Comparison of Random Forest Classification and Regression on Zigzag-Labeled Data for Equity Trading Strategies

Julius Mikelevičius, Ilja Jurčenko, Igoris Belovas

Vilniaus universitetas, Matematikos ir informatikos fakultetas
Naugarduko g. 24, Vilnius
julius.mikelevicius@mif.stud.vu.lt

Abstract. This paper compares Random Forest classification and regression approaches for equity trading strategy development using Zigzag-based price labeling. Both models are trained to identify buy and sell signals derived from Zigzag pivot points and are evaluated against a buy-and-hold benchmark using three exchange-traded funds: S&P 500, Hang Seng, and MSCI UK. A walk-forward framework is implemented using five-year training windows and two-year out-of-sample test windows. The results indicate that the Random Forest classification strategy delivers the strongest performance, particularly in the Hong Kong market, and is found to statistically significantly outperform the regression strategy. Further analysis of pivot timing shows that both models exhibit higher accuracy in predicting buy signals than sell signals, suggesting an asymmetry in model effectiveness across market turning points. As Zigzag labels are retrospective by construction, thus results should be interpreted as an upper bound on achievable performance rather than a direct estimate of live trading returns.

Keywords: random forest, zigzag labeling, equity trading, classification, regression.

1 Introduction

Machine learning has become more common in quantitative finance, with supervised learn-ing methods applied to stock return prediction, portfolio optimization, and systematic trading strategy development [1]. Among these methods, Random Forests (RF) have proven robust in financial applications due to their resistance to overfitting, ability to handle nonlinear relationships, and built-in feature importance measures [2].

This paper addresses the limited systematic exploration of Zigzag-based labeling in supervised learning [4, 9] by directly comparing two

DOI: https://doi.org/10.15388/LMITT.2026.18

complementary Random Forest mod-eling approaches applied to Zigzag-derived labels: a classification model that assigns binary trend-leg labels and a regression model that applies Gaussian density targets cen-tered on pivots. The primary goal is to determine which approach better captures Zigzag pivot structure and produces more effective trading signals. A buy-and-hold benchmark serves as a common reference point for evaluating the resulting strategies across three eq-uity markets. The specific contributions are: (1) a symmetric labeling framework where classification labels trend legs and regression applies Gaussian function around pivots, (2) a walk-forward evaluation with consistent train-test splits across markets, and (3) a di-rect comparison of both strategies using Wilcoxon signed-rank significance testing. Since Zigzag labels are forward-looking by design, the walk-forward framework serves as the safeguard with multiple out-of-sample test periods.

2 Literature review

In 2017, Krauss et al. [5] provided one of the most cited empirical benchmarks in this area, demonstrating that Random Forests, gradient-boosted trees, and deep neural networks can all generate statistically significant excess returns on the S&P 500 through daily return prediction, though profitability erodes over time as market efficiency improves.

Recently, Deep et al. [6] offered a related point of comparison, evaluating Random Forest regression models on SPY data using traditional technical indicators and finding a stark contrast between in-sample R^2 values of 0.749–0.812 and negative out-of-sample values, with all model configurations underperforming a buy-and-hold benchmark. Their feature importance analysis shows that raw price-based inputs dominate over oscillator-type indicators such as relative strength index (RSI), a finding broadly consistent with the trend-relative feature dominance observed in this study.

Wolff and Echterling [17] extended the Random Forest literature to individual stock selection, finding that predictive power is retained most strongly for smaller and less liquid equities, a pattern consistent with the cross-market performance asymmetry observed here.

A critical and often underappreciated dimension of supervised trading research is the labeling problem: how to assign a meaningful target to historical price series [3, 7]. The most widely used approach, fixed-horizon labeling [7, 8], assigns a positive class to peri-ods followed by a positive

return over a fixed window, but this ignores the path of prices and conflates distinct market regimes. López de Prado [3] formalized several alternatives, most notably the triple-barrier method, which accounts for stop-loss, take-profit, and time-expiry conditions simultaneously. However, both fixed-horizon and triple-barrier la-bels are highly sensitive to their parametrisation, prone to overfitting when thresholds are selected in-sample [3, 7], and do not explicitly target market turning points. Compared to the triple-barrier method, Zigzag labeling is a simpler approach requiring fewer parame-ters that directly target significant price turning points. Triple-barrier parameters require the specification of stop-loss and take-profit thresholds, which are themselves prone to in-sample overfitting [3]. The sensitivity of model performance to labeling design has been confirmed empirically: Song [7] proposed an *N* -period volatility labeling scheme for stock trading systems, arguing that standard up-down labeling fails to capture continu-ous trend properties and introduces substantial noise into training data; Han et al. [8] independently demonstrated that labeling method choice significantly affects downstream model performance through a genetic algorithm-based labeling approach. These findings motivate the use of Zigzag-based labeling in this study as a structurally distinct alterna-tive that anchors labels to significant price turning points rather than fixed horizons or arbitrary thresholds.

