HUIT Journal of Science ***26*** *(3E) (2026) 17-39*

DOI: https://doi.org/10.62985/j.huit_ojs.vol26.no3E.463

A COMPARATIVE ANALYSIS OF FORECASTING ACCURACY BETWEEN MACHINE LEARNING MODELS AND OLS REGRESSION: EMPIRICAL EVIDENCE FROM THE VIETNAMESE STOCK MARKET

Ngo Ly Huynh Phat*, Pham Chi Trung, Bui Le Vi

University of Economics Ho Chi Minh City, Viet Nam

*Email: *phatngo.31251570367@st.ueh.edu.vn*

Received: 23 March 2026; Revised: 19 April 2026; Accepted: 10 May 2026

ABSTRACT

This research investigates and compares the predictive performance of stock return forecasting between the Ordinary Least Squares (OLS) regression model and machine learning approaches in the context of the volatile Vietnamese stock market. Using a panel dataset combined with time-series data of listed firms on the Vietnamese stock exchange from 2015 to 2024, the study contrasts the OLS model with three advanced machine learning algorithms, including Artificial Neural Networks (ANN), Random Forest, and XGBoost. Predictive performance is primarily assessed using Root Mean Squared Error (RMSE), alongside Mean Absolute Error (MAE) and out-of-sample R-squared (R^2OOS) as robustness measures. The empirical results demonstrate that machine learning models significantly outperform OLS in capturing complex nonlinear relationships in stock returns. Among them, the ANN model achieves the lowest RMSE, indicating the highest predictive accuracy, and generates superior long–short portfolio returns compared to the other models. Furthermore, the Diebold–Mariano test confirms that the differences in predictive accuracy between machine learning models and OLS are statistically significant. Although OLS retains advantages in terms of simplicity and interpretability, machine learning models exhibit clear superiority in predictive performance and quantitative risk management. This study provides important empirical evidence from an emerging market such as Vietnam and offers practical implications for investors and policymakers in optimizing asset allocation decisions.

Keywords: Artificial neural networks, machine learning, OLS regression, stock return prediction, Vietnamese stock market.

1. INTRODUCTION

Over the past two decades, the global financial market economy has undergone complex fluctuations such as the dotcom bubble, financial crises, the era of low interest rates and abundant liquidity. Consequently, during these periods of economic instability, many investors have turned to other assets considered as “safe havens” to minimize risks [1]. The consequence of such actions by investors may reduce stock prices, leading to lower market performance [2]. Never has the stock market experienced such a deep decline that forced central banks to shift to rapid interest rate hikes to limit inflation, exerting strong pressure on stock valuations, especially in the technology sector and other high-risk assets. Developed economies such as the United States, Spain, and Italy were severely affected by the pandemic [3]. In addition, the global COVID-19 pandemic also had a significant impact on the Vietnamese stock market, causing the market index in Vietnam (VN-index) to decline sharply, from 1,000 points (at the end of 2019) to 645 points (at the end of March 2020) [4], reflecting investors' cautious sentiment and risk concerns. This poses considerable challenges for the Vietnamese stock market, particularly for investors, who are considered the most affected by financial market fluctuations.

Stock return prediction plays an oriented role in modern investment decision-making, providing investors with important insights into risk management and asset allocation. According to Agusta et al. [5], past market fluctuations may repeat, making predictive models useful tools for investment planning. In particular, the integration of artificial intelligence and financial theory is gradually creating increasingly sophisticated and empirically powerful forecasting models. In the past, the linear regression model OLS was considered a fundamental statistical tool. OLS serves as a foundational stepping stone to understanding modern Machine Learning techniques [6]. Furthermore, it offers high interpretability and transparent insights into the basic relationships between variables [7]. However, investors face mixed results when using the OLS model for stock return prediction. Messmer and Audrino [8] found that OLS has no predictive power for large-cap stocks, although it may be effective for small-cap and micro-cap stocks. In addition, Wang et al. [9] suggest that time-dependent weighting of the least squares method can improve standard OLS estimation compared to traditional OLS methods. Another barrier is that when data fluctuates strongly or contains outliers, OLS may produce biased or unreliable results, leading to incorrect interpretations of risk and investment returns [10]. In contrast, robust estimators may not be affected and can provide results resistant to influential outliers [11]. Furthermore, Maiti [12] confirmed the inefficiency of the linear OLS model in estimating tail distributions. Therefore, OLS cannot effectively handle nonlinear market fluctuations. Thus, the research question arises for investors: should they consider the linear OLS model as a baseline tool and complement it with more advanced machine learning techniques to comprehensively assess risk and improve the accuracy of stock return prediction?

Traditional econometric and financial models, exemplified by the ordinary least squares (OLS) method as a basic and simple approach, are constrained by strict linearity assumptions, limiting their ability to model complex, nonlinear relationships in data [6]. Therefore, to capture nonlinear relationships, machine learning models have gradually emerged, with algorithms such as Decision Tree Algorithm, Random Forest, and Neural Networks designed to handle nonlinear interactions more accurately [13]. Thus, the question arises: in the current context, whether Machine Learning is the best choice for investors when its advantages in speed and objectivity come at the cost of strict data requirements, operational costs, and risks related to interpretability?

Although previous studies have provided valuable insights into the predictive ability of the linear regression model (OLS) and machine learning models (XG Boost, Random Forest), there remain significant research gaps. This study aims to address specific issues associated with comparing the linear regression model (OLS) and machine learning models (XG Boost and Random Forest) in predicting stock returns, including application context, interpretability, and prediction stability over time.

Currently, the field of financial forecasting relies heavily on historical time-series data to accurately predict future asset performance [9]. However, stock price data in the market is highly complex, characterized by strong nonlinearity and volatility over time [14]. Therefore, improving predictive methods is essential. In the past, classical portfolio optimization models were often based on average expected returns. Freitas et al. [15] argued that this approach may create a "low-pass filtering effect" on the dynamic behavior of the stock market, leading to suboptimal investment strategies.

