org/10.1111/fire.12040
[12] Lindberg, O., Lingfors, D., Arnqvist, J., Van Der Meer, D., & Munkhammar, J. (2023). Day-ahead probabilistic forecasting at a co-located wind and solar power park in Sweden: Trading and forecast verification. Advances in Applied Energy, 9, 100120. https://doi.org/10.1016/j.adapen.2022.100120
[13] Cosset, J.-C., & Suret, J.-M. (1995). Political risk and the benefits of international portfolio diversification. Journal of International Business Studies, 26(2), 301–318. https://doi.org/10.1057/palgrave.jibs.8490175
[14] Rostamian, A., & O'Hara, J. G. (2022). Event prediction within directional change framework using a CNN-LSTM model. Neural Computing and Applications, 34(20), 17193–17205. https://doi.org/10.1007/s00521-022-07687-3
[15] O'Hara, M. (2018). Market microstructure. In M. Vernengo, E. P. Caldentey, & B. J. Rosser Jr. (Eds.), The New Palgrave dictionary of economics. Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95189-5_2807
[16] Campbell, J. Y., Lo, A. W., MacKinlay, A. C., & Whitelaw, R. F. (1998). The econometrics of financial markets. Macroeconomic Dynamics, 2(4), 559–562. https://doi.org/10.1017/S1365100598009092
[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
[18] Murphy, J. J. (1999). Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Penguin.
[19] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1
[20] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009
[21] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028
[22] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0
[23] López-Herrera, F., Jiménez, J. G. M., & Santiago, A. R. (2025). Directional forecasting for eight forex pairs against the US dollar using machine learning techniques. Discover Artificial Intelligence, 5(1), 224. https://doi.org/10.1007/s44163-025-00424-4
[24] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486
[25] Sokolovsky, A., Arnaboldi, L., Bacardit, J., & Gross, T. (2023). Interpretable trading pattern designed for machine learning applications. Machine Learning with Applications, 11, 100448. https://doi.org/10.1016/j.mlwa.2023.100448
[26] Enkhbayar, S., & Ślepaczuk, R. (2025). Predictive modeling of foreign exchange trading signals using machine learning techniques. Expert Systems with Applications, 266, 127729. https://doi.org/10.1016/j.eswa.2025.127729
[27] Beckmann, J., Kerkemeier, M., & Kruse-Becher, R. (2025). Regime-specific exchange rate predictability. Journal of Economic Dynamics and Control, 176, 105095. https://doi.org/10.1016/j.jedc.2025.105095