The Zigzag labeling offers precisely this philosophy: rather than projecting forward from each observation, it retrospectively identifies significant price pivots by requiring that reversals exceed a minimum percentage threshold, producing a sequence of alternating buy and sell pivots that corresponds closely to the turning points a technical trader would manually identify. Despite its intuitive nature and appeal among technical analysts, Zigzag-based labeling has received limited systematic academic exploration as a supervised learning target. Saberi et al. [4] adopted a related approach using dynamic threshold breakout labeling, demonstrating that threshold-based label construction can capture trend reversals more effectively than fixed-horizon alternatives. Qi et al. [9] incorporated the Zigzag indicator as an event marker within an event-driven LSTM framework for Forex prediction, explicitly acknowledging its look-ahead property as a structural characteristic rather than a flaw — though they use it to define input events rather than as a direct supervised learning target. To the authors' knowledge, no prior work has applied Zigzag labels as the direct supervised target for a

Random Forest trading framework, which constitutes a primary contribution of this paper. Masters [10] addressed the look-ahead bias inherent in any retrospective labeling scheme and advocates for permutation and randomisation testing as the primary safeguard, an approach adopted directly in this study.

A further dimension of the labeling question concerns the choice between classification and regression formulations. Breitung [11] directly compared Random Forest classifica-tion against Random Forest regression in a stock selection context, finding substantially higher returns and lower risk from the classification approach, a result directionally con-sistent with the classification advantage observed in this study, though the setting differs in that Breitung's framework targets stock picking rather than trade signal generation from price labels. This motivates the direct comparison undertaken here, where binary trend-leg labels (classification) are contrasted with continuous Gaussian function targets (regression) on the same underlying Zigzag structure.

Beyond labeling, the choice of input features is a central concern. Technical indicators derived from price and volume data, including momentum oscillators, moving average ra-tios, and volatility measures, remain competitive inputs to machine learning models [12]. The asymmetry between buy and sell signal predictability observed in this study is consistent with the behavioural finance literature: Khan et al. [13] noted that fear-driven selling is associated with emotional instability and tends to be more abrupt and less foreseeable than the gradual accumulation phases that precede market bottoms.

Walk-forward evaluation is the methodological standard for assessing out-of-sample strategy performance in supervised trading research, as it prevents the inadvertent use of future data in model selection or feature computation [3]. The use of multiple geograph-ically distinct markets follows the approach of Krauss et al. [5] and is further supported by Bustos et al. [14], who apply seven ML algorithms across 55 markets and 65 periods, finding that algorithm accuracy varies systematically with market efficiency, providing methodological justification for multi-market evaluation designs and for interpreting per-formance differences across the US, Hong Kong, and UK indices examined here.

3 Data and Feature Computation

The data for this study are obtained from the yfinance library for three instruments: SPY (S&P 500 ETF, US market), 2800.HK (Hang Seng ETF, Hong Kong market), and EWU (MSCI UK ETF, UK market) [15]. These indices were selected to provide geographic diversity across developed markets with distinct economic cycles, following the multi-market evaluation approach commonly used in the literature (cf. [5]).

To characterise the training data, eight technical indicators are employed: RSI (14-period), moving average convergence/divergence (MACD) line and signal line (12-day fast, 26-day slow, 9-day signal), 10-day price momentum, price relative to the 50-day simple moving average (SMA), price relative to the 200-day SMA, price relative to the 52-week high, and average true range (ATR) volatility ratio (14-day ATR divided by its 50-day mean). These features capture the momentum, trend, and volatility dimensions of price behaviour and are calculated separately for each training window to prevent information leakage [3, 12].

Table 1: Market characteristics: lag autocorrelation of daily returns, annualised volatility, and average daily dollar volume.

Ticker	Lag-1	Lag-5	Lag-10	Ann.	Volatility	Avg Daily Volume
SPY	−0.1066	−0.0143	−0.0036		0.1976	$23.4B
2800.HK	−0.0305	−0.0031	−0.0469		0.2412	$19.1B
EWU	−0.0945	−0.0040	−0.0221		0.2308	$0.04B

Dollar volume for 2800.HK converted from HKD to USD using daily closing exchange rates.