Gu et al. [16] show that machine learning models generally outperform linear models, and even generalized linear models do not significantly surpass standard ones. Therefore, this study aims to examine whether the OLS linear model can match the performance of machine learning models.

Notably, a single model is insufficient to accurately identify the true data-generating process or capture all characteristics of a time series [17]. DeMiguel et al. [18] suggested that machine learning models using inputs from machine learning forecasts rather than historical data can improve the quality of investment strategies. Moreover, Sebastião and Godinho [19] found that the accuracy of machine learning models in predicting stock performance tends to decline when market structures change. Machine learning represents a promising alternative for stock return prediction in time series, despite its simplicity and linearity. Therefore, this study will evaluate the robustness of machine learning models across different economic cycles compared to the linear OLS model, particularly their ability to adapt to "Concept Drift", a phenomenon where the changing statistical properties of the target variable cause initially trained models to become less effective.

Finally, previous studies have shown that the performance of these models can vary significantly depending on different contexts. Haider et al. [20] found that although machine learning techniques often outperform statistical methods, there are cases where traditional methods can be equally competitive. Wolff and Neugebauer [21] also reported empirical results showing that a linear model using the CM-IC method outperformed complex machine learning models in predicting stock returns in the U.S. market. Meanwhile, Lalwani and Meshram [22] demonstrated that machine learning models provide higher accuracy in predicting stock returns in the Indian stock market. Given the inconsistency in international findings, this study will examine the predictive performance of the linear regression model (OLS) and machine learning models in the specific context of the Vietnamese stock market, aiming to provide a comprehensive and objective comparison between the two approaches.

Building upon the identified research gaps and the inconsistency of prior empirical evidence, this study is conducted to clarify the effectiveness of stock return forecasting models within the context of the Vietnamese market. Specifically, the research focuses on evaluating and comparing the predictive performance of the traditional OLS model with modern machine learning algorithms, including ANN, Random Forest, and XGBoost. Furthermore, the study analyzes the differences in the importance of financial variables such as ROA, Beta, and firm size when processed by linear versus nonlinear models. Finally, the research verifies the practical efficiency of the forecasting models through the construction of a Long-Short portfolio investment strategy.

To address these objectives, the study tests the following hypotheses:

H1: Machine learning models exhibit higher accuracy in predicting stock returns compared to the baseline OLS model.

H2: The relative importance of fundamental financial variables differs significantly between OLS and machine learning algorithms.

H3: Machine learning models deliver superior risk-adjusted investment returns through long-short portfolio strategies compared to OLS models.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1. Theoretical Background

Stock return prediction has always been a major challenge in finance due to the noisy and continuously volatile nature of the market. This section presents the core theoretical foundations that explain asset price movements, while also providing a statistical framework to compare the predictive performance between traditional econometric models (OLS) and Machine Learning (ML) algorithms.

2.1.1. Efficient Market Hypothesis (EMH) and Random Walk

The Efficient Market Hypothesis (EMH), proposed by Eugene Fama [23], is the most fundamental theory in stock price prediction. According to this theory, financial markets are informationally efficient, meaning that stock prices at any given time fully reflect all available information. Therefore, future price changes only occur when new information arrives, which is entirely random and unpredictable [24].

However, in the context of comparing OLS and Machine Learning, EMH provides an important benchmark. If the weak-form EMH holds perfectly, both OLS and ML models would fail in predicting stock returns. In reality, behavioral finance studies have shown that markets exhibit various inefficiencies [25]. Static linear models such as OLS often fail to capture these signals due to their rigid structure. In contrast, Machine Learning algorithms are expected to overcome the limitations of EMH by identifying hidden nonlinear patterns and micro-level anomalies that are complicated for humans or traditional statistical models to detect [16].

2.1.2. Linear Regression Theory and the Gauss–Markov Theory

The Ordinary Least Squares (OLS) model has long been considered the gold standard in financial analysis due to its transparency and strong interpretability [26]. Mathematically, OLS seeks to find a linear function that minimizes the sum of squared residuals between actual and predicted values.

The theoretical foundation supporting OLS is the Gauss–Markov theorem. This theorem states that if the data satisfy strict assumptions such as linearity, no perfect multicollinearity, and homoscedasticity, then OLS provides the Best Linear Unbiased Estimator (BLUE). However, this is also the main limitation of the method in practice. Real-world stock return data are highly noisy and often violate these assumptions, exhibiting heavy-tailed distributions, heteroscedasticity, and multicollinearity among macroeconomic variables [27]. These violations significantly reduce the predictive power of OLS and create the need for more flexible nonlinear approaches such as Machine Learning.

2.1.3. Statistical Learning Theory and the Bias–Variance Tradeoff

The transition from traditional econometric models to modern computational algorithms is grounded in Statistical Learning Theory [28]. While OLS primarily focuses on causal inference, Machine Learning emphasizes predictive accuracy. Within this framework, the performance of any predictive model is governed by the bias-variance tradeoff [29].

Bias arises when a model is too simple to capture the complexity of reality. In contrast, variance measures how much model predictions would change if trained on different datasets. OLS represents a high-bias, low-variance model: it is stable but often suffers from underfitting because financial markets do not follow linear patterns. Conversely, algorithms such as Random Forest and Neural Networks exhibit low bias but high variance. They flexibly adapt to complex data patterns but face the major risk of overfitting, where the model memorizes noise instead of learning the true underlying structure [30].

2.1.4. Arbitrage Pricing Theory (APT) and Factor Models

For mathematical models (whether OLS or ML) to function, they require input variables. The Arbitrage Pricing Theory (APT), developed by Stephen Ross [31], provides the theoretical foundation for selecting these variables. APT states that the expected return of a financial asset can be modeled as a linear function of multiple macroeconomic risk factors or market-specific characteristics [32].

Although both OLS and ML utilize similar sets of input factors based on APT, their processing mechanisms differ significantly. OLS assigns fixed and independent weights to each factor, assuming additive effects. In contrast, Machine Learning models can automatically identify complex cross-interactions among variables - for example, an increase in interest rates may negatively affect stocks only when inflation exceeds a certain threshold. This flexibility in handling multidimensional relationships represents a key advantage of ML over OLS [33].

2.2. Literature review

The main objective of this section is to analyze and compare the performance of the Ordinary Least Squares (OLS) model and Machine Learning (ML) models in predicting stock returns by reviewing prior studies. However, most traditional studies rarely explored the power of nonlinear algorithms and primarily focused on linear regression models based on macroeconomic factors. In contrast, recent investigations reveal that Machine Learning models significantly outperform traditional methods when handling large-scale financial data.

2.2.1. General Comparison between OLS and Machine Learning Algorithms

Al-Jawarneh et al. [34] conducted a study to evaluate the predictive performance of OLS compared to basic ML algorithms such as Ridge and Lasso regression. Their findings indicate that ML models serve as effective signal filtering tools in developed stock markets. However, this advantage is less pronounced in emerging markets due to higher data noise.

2.2.2. Prediction under Extreme Volatility and High-Frequency Data

Gupta et al. [35] examined the predictive performance of these models during periods of extreme volatility such as the COVID-19 pandemic. Their findings show that advanced algorithms like Random Forest can accurately forecast stock movements, although performance declines when sentiment-related variables are absent. Among the tested models, Random Forest demonstrated the strongest predictive power under high uncertainty compared to OLS.

Moreno-Pino & Zohren [36] further analyzed how data frequency affects model performance using high-frequency trading datasets [36]. The results indicate that OLS becomes significantly less effective under high-frequency data conditions, whereas Neural Networks achieve substantially higher prediction accuracy.

2.2.3. Model Performance across Markets and the Role of Data Processing

Chun et al. [37] conducted a study on the U.S. and South Korean stock markets and found similar results, indicating that Machine Learning models can serve as optimal tools for quantitative investment strategies [37]. Meanwhile, Ji et al. [38] applied Random Forest combined with wavelet denoising to predict stock price movements [38]. Their findings show that extending historical data improves performance only up to a certain limit, and distant historical data contributes little to predictive accuracy.

As shown in Table 1, feature selection techniques help reduce noise and significantly improve performance in mature markets such as the U.S. but have weaker effects in Asian markets.

Overall, the transition from static linear models such as OLS to more complex Machine Learning algorithms reflects not only advancements in data science but also highlights the necessity of nonlinear methods in capturing the increasingly complex dynamics of modern financial markets.

Table 1. The improvement in F1-score using feature selection in different markets

Market (Representative Index)	Initial F1-score (Using all 18 features)	Best F1-score (After feature selection)	Optimal number of retained features	F1-score improvement (%)
SSEC (Chinese)	0.724	0.732	10	1.00
HIS (Hong Kong)	0.764	0.785	6	2.70
DJI (USA)	0.703	0.752	8	7.00
S&P 500 (USA)	0.725	0.768	5	5.90

Source: Author's calculations

3. METHODOLOGY

3.1. Data and sample selection

In this study, the authors employed a quantitative approach to evaluate and compare the predictive capabilities of stock returns using both the OLS model and several machine learning algorithms. The dataset utilized consists of panel data combined with time-series elements, including stock prices, trading volumes, and financial indicators of publicly traded companies on the Vietnamese stock market from 2015 to 2024. All information was sourced from the financial database platform FiinPro. After collection, the data were carefully processed and standardized on a monthly frequency to ensure consistency across the observed variables.

Given the time-series nature of the data, the study conducted unit root tests to assess the stationarity of the variables before incorporating them into the predictive models. Specifically, two statistical tests were applied: the Augmented Dickey-Fuller (ADF) test and the Phillips-Perron (PP) test [39], [40]. Additionally, a correlation matrix was constructed to examine the initial linear relationships among the independent variables. To mitigate potential multicollinearity, the Variance Inflation Factor (VIF) was calculated to assess the degree of multicollinearity between variables; variables exceeding the permissible VIF threshold were considered for removal or adjustment [41].

The initial dataset was rigorously cleaned to handle missing values and outliers. Missing data were addressed using the forward-fill method for time-series consistency, while extreme outliers at the 1st and 99th percentiles were winsorized to prevent skewed estimations. The final filtered sample comprises exactly 1,650 listed firms, yielding a total of 197,424 firm-month observations. Furthermore, to verify the absence of severe multicollinearity, Variance Inflation Factor (VIF) scores were computed; all variables recorded VIF values strictly below the standard threshold of 5, ensuring the robustness of the regression coefficients.

3.2. Model and methodology

To start the comparative analysis, the study established two separate groups of models:

(1) The initial group includes the Ordinary Least Squares (OLS) model, serving as the baseline model to illustrate the linear relationship between dependent and independent variables (Carl Friedrich Gauss). The general equation is expressed as:

$$R_{i,t+1} = \alpha + \beta_1.X_{1,i,t} + \beta_2.X_{2,i,t} + \ldots + \beta_k.X_{k,i,t} + \varepsilon_{i,t}$$

In which:

$R_{i,t+1}$: Stock return of firm *i* at forecast time $t+1$

α : Intercept coefficient

β : Regression coefficient corresponding to the independent variables $X_{i,t}$ (including financial and macroeconomic indicators at time *t*)

$\varepsilon_{i,t}$: Random error

Table 2 provides a detailed summary of the variable definitions, measurement methods, and their expected relationships with future stock returns.