Table 1 reports the lag-1, lag-5, and lag-10 autocorrelation of daily returns for each market. All values are small in magnitude, consistent with the general efficiency of de-veloped markets, and do not show a clear ordering across markets. A more informative distinction lies in volatility and liquidity: SPY is the least volatile of the three markets (annualized volatility 0.1976), while 2800.HK (0.2412) and EWU (0.2308) exhibit more pronounced price swings. All three markets are highly liquid by dollar volume; SPY averages $23.4B and 2800.HK $19.1B in daily dollar volume (the latter converted from HKD to USD using daily closing exchange rates), with EWU considerably smaller at $0.04B. Since Zigzag-based strategies depend on significant price reversals to generate pivots, the higher volatility of 2800.HK and EWU could

provide more signal opportunities for the models to learn from, which can be connected with stronger strategy performance observed in those markets.

4 Methodology

Zigzag Labeling

The Zigzag indicator identifies price pivots by scanning for reversals exceeding a minimum movement threshold, set at 5% throughout this study. Starting from the first price, the algorithm tracks the running maximum or minimum in the current direction and records a pivot when price reverses by at least 5% from that extreme. This produces an alternating sequence of buy pivots (local minima) and sell pivots (local maxima).

For the classification models, the 20 trading days leading up to each buy pivot are labeled 1 for the buy model, and the 20 days leading up to each sell pivot are labeled 1 for the sell model, with all other days labeled 0; the model output is therefore a class probability. For the regression models, the target is a Gaussian function value, where higher values indicate proximity to a pivot; the model output is therefore a continuous score rather than a discrete class. Gaussian function centered on each pivot is

$$w_k = \exp\left(-\frac{1}{2}\left(\frac{k-p}{\sigma}\right)^2\right), \quad k \le p, \tag{1}$$

where p is the pivot position in trading days and σ = 20. This value was selected empirically, as it produced the most consistent results across markets, and was aligned with the classification window length to ensure both models operate on the same temporal scale, isolating the modeling approach as the variable under study. The density is zero for all days after the pivot, making the signal purely anticipatory. Separate buy and sell density series are computed by taking the element-wise maximum over all pivots of the respective type.

It is known that Zigzag labels involve inherent look-ahead bias [10], as the classification of a day as part of an upward or downward leg requires knowledge of the subsequent pivot. This look-ahead is structurally unavoidable in any supervised labeling scheme where tar-gets are derived from future prices, including common alternatives such as fixed-horizon

returns or triple-barrier labels. Rather than treating this as a limitation, the paper frames it as a deliberate design choice: the model is trained to approximate hindsight, learning what market conditions characteristically precede significant turning points, with the walk-forward framework then testing whether this learned approximation generalizes to unseen data [9]. Results should nonetheless be interpreted as an upper bound on achievable performance rather than a direct estimate of live trading returns; further assumptions and constraints of the evaluation framework are discussed in Section 6.

Model Training and Evaluation

A walk-forward framework is applied with 5-year training windows and 2-year test win-dows, stepping forward by 2 years at each iteration, yielding 6 non-overlapping test periods per ticker. Separate Random Forest models are trained for buy and sell signals for both classification and regression, giving four models per window (see Fig. 1). All models share the same hyperparameters: 1000 trees, maximum depth 10, minimum samples per split 10, minimum samples per leaf 5, and square root feature subsampling. Classification mod-els use balanced class weights to address the label imbalance inherent in the windowed labeling scheme [12].

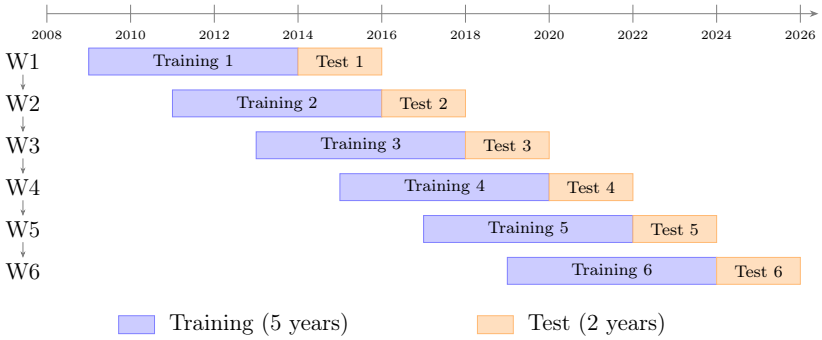

Figure 1: Walk-forward evaluation framework. Each window advances by two years, resulting in six non-overlapping test periods.

Decision thresholds are set at the 80th percentile of training set predictions, indepen-dently for each model. While alternative approaches such as directly optimising Sharpe ratio, return, or pivot timing accuracy

were explored, these consistently resulted in ex-tremely low trade counts. Such solutions, occasionally exhibiting favourable performances in bull markets, are not economically meaningful due to their minimal market participa-tion.

The percentile-based thresholding therefore serves as a practical constraint on signal frequency, ensuring that each model generates a sufficient number of trading opportunities. This allows for a more balanced evaluation of predictive quality under comparable trading intensity, rather than favouring models that achieve performance metrics through sparse and highly selective signals.

Trading Strategy and Performance Metrics

The trading strategy operates as follows. On each day of the test period, the buy model fires if its prediction exceeds the buy threshold, triggering entry into a long position if not already held. The sell model fires if its prediction exceeds the sell threshold, triggering exit from the position. The models act independently, there is no requirement for both to agree. A commission of 0.1% is applied on each trade; this reflects institutional execution costs and results may be sensitive to this assumption at higher retail commission rates given typical strategy trade frequencies. Signals are generated using end-of-day prices and executed at the following day's open, reflecting realistic trade execution constraints. The strategy is long-only. Long positions are more accessible to individual investors than short positions, which typically require margin accounts and carry theoretically unlimited downside risk; for this reason, the majority of systematic retail trading research focuses on the long side [17]. Extension to short selling is left for future work.

Performance is measured by total return, annualized return, Sharpe ratio, maximum drawdown (Max DD), and Calmar ratio. The benchmark is buy-and-hold over the same test period. Strategy returns are compared to buy-and-hold using the Wilcoxon signed-rank test, pooled across all ticker-period observations, with a significance level of $\alpha = 0.05$; this determines whether the observed performance differences are statistically meaningful or could plausibly arise by chance across the walk-forward periods. To further quantify how much of the observed performance may be attributable to the Zigzag labeling struc-ture rather than true predictive power, a Monte Carlo permutation test is implemented, in which pivot labels are randomly shuffled and the full walk-forward procedure is re-run; this distinguishes

genuine signal timing skill from returns that could arise from any sufficiently frequent trading strategy on the same instrument.

5 Results

Strategy Performance

Tables 2, 3, and 4 present the full walk-forward results for SPY, 2800.HK, and EWU respectively. Tables shows total return, Sharpe ratio, Calmar ratio, and maximum draw-down for buy-and-hold, the classification strategy, and the regression strategy across all six test periods. Table legend: B&H stands for buy-and-hold, CLF stands for classification strategy, REG stands for regression strategy, asterisks denote whether the model outperformed buy-and-hold on respective metrics, figures in bold signify which RF performed better that period.

Note that the Hang Seng market presents the most competitive results, with the clas-sification strategy averaging 13.3% against a buy-and-hold average of 13.5%, while the regression strategy trails considerably at 1.3%. Notably, the classification strategy outper-forms buy-and-hold in three of six Hang Seng periods, including the difficult 2022–2023 period where buy-and-hold lost over 20% while classification lost only 4.4%. The classification strategy also achieves a higher average Sharpe ratio (0.51 vs 0.38) and Calmar ratio (0.71 vs 0.43) than buy-and-hold on the Hang Seng, the only market where risk-adjusted outperformance is observed on average. US and UK markets show weaker strategy per-formance in absolute terms, with both strategies averaging below buy-and-hold for total return.

Maximum drawdown results are more favourable for the strategies. The classification strategy consistently shows lower maximum drawdown than buy-and-hold on the Hang Seng (average −13.3% vs −24.3%), suggesting that partial market participation reduces downside exposure even when it also reduces upside capture.

Examining trade-level win rates across all three assets reveals a nuanced picture. For SPY, the classifier achieves an average win rate of 48.6%, marginally above the regres-sion's 47.9%. For 2800.HK, the regression holds a modest edge at 54.7% against the classifier's 52.6%, and a similar pattern emerges for EWU, where the regression averages 52.4% compared to the classifier's 50.9%. Taken together, the regression model maintains marginally higher average win rates on two of the three assets, yet this

Table 2: SPY walk-forward results.