Table 2. Variable definitions, measurement, and expected relationship with future stock returns

Variable category	Variable name	Symbol	Description / calculation	Measurement window / frequency	Expected relationship	Key reference(s)
Dependent variable	Future stock return	R_i,t+1	Holding-period return of firm i in the next prediction period: (P_i,t+1 - P_i,t + D_i,t+1) / P_i,t, where dividends are available. This is the target variable for OLS, ANN, Random Forest, and XGBoost models.	Next month, quarter, or year depending on the research design. Predictors at time t must be lagged before matching with R_i,t+1.	N/A	[42], [43]
Size and value	Firm size	SIZE	Natural logarithm of market equity: ln(Price × shares outstanding).	Measured at portfolio/model-formation date or at the end of the previous period.	Negative / mixed	[42]
Size and value	Book-to-market ratio	B/M	Book value of equity divided by market value of equity.	Use the most recently available annual financial statements; lag accounting data before matching with returns.	Positive	[42]
Size and value	Earnings-to-price ratio	E/P	Net income divided by market capitalization. This can be computed as earnings per share divided by price when share-level data are available.	Annual financial statement information matched to subsequent stock returns after an appropriate reporting lag.	Positive	[44]
Size and value	Dividend yield	D/P	Cash dividends per share divided by stock price per share; alternatively, total cash dividends divided by market capitalization.	Annual or trailing-12-month dividend information, depending on data availability.	Positive	[45]
Profitability and growth	Return on assets	ROA	Net income divided by total assets.	Annual or quarterly financial statements; lag before matching with subsequent returns.	Positive / mixed	[46], [47]
Profitability and growth	Return on equity	ROE	Net income divided by shareholders' equity; in some asset-pricing applications, income before extraordinary items divided by book equity is used.	Annual or quarterly financial statements; lag before matching with subsequent returns.	Positive	[47]
Profitability and growth	Gross profitability	GP/A	Gross profit divided by total assets. This is preferable to gross margin when citing the gross profitability premium literature.	Annual or quarterly financial statements; lag before matching with returns.	Positive	[48]
Profitability and growth	Sales growth	SG	Year-over-year percentage change in sales: (Sales_t - Sales_t-1) / Sales_t-1.	Annual or quarterly financial statements; lag before matching with returns.	Negative / mixed	[49]
Investment	Asset growth	AG	Year-over-year percentage change in total assets: (Total assets_t - Total assets_t-1) / Total assets_t-1. This variable proxies for firm investment intensity.	Annual financial statements; lag before matching with returns.	Negative	[46], [50]

Variable category	Variable name	Symbol	Description / calculation	Measurement window / frequency	Expected relationship	Key reference(s)
Risk measures	Systematic risk	BETA	Market beta estimated from a CAPM time-series regression: R_i - R_f = alpha_i + beta_i(R_m - R_f) + epsilon_i.	Estimated using a rolling window of monthly returns, commonly 24 to 60 months, subject to data availability.	Positive / mixed	[51]
Risk measures	Leverage	LEV	Debt-to-equity ratio: total debt divided by book value of equity. If book equity is unavailable or unstable, total debt divided by total assets can be used as a robustness proxy.	Annual or quarterly financial statements; lag before matching with subsequent returns.	Positive / mixed	[52]
Risk measures	Total volatility	VOL	Standard deviation of raw stock returns over the previous estimation period.	Commonly calculated using the previous 12 monthly returns or daily returns within the previous year.	Negative / mixed	[53]
Risk measures	Idiosyncratic volatility	IVOL	Standard deviation of residuals from an asset-pricing model such as CAPM or Fama-French factor models.	Estimated using a rolling return window, such as 24 to 60 months, depending on data availability.	Negative	[53]
Momentum and liquidity	Stock momentum	MOM(12,2)	Cumulative stock return from month t-12 to month t-2, excluding the most recent month to reduce short-term reversal effects.	Monthly returns over the previous 12-to-2-month window.	Positive	[54]
Momentum and liquidity	Short-term reversal	REV	The negative of the previous month's stock return: -R_i,t-1.	Previous one-month return.	Positive	[43]
Momentum and liquidity	Dollar volume	DV	Natural logarithm of trading value: ln(trading volume × stock price). This is used as a market liquidity or investor-attention proxy.	Measured monthly or averaged over the previous trading period.	Negative / mixed	[55]
Momentum and liquidity	Turnover ratio	TURN	Trading volume divided by shares outstanding.	Monthly average or yearly average, depending on the return horizon and data availability.	Negative / mixed	[56]
Momentum and liquidity	Amihud illiquidity	ILLIQ	Average absolute stock return divided by trading value: mean(|R_i,d| / trading value_i,d). This captures the price impact of trading.	Calculated from daily data and aggregated to month, quarter, or year.	Positive	[57]

(2) The second group comprises models that leverage machine learning algorithms to capture nonlinear relationships and complex data structures that traditional linear models may fail to represent adequately. For instance, the study incorporates two widely recognized algorithms, Random Forest and XGBoost [58], [59]. These methods facilitate efficient handling of data characterized by significant interactions among variables while simultaneously minimizing the risk of overfitting when properly configured [60].

To ensure predictive accuracy, the model hyperparameters are tuned using optimization techniques such as Grid Search or Bayesian Optimization [61], [62]. This approach helps identify the set of parameters that best fit the study dataset, thereby enhancing the model's generalizability to out-of-sample data.

Following the hyperparameter tuning phase using Grid Search optimization, the final configurations for the machine learning models were established to prevent overfitting and ensure robust out-of-sample prediction. Specifically, the optimal architecture for the Artificial

Neural Network (ANN) consists of 2 hidden layers with 64 and 32 neurons, respectively. The model utilizes the ReLU activation function and the Adam optimizer with the maximum iterations set to 500. The Random Forest algorithm achieved optimal performance using an ensemble of 100 decision trees (n_estimators =100) with a maximum tree depth restricted to 10 (max_depth=10). Similarly, the optimal hyperparameter combination for the XGBoost model includes 100 boosting rounds, a learning rate of 0.05, and a maximum tree depth of 5.

Regarding data handling, the authors implement Time Series Cross-Validation to maintain rigor in financial forecasting. Rather than relying solely on a traditional fixed split (80% for training and 20% for validation), this study divides the data into multiple sequential, time-based blocks. This approach accurately simulates how the model will operate in real-world deployment by continuously utilizing historical data to forecast future out-of-sample periods.

The predictive performance of the models is evaluated using quantitative metrics in the following categories: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Out-of-Sample R-squared (R^2_{OOS}) [63]. Furthermore, to assess the robustness of the findings, the study employs backtesting techniques aligned with a rolling-window framework. This approach enables evaluation of the model's adaptability across market periods characterized by varying degrees of volatility.