Period	Total Return (%)			Sharpe			Calmar			Max DD (%)			Win Rate (%)	
	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	CLF	REG
2014–15	16.0	**10.9**	10.5	0.62	**0.49**	0.47	0.65	0.43	**0.44**	-11.9	-12.2	**-11.6***	**53.8**	53.3
2016–17	38.2	**19.4**	12.5	1.62	**1.12**	0.74	1.92	**1.49**	0.57	-9.2	**-6.3***	-10.7	**55.0**	50.0
2018–19	24.3	-9.6	**-9.3**	0.81	-0.34	**-0.33**	0.60	**-0.21**	-0.25	-19.4	-23.0	**-18.9***	41.4	**57.5**
2020–21	50.9	-4.2	**9.1**	0.94	0.01	**0.31**	0.68	-0.08	**0.21**	-33.7	-26.6*	**-21.1***	47.0	**53.7**
2022–23	2.6	6.5*	**11.7***	0.16	0.26*	**0.39***	0.05	0.15*	**0.24***	-24.5	**-22.3***	-23.5*	37.5	**46.2**
2024–25	49.0	**34.4**	-2.5	1.31	**1.18**	-0.03	1.18	**0.93**	-0.05	-18.8	**-17.4***	-25.2	**57.1**	26.7
Average	30.2	**9.6**	5.3	0.91	**0.45**	0.26	0.85	**0.45**	0.19	-19.6	**-17.9***	-18.5*	**48.6**	47.9

Table 3: 2800.HK walk-forward results.

Period	Total Return (%)			Sharpe			Calmar			Max DD (%)			Win Rate (%)	
	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	CLF	REG
2014–15	0.7	**12.5***	10.9*	0.11	**0.50***	0.43*	0.01	**0.29***	0.24*	-25.8	**-21.5***	-22.2*	**66.7**	50.0
2016–17	49.8	18.6	**19.2**	1.42	**1.09**	1.06	1.68	**1.26**	0.88	-13.7	**-7.2***	-10.6*	50.0	**53.8**
2018–19	-0.1	**-0.3**	-4.9	0.09	**0.05**	-0.12	-0.00	**-0.01**	-0.12	-22.8	**-13.5***	-22.1*	**35.7**	42.9
2020–21	-13.4	**-9.7***	-16.1	-0.23	**-0.28**	-0.45	-0.28	**-0.30**	-0.44	-24.8	**-17.1***	-19.4*	50.0	**52.4**
2022–23	-20.8	**-4.4***	-22.7	-0.30	**-0.15***	-0.44	-0.29	**-0.20***	-0.38	-38.5	**-11.6***	-32.4*	**33.3**	40.0
2024–25	64.7	**62.9**	21.3	1.17	**1.87***	0.73	1.46	**3.25***	0.63	-20.1	**-8.8***	-16.6*	80.0	**88.9**
Average	13.5	**13.3**	1.3	0.38	**0.51***	0.20	0.43	**0.71***	0.14	-24.3	**-13.3***	-20.6*	52.6	**54.7**

Table 4: EWU walk-forward results.

Period	Total Return (%)			Sharpe			Calmar			Max DD (%)			Win Rate (%)	
	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	B&H	CLF	REG	CLF	REG
2014–15	-13.0	**-15.4**	-23.7	-0.35	**-0.54**	-0.85	-0.31	**-0.32**	-0.40	-21.9	**-25.2**	-31.9	**44.4**	30.4
2016–17	22.0	-8.3	**6.7**	0.67	-0.27	**0.32**	0.68	-0.24	**0.30**	-15.5	-17.5	**-11.1***	25.0	**52.6**
2018–19	3.0	**11.8***	0.2	0.18	**0.54***	0.07	0.07	**0.41***	0.01	-21.1	**-13.9***	-18.9*	**55.6**	43.8
2020–21	4.1	**19.4***	2.5	0.21	**0.45***	0.19	0.05	**0.34***	0.05	-42.4	**-27.6***	-27.6*	47.1	**54.5**
2022–23	6.2	**17.6***	3.9	0.25	**0.71***	0.21	0.12	**0.65***	0.08	-24.9	**-13.2***	-23.6*	**58.3**	50.0
2024–25	46.2	17.1	**17.2**	1.45	0.82	**0.82**	1.67	0.66	**0.66**	-12.6	-12.6	**-12.6**	75.0	**83.3**
Average	11.4	**7.0**	1.1	0.40	**0.28**	0.13	0.38	**0.25**	0.12	-23.1	**-18.3***	-21.0*	**50.9**	52.4

does not translate into superior performance. The classifier consistently outperforms on total return, Sharpe, and Calmar ratios. This apparent paradox is largely explained by trade frequency (see Table 7): the classifier makes fewer, more selective trades, whereas the regression trades more often and, despite winning slightly more of them individually, accumulates worse aggregate outcomes, suggesting that trade selectivity and the sizing of winning trades matter considerably more than raw win rate alone.

Performance in Extreme Market Conditions

An inspection across all three markets suggests that the classification strategy tends to perform relatively better during periods of unusual market conditions, such as sharp drawdowns, high volatility, or sudden reversals. For instance, in the Hang Seng and UK markets, the classification strategy outperformed buy-and-hold in periods of market stress (2022–23 for 2800. HK and 2018–19, 2020–21, 2022–23 for EWU) despite generally lagging the benchmark in calmer periods. This pattern is consistent with the pivot-based Zigzag labeling approach, which focuses on price turning points, and suggests that the model is most effective when market movements are abrupt or non-trending.

Statistical Significance

Table 5 presents Wilcoxon signed-rank test results pooled across all ticker-period obser-vations (18 observations per strategy pair).

Table 5: Wilcoxon signed-rank test results (two-sided), pooled across all tickers and periods.

Comparison	W statistic	*p*-value	Mean difference
CLF vs B&H	56.0	0.2121	−8.41%
REG vs B&H	17.0	**0.0016**	−15.78%
CLF vs REG	40.0	**0.0483**	+7.37%

Only the regression strategy underperforms buy-and-hold at conventional significance levels (p = 0.0016, mean difference −15.78%). The classification strategy also trails buy-and-hold directionally (mean difference −8.41%) but does not reach statistical significance (p = 0.2121). The difference between classification and regression is statistically significant (p = 0.0483), with classification outperforming regression by +7.37%

on average. It should be noted that pooling observations across three markets with materially different return distributions strengthens sample size but weakens the exchangeability assumption underlying the Wilcoxon test. Results should therefore be interpreted with caution at the individual market level.

Monte Carlo Permutation Test and Pivot Timing Analysis

Table 6 reports the Monte Carlo permutation test *p*-values for each ticker-period combi-nation (n = 1000 permutations per period). The *p*-value is the fraction of random signal permutations that matched or exceeded the actual strategy return; low values indicate that the observed timing is unlikely to arise by chance.

Table 7 summarises the mean absolute timing error between model signal peaks and actual Zigzag pivots in the test periods, averaged across windows, alongside the average number of trades generated per test window.

A consistent pattern emerges across all tickers and both model types: buy pivot tim-ing error is consistently lower than sell pivot timing error, with overall averages of 9.5 vs

15.4 trading days for the classification model and 10.0 vs 15.3 trading days for the regres-sion model. This asymmetry suggests that market bottoms are more predictable from technical indicators than market tops, which is consistent with the general observation that fear-driven selling is more abrupt and harder to anticipate than accumulation-driven buying [13]. The regression model generates on average more trades per window than the classification model (24.8 vs 18.5), suggesting that the continuous density target produces more frequent but not more accurate signals.

Table 6: Monte Carlo permutation test *p*-values (n = 1000). Bold entries indicate $p < 0.05$.

Period	SPY		2800.HK		EWU	
	CLF *p*	REG *p*	CLF *p*	REG *p*	CLF *p*	REG *p*
2014–15	0.347	0.323	0.132	0.161	0.628	0.868
2016–17	0.091	0.358	0.137	0.093	0.905	0.387
2018–19	0.825	0.821	0.341	0.383	0.070	0.316
2020–21	0.418	0.100	0.412	0.564	0.198	0.360
2022–23	0.259	0.116	0.176	0.637	0.086	0.344
2024–25	0.100	0.946	**0.009**	0.318	0.239	0.187
Average	*0.340*	*0.444*	*0.201*	*0.359*	*0.354*	*0.410*

Table 7: Mean absolute pivot timing error (trading days, lower is better) and average trade count per test window.