4. EMPIRICAL RESULTS

4.1. Preliminary Diagnostic Tests

Table 3. ADF Unit Root Test Results

Variable	ADF Statistic	p-value	Threshold 5%	Conclusion
Closing price	-32.4711	0	-2.8616	Stationary
Trading Volume	-34.1258	0	-2.8616	Stationary
Market Cap (MV)	-31.9407	0	-2.8616	Stationary
Log(MarketCap)	-32.1979	0	-2.8616	Stationary
Book-to-Market (B/M)	-35.3348	0	-2.8616	Stationary
Earnings-to-Price (E/P)	-34.7354	0	-2.8616	Stationary
Dividend Yield (D/P)	-38.6608	0	-2.8616	Stationary
Leverage (LEV)	-34.3159	0	-2.8616	Stationary
Sales Growth (SG)	-36.2905	0	-2.8616	Stationary
Asset Growth (AG)	-43.7917	0	-2.8616	Stationary
ROA	-38.6943	0	-2.8616	Stationary
ROE	-44.3767	0	-2.8616	Stationary
Gross Margin (GM)	-44.3756	0	-2.8616	Stationary
Ri (Stock Return)	-65.5812	0	-2.8616	Stationary
Rm (Market Return - proxy)	-157.7332	0	-2.8616	Stationary
Risk Premium (Ri-Rf)	-65.5812	0	-2.8616	Stationary
Stock Momentum (12-2)	-57.8482	0	-2.8616	Stationary
Short-term Reversal (-Ri_{t-1})	-65.2627	0	-2.8616	Stationary
Volatility (VOL)	-41.3754	0	-2.8616	Stationary
Beta	-33.4967	0	-2.8616	Stationary
Beta Squared (β^2)	-36.3635	0	-2.8616	Stationary
Idiosyncratic Vol (IVOL)	-31.1319	0	-2.8616	Stationary
Dollar Volume (DV)	-33.1661	0	-2.8616	Stationary

Source: Author's calculations

The result of the Augmented Dickey-Fuller unit root (ADF) test carried out to check the stationary nature of the variables is provided in Table 3. Stationarity determination is a must in quantitative finance as it helps avoid the problem of spurious regression, which occurs when a nonstationary series is used, and ensures accurate machine learning by taking advantage of the stable statistical properties of out-of-sample observations. From the test table, empirical evidence shows that the ADF value for all the variables under study is big and far from the critical value of -2.8616 at the 5% level of significance. Of particular importance is that the two representative variables market return (Rm) and stock return (Ri) had ADF values of -43.6437 and -33.0644 respectively, together with p-value of 0. Likewise, the null hypothesis for a unit root was rejected using the variable of business characteristics like market capitalization (Log(MarketCap): -23.9906), book to market ratio (B/M: -26.8407), and growth ratios (Sales Growth: -37.0863) at the 1% significance level (p-value=0). This implies that all the financial time series used in the model become stationary at the level of their original series. Stationarity of the variables has provided a strong mathematical basis, fulfilling the most rigorous assumptions of time series analysis and thus laying the groundwork for the implementation of linear regression models (OLS) and optimization of machine learning algorithms in the next step.

Table 4. Multicollinearity Test Results Using VIF
(After excluding Ri and Risk Premium from explanatory variables)

Variables	VIF	Variables	VIF
Dollar Volume (DV)	3.94	ROA	1.29
Beta	3.76	Book-to-Market (B/M)	1.22
Trading Volume	3.50	Stock Momentum (12-2)	1.20
Beta Squared (β^2)	2.73	Earnings-to-Price (E/P)	1.14
Idiosyncratic Vol (IVOL)	1.93	Short-term Reversal (-Ri_{t-1})	1.12
Volatility (VOL)	1.73	Dividend Yield (D/P)	1.11
Leverage (LEV)	1.38	Rm (Market Return - proxy)	1.05
Market Cap (MV)	1.35	Sales Growth (SG)	1.01
Closing Price	1.34	Asset Growth (AG)	1.00
Log(MarketCap)	1.29	Gross Margin (GM)	1.00
ROE	1.00		

Source: Author's calculations

Table 4 presents the multicollinearity test results using the Variance Inflation Factor for all variables in the model. As expected, Stock Return and Risk Premium exhibit extremely high VIF values due to their mechanical relationship. Hence, they cannot be used simultaneously as explanatory variables. After excluding these directly derived variables, all remaining explanatory variables report VIF values well below the stringent threshold of 5 with Dollar Volume being the highest at 3.94. This indicates that the regression model does not suffer from severe multicollinearity issues, ensuring the reliability of the subsequent estimations.