Ticker	CLF Buy	CLF Sell	REG Buy	REG Sell	CLF Trades	REG Trades
SPY	7.9	14.4	9.2	12.8	33.7	44.7
2800.HK	11.4	15.1	11.9	15.4	8.7	13.2
EWU	9.1	16.6	8.8	17.7	13.2	16.5
Overall	9.5	15.4	10.0	15.3	18.5	24.8

Feature Importance

A consistent pattern in feature importance is observed across all tickers and both model types. For sell models, Price_vs_SMA200 is the dominant feature, averaging approxi-mately 0.17 importance across all tickers and model types, followed by Price_vs_52w_high and Vol_Ratio. For buy models, Price_vs_SMA50 and Price_vs_SMA200 dominate, with Momentum_10 also contributing meaningfully. This suggests that trend-relative position-ing is the primary driver of both buy and sell signal generation, with shorter-term trend measures more relevant for buy decisions and longer-term measures for sell decisions.

Summary of Findings

Three major findings emerge. First, the regression strategy significantly underperforms buy-and-hold (p = 0.0016, mean difference −15.78%), while the classification strategy does not (p = 0.2121, mean difference −8.41%); classification significantly outperforms regression (p = 0.0483, mean difference +7.37%). Second, the results depend strongly on *market/strategy selection*: the Hang Seng market shows the most competitive results, with the classification strategy outperforming buy-and-hold in three of six test periods and delivering substantially lower maximum drawdowns (average −13.3% vs −24.3%). Third, buy pivot timing error is consistently lower than sell timing error across all models and markets, with overall averages of 9.5 vs 15.4 trading days for the classification model and 10.0 vs 15.3 trading days for the regression model.

6 Conclusions

This paper compared Random Forest classification and regression strategies trained on Zigzag-labeled price data across three equity markets, with the

primary goal of determin-ing which modeling approach better captures Zigzag pivot structure. The results consis-tently favour the classification approach: it outperforms regression in all three markets on average Sharpe ratio, produces lower maximum drawdowns, and is the only strategy that does not significantly underperform buy-and-hold at conventional significance levels. The classification strategy is also found to significantly outperform the regression strat-egy (p = 0.0483), strengthening the case for binary trend-leg labeling over continuous Gaussian function targets. However, the Monte Carlo permutation tests indicate that the observed timing skill is largely indistinguishable from chance: only one of eighteen ticker-period combinations achieves statistical significance (2800.HK CLF 2024–25, p = 0.009), suggesting that the strategies' returns could arise from trading frequency alone rather than genuine predictive power. This conclusion calls for further investigation of this aspect (see future research directions for a more detailed discussion). Even though strategies most of the time are not statistically better than random chance, their win rates for singular trades reveal that they have predictive power above 50% in the case for 2800.HK classifier and regressor (52.6% and 54.7% respectively) and EWU (50.9% and 52.4%respectively). These results alone do not prove the strategies' predictive powers, but they are nonetheless consistent with the presence of a modest informational edge, particularly when considered alongside the favourable risk-adjusted returns observed across multiple periods.

The strategies show a clear performance asymmetry across markets. In the Hang Seng market, the classification strategy demonstrates genuine competitiveness with buy-and-hold, particularly during periods of market stress. The SPY results are notably weaker, which is consistent with the efficiency of large-cap US equity markets (see Table 1): SPY exhibits the lowest annualised volatility of the three markets (0.1976), meaning price reversals are shallower and less structurally pronounced, giving the Zigzag indicator less signal to work with. Furthermore, in a more efficiently priced market, technical indicators carry less predictive signal, resulting in noisier and more uniformly distributed model predictions. This causes more days to cross the 80th percentile threshold, producing the highest trade frequency of the three markets (Table 7), which in turn increases commission drag and false signals, ultimately leading to weaker performance. However, this conclusion requires additional validation. The UK market falls between these two extremes, with both strategies showing competitive risk-adjusted performance in several periods despite lower average total returns.

Remarks and Further Research Directions

The Zigzag labeling scheme involves look-ahead bias by construction, as pivot classifica-tion requires knowledge of subsequent price movements. This means the labels represent an idealized historical signal rather than a truly causal one, and strategy results should be interpreted as an upper bound on achievable performance rather than a direct estimate of live trading returns. Furthermore, most strategies did not pass the Monte Carlo permu-tation tests, indicating that their apparent timing skill could arise largely from chance. The study is also limited to three instruments, long-only positions, and does not include a validation set for hyperparameter tuning, which is left for future work. The full source code is publicly available to support reproducibility [16].

The future work is to investigate the performance of Zigzag-based Random Forest strategies conditional on market regimes, as the observed results suggest that strate-gies perform relatively better during high-volatility or drawdown periods. Incorporating regime classification, volatility-adjusted thresholds, or tail-event detection may improve overall strategy robustness and allow for adaptive deployment depending on prevailing market conditions. Given that only one of eighteen tests achieved better statistically significant results than random permutations, a more rigorous approach could include adding a validation set to each walk forward period to optimize parameters of the strate-gies. Trading volume is not currently used as a feature; given its established role in technical analysis, its inclusion may improve signal quality [5]. Feature selection methods such as Boruta or Recursive Feature Elimination [2] were not applied here and could re-duce indicator noise, potentially improving signal quality further. Alternative ensemble models, most notably gradient-boosted trees such as XGBoost, could also be explored as replacements for or complements to the Random Forest, given their strong empirical track record in financial classification tasks [5]. Extension to short selling would allow a more complete evaluation of the framework, as long-only constraints limit strategy flexibility in bear markets. Finally, the signals generated by this framework could be incorporated into a broader portfolio optimisation setting, where classification and regression outputs serve as inputs to position sizing or multi-asset allocation models [12]. The Information Ra-tio could also be incorporated as an additional performance metric in future evaluations, quantifying active return relative to active risk against the buy-and-hold benchmark, which would better disentangle genuine alpha generation from risk reduction.

References

[1] Valencia-Arias, A., Gaviria Rodríguez, D.Y., Verde Flores, L., et al. (2025). Use of machine learning in the financial sector: an analysis of trends and the research agenda. *Discover Artificial Intelligence*, 5, 280. https://doi.org/10.1007/s44163-025-00539-8

[2] Dewi, C., Andika, R.A., Haryani, E., Riantama, D., Sajid, A., Alam, M.M., & Su'ud M.M. (2025). Feature selection for financial data classification using Random Forest, Boruta, and Recursive Feature Elimination. *Ingénierie des Systèmes d'Information*, 30(8), 2165–2173. https://doi.org/10.18280/isi.300822

[3] López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

[4] Saberi, E., Pirgazi, J., & Ghanbari Sorkhi, A. (2024). A machine learning approach for trading in financial markets using dynamic threshold breakout labeling. *The Journal of Supercomputing*, 80(17), 25188–25221. https://doi.org/10.1007/s11227-024-06403-3

[5] Krauss, C., Do, X.A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689–702.

[6] Deep, A., Shirvani, A., Monico, C., Rachev, S., & Fabozzi, F. (2025). Risk-adjusted performance of Random Forest models in high-frequency trading. *Journal of Risk and Financial Management*, 18(3), 142. https://doi.org/10.3390/jrfm18030142

[7] Song, Y.H., Park, M., & Kim, J. (2024). Improving the machine learning stock trading system: An N-period volatility labeling and instance selection technique. *Complexity*, 2024, 5036389. https://doi.org/10.1155/2024/5036389

[8] Han, Y., Kim, J., & Enke, D. (2024). Selective genetic algorithm label-ing: A new data labeling method for machine learning stock market trad-ing systems. *Engineering Applications of Artificial Intelligence*, 135, 108680. https://doi.org/10.1016/j.engappai.2024.108680

[9] Qi, L., Khushi, M., & Poon, J. (2021). Event-driven LSTM for Forex price prediction. *arXiv preprint*, arXiv:2102.01499. https://arxiv.org/abs/2102.01499

[10] Masters, T. (2020). *Permutation and Randomization Tests for Trading System De-velopment: Algorithms in C++*. Apress.

[11] Breitung, C. (2023). Automated stock picking using random forests. *Journal of Em-pirical Finance*, 72, 532–556.

[12] Ahlawat, S. (2025). *Statistical Quantitative Methods in Finance: From Theory to Quantitative Portfolio Management*. Apress.

[13] Khan, M.S.R., Yoshimura, H., & Kadoya, Y. (2024). Emotional instability and financial decisions: How neuroticism fuels panic selling. *Risks*, 12(12), 203. https://doi.org/10.3390/risks12120203

[14] Bustos, O., Pomares-Quimbaya, A., & Stellian, R. (2025). Machine learning, stock market forecasting, and market efficiency: a comparative study. *International Journal of Data Science and Analytics*, 20, 6815–6839. https://doi.org/10.1007/s41060-025-00854-4

[15] Aroussi, R. (2024). yfinance (Version 1.1.0) [Python library]. Retrieved from https: //github.com/ranaroussi/yfinance

[16] Mikelevičius, J. (2026). rf-clf-vs-rf-reg [Source code]. GitHub. https://github.com/juliusmikel/rf-clf-vs-rf-reg

[17] Wolff, D., & Echterling, F. (2024). Stock picking with machine learning. *Journal of Forecasting*, 43(5), 1385–1402.