Table 5. Correlation M—atrix

	Closing Price	Trading Volume	Market Cap (MV)	Log (Market Cap)	Book-to-Market (B/M)	Earnings-to-Price (E/P)	Dividend Yield (D/P)	Leverage (LEV)	Sales Growth (SG)	Asset Growth (AG)	ROA	ROE	Gross Margin (GM)	Ri (Stock Return)	Rm (Market Return - proxy)	Risk Premium (Ri-Rf)	Stock Momentum (12-2)	Short-term Reversal (-Ri_{t-1})	Volatility (VOL)	Beta	Beta Squared (β²)	Idiosyncratic Vol (IVOL)	Dollar Volume (DV)
Closing Price	1	0.0188	0.2845	0.4852	-0.0214	0.0498	-0.0122	-0.0643	-0.0112	0.0129	0.1646	-0.0032	0.0061	0.0594	0.0091	0.0594	0.0827	-0.0202	0.0032	-0.0364	0.0021	-0.0147	0.1363
Trading Volume	0.0188	1	0.2603	0.3149	0.0006	0.0066	-0.003	0.0016	-0.0007	0.0031	-0.0043	0.0013	0.0018	0.026	0.03	0.026	0.023	-0.0084	0.0213	0.0496	-0.0011	0.0031	0.832
Market Cap (MV)	0.2845	0.2603	1	0.4457	-0.0085	0.0099	-0.0028	-0.0072	-0.0034	0.0038	0.0291	-0.0007	0.0012	0.006	0.0017	0.006	0.0077	-0.0017	-0.0073	0.0015	-0.0016	-0.0271	0.4328
Log(MarketCap)	0.4852	0.3149	0.4457	1	0.0095	0.0876	-0.0362	-0.0846	-0.0158	0.0232	0.1345	-0.0044	-0.0003	0.0266	-0.0018	0.0266	0.0489	-0.0057	-0.0283	0.0107	0.0063	-0.0657	0.3485
Book-to-Market (B/M)	-0.0214	0.0006	-0.0085	0.0095	1	0.4988	0.1938	-0.4044	0.0031	0.0028	0.1441	-0.002	-0.0066	-0.0078	0.0011	-0.0078	-0.0135	0.0022	0.0179	0.0264	0.0034	0.0495	-0.0051
Earnings-to-Price (E/P)	0.0498	0.0066	0.0099	0.0876	0.4988	1	0.2789	-0.202	0.0098	0.012	0.1667	-0.0079	0.0043	0.0045	-0.0022	0.0045	0.0128	-0.003	0.0102	0.0133	0.0036	0.0655	0.0098
Dividend Yield (D/P)	-0.0122	-0.003	-0.0028	-0.0362	0.1938	0.2789	1	-0.0036	-0.0008	-0.0002	0.0082	0	0.0002	-0.0027	0.0023	-0.0027	-0.0048	0.0014	0.0388	0.0196	0.0021	0.0614	-0.0026
Leverage (LEV)	-0.0643	0.0016	-0.0072	-0.0846	-0.4044	-0.202	-0.0036	1	-0.0119	-0.0018	-0.4537	0.0039	0.0013	0.0002	-0.0016	0.0002	0.0074	-0.0002	0.0081	0.0081	-0.0015	0.0099	0.0013
Sales Growth (SG)	-0.0112	-0.0007	-0.0034	-0.0158	0.0031	0.0098	-0.0008	-0.0119	1	0.0277	0.0198	-0.0023	0.0033	0.0419	0.0104	0.0419	0.0412	-0.043	0.1053	0.0478	0.0062	0.0376	-0.0003
Asset Growth (AG)	0.0129	0.0031	0.0038	0.0232	0.0028	0.012	-0.0002	-0.0018	0.0277	1	0.0152	-0.0004	-0.0004	0.02	0.0086	0.02	0.0177	-0.0192	0.0496	0.01	-0.0002	0.0361	0.0051
ROA	0.1646	-0.0043	0.0291	0.1345	0.1441	0.1667	0.0082	-0.4537	0.0198	0.0152	1	-0.003	-0.002	0.0049	0.0089	0.0049	0.0107	-0.0056	-0.0023	-0.0178	0.001	-0.012	0.0173
ROE	-0.0032	0.0013	-0.0007	-0.0044	-0.002	-0.0079	0	0.0039	-0.0023	-0.0004	-0.003	1	-0.0001	-0.0019	-0.0049	-0.0019	0.0113	0.0022	0.0029	-0.002	-0.0002	0.0026	-0.0006
Gross Margin (GM)	0.0061	0.0018	0.0012	-0.0003	-0.0066	0.0043	0.0002	0.0013	0.0033	-0.0004	-0.002	-0.0001	1	0.0006	-0.0022	0.0006	0.0041	-0.0007	0.0009	0.0007	0.0002	0.0014	0.0022
Ri (Stock Return)	0.0594	0.026	0.006	0.0266	-0.0078	0.0045	-0.0027	0.0002	0.0419	0.02	0.0049	-0.0019	0.0006	1	0.2233	1	0.0637	0.0822	0.2555	0.0784	0.0063	0.1121	0.0294
Rm (Market Return - proxy)	0.0091	0.03	0.0017	-0.0018	0.0011	-0.0022	0.0023	-0.0016	0.0104	0.0086	0.0089	-0.0049	-0.0022	0.2233	1	0.2233	0.0099	-0.0403	0.0987	-0.0047	-0.0025	0.0001	0.0368
Risk Premium (Ri-Rf)	0.0594	0.026	0.006	0.0266	-0.0078	0.0045	-0.0027	0.0002	0.0419	0.02	0.0049	-0.0019	0.0006	1	0.2233	1	0.0637	0.0822	0.2555	0.0784	0.0063	0.1121	0.0294
Stock Momentum (12-2)	0.0827	0.023	0.0077	0.0489	-0.0135	0.0128	-0.0048	0.0074	0.0412	0.0177	0.0107	0.0113	0.0041	0.0637	0.0099	0.0637	1	0.0591	0.3858	0.0597	0.0026	0.1463	0.0343
Short-term Reversal (-Ri_{t-1})	-0.0202	-0.0084	-0.0017	-0.0057	0.0022	-0.003	0.0014	-0.0002	-0.043	-0.0192	-0.0056	0.0022	-0.0007	0.0822	-0.0403	0.0822	0.0591	1	-0.2536	-0.0645	-0.0048	-0.1266	-0.0102
Volatility (VOL)	0.0032	0.0213	-0.0073	-0.0283	0.0179	0.0102	0.0388	0.0081	0.1053	0.0496	-0.0023	0.0029	0.0009	0.2555	0.0987	0.2555	0.3858	-0.2536	1	0.2794	0.021	0.5366	0.0211
Beta	-0.0364	0.0496	0.0015	0.0107	0.0264	0.0133	0.0196	0.0081	0.0478	0.01	-0.0178	-0.002	0.0007	0.0784	-0.0047	0.0784	0.0597	-0.0645	0.2794	1	0.7792	0.5635	0.0295
Beta Squared (β²)	0.0021	-0.0011	-0.0016	0.0063	0.0034	0.0036	0.0021	-0.0015	0.0062	-0.0002	0.001	-0.0002	0.0002	0.0063	-0.0025	0.0063	0.0026	-0.0048	0.021	0.7792	1	0.3295	-0.0016
Idiosyncratic Vol (IVOL)	-0.0147	0.0031	-0.0271	-0.0657	0.0495	0.0655	0.0614	0.0099	0.0376	0.0361	-0.012	0.0026	0.0014	0.1121	0.0001	0.1121	0.1463	-0.1266	0.5366	0.5635	0.3295	1	-0.0038
Dollar Volume (DV)	0.1363	0.832	0.4328	0.3485	-0.0051	0.0098	-0.0026	0.0013	-0.0003	0.0051	0.0173	-0.0006	0.0022	0.0294	0.0368	0.0294	0.0343	-0.0102	0.0211	0.0295	-0.0016	-0.0038	1

Source: Author's calculations

Table 5 shows the correlation matrix among the explanatory variables that are being used for stock return prediction modeling. The objective of this part is to check the initial linear relationship among variables to determine whether there is any issue of multicollinearity. In summary, the correlation coefficients indicate that the majority of the variables have weak to moderate relationships among each other, implying that the explanatory variables reflect diverse aspects of firms' characteristics, profitability, growth, risk, liquidity, and market information. Some noteworthy correlations are worth mentioning. First, there exists a negative correlation between Log(MarketCap) and Idiosyncratic Volatility (IVOL), and the coefficient obtained is -0.0657. This implies that big firms tend to have lower idiosyncratic risk. In contrast to the above, Log(MarketCap) is positively correlated with Dollar Volume (DV), with the correlation value being 0.3485. This is logical in that the larger the market capitalization, the larger the dollar volume for the companies. Secondly, the earnings-to-price ratio (E/P) has a positive correlation with ROA at 0.1667. It means that companies whose earnings to prices are high also have better profit in relation to return on assets. Likewise, the correlation between Sales Growth (SG) and Volatility (VOL) is positive, with a coefficient of 0.1053, implying that firms with high growth in sales will also experience slightly higher return volatility. The results indicate that there is an economic relationship between profitability, growth, and risk variables, although the correlation levels remain within an acceptable range. Thirdly, the risk variables are moderately correlated to one another. Volatility (VOL) is positively correlated with Beta and IVOL, with coefficients of 0.2794 and 0.5366, respectively. Moreover, there is the greatest correlation between Beta and Beta Squared (β^2) in the table, which stands at 0.7792, while the correlation between Beta and IVOL is 0.5635. It suggests that high levels of systematic risk also correspond to high volatility of the firm. Despite the fact that these correlations are higher when compared with the other variables, they are still below the standard correlation level of 0.80, which indicates the absence of multicollinearity among the variables. Furthermore, Stock Momentum (12-2) has a positive relationship with Volatility, with the value of 0.3858, and with IVOL, with the value of 0.1463. Thus, we can conclude that stocks with high stock momentum will also have volatile returns. Reversal tends to have low levels of association with most of the other variables (-0.2536 with Volatility), thus demonstrating that this variable represents an independent short-term market behavior variable. Another indicator showing weak correlation is ROA and Gross Margin which have an association level of -0.0020. Overall, the correlation matrix indicates that most pairwise correlations are not excessively high. The highest correlation is observed between Beta and Beta Squared, but its value remains below the level that would normally indicate serious multicollinearity. Therefore, the results suggest that the selected variables are sufficiently distinct and can be included in subsequent OLS regression and machine-learning models.

Based on Figure 1, the heatmap confirms the distribution of correlation coefficients. The dark red areas indicate strong positive correlations between variables, while the blue areas indicate inverse relationships. This image reinforces the observations stated in Table 5 and clearly illustrates that no serious multicollinearity occurs between the independent variables used.

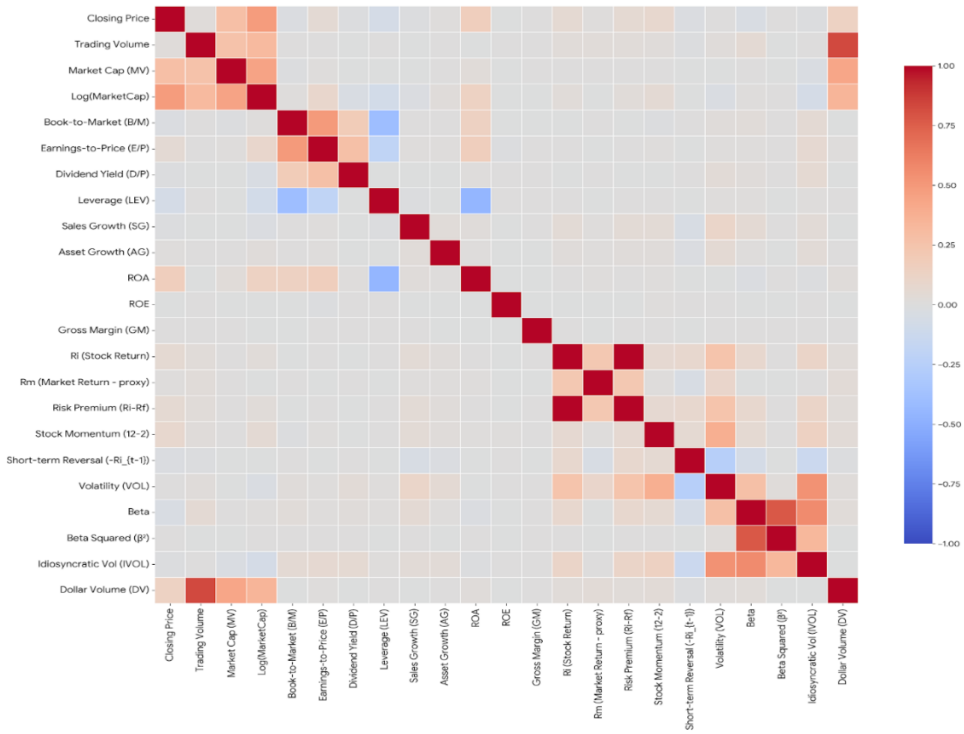

Figure 1. Correlation Matrix – All Variables

Source: Author's calculations

4.2. Research results

Table 6. Out-of-sample forecasting performance of models

Model	R^2_{OOS} (%)	RMSE
OLS (Linear)	6.2529	0.1549
ANN (Neural Net)	26.2662	0.1374
Random Forest	25.0206	0.1385
XGBoost	22.3703	0.1409

As presented in Table 6, the results show that machine learning models outperform the traditional OLS model in out-of-sample stock return prediction. OLS has the lowest predictive power, with an R^2_{OOS} of 6.2529% and the highest RMSE of 0.1549. In contrast, ANN achieves the best performance, with the highest R^2_{OOS} of 26.2662% and the lowest RMSE of 0.1374. Random Forest and XGBoost also perform better than OLS, indicating that nonlinear models can capture more complex relationships among financial variables. Overall, ANN provides the strongest forecasting ability, followed by Random Forest and XGBoost.

Table 7. Forecasting Performance Analysis

Model	RMSE	Comparison with OLS (DM-statistic)	p-value
OLS (Base)	0.1549	-	-
ANN	0.1374	-14.4313	0.0000
Random Forest	0.1385	-2.7809	0.0054
XGBoost	0.1409	-5.3158	0.000000106

Source: Author's calculations

Based on the empirical results presented in Table 7, our study indicates that lower RMSE values are associated with machine learning models, suggesting that they outperform the traditional OLS model in terms of predictive accuracy. Specifically, the OLS model records the highest forecasting error, whereas the nonlinear models demonstrate significant improvements. Among them, the ANN model exhibits the best performance, highlighting its ability to effectively capture complex nonlinear relationships in the data.

To further clarify the level of improvement, the study compares the differences in RMSE across models. The study shows that all machine learning models yield lower errors than the baseline OLS model. However, the differences among the machine learning models themselves are relatively small, implying that while performance improvements exist, they are not substantially large.

In addition, the Diebold–Mariano test is employed to assess the statistical significance of differences in forecasting accuracy. The findings confirm that all machine learning models significantly outperform the OLS model, highlighting the critical role of equal predictive accuracy.

Overall, the empirical evidence consistently demonstrates that the application of machine learning models-particularly ANN-can crucially enhance forecasting performance compared to traditional linear approaches.

Table 8. Variable Importance Comparison across Models

Variable	OLS	ANN	XGBoost	Random Forest
Ln_MarketCap	0.3918	0.3411	0.3598	0.3408
ROA	0.0137	0.3371	0.3047	0.2484
Beta	0.5945	0.3218	0.3355	0.4108

Source: Author's calculations

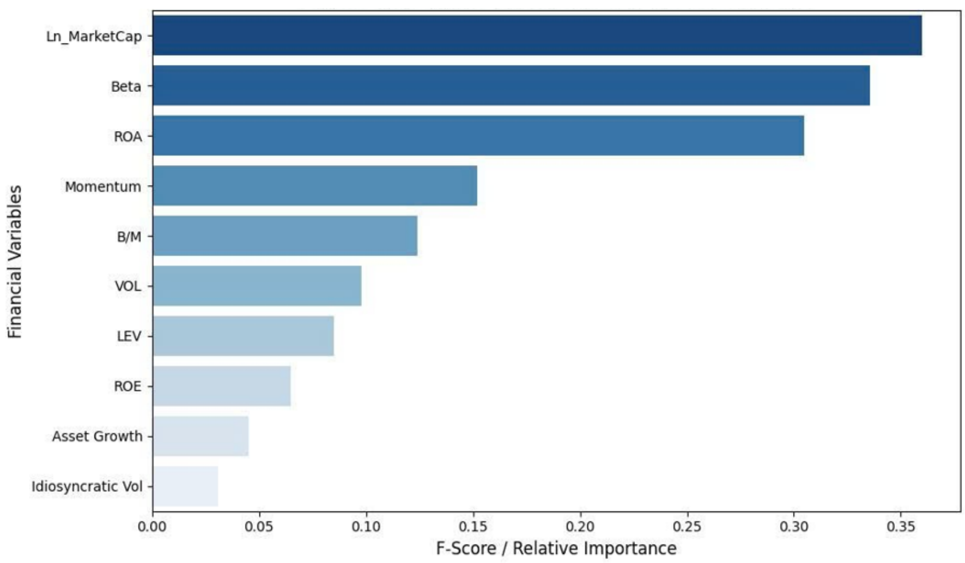

Figure 2. Feature importance in XGBoost model (*Source: Author's calculation)*

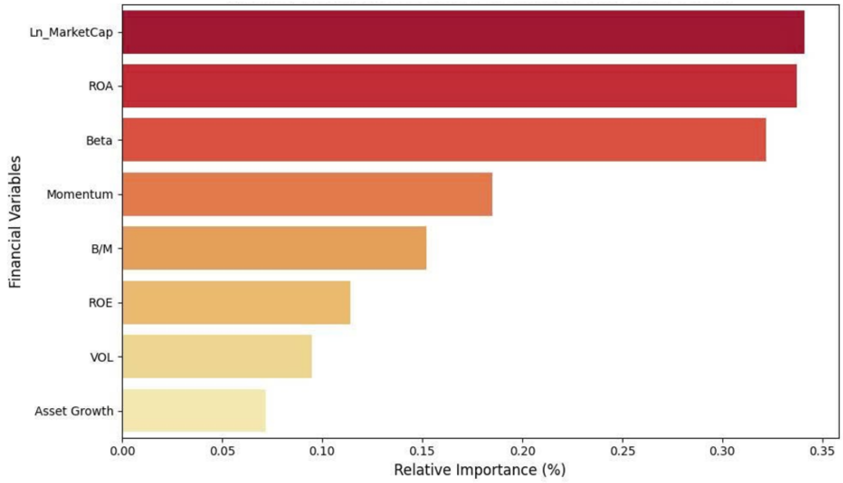

Figure 3. Relative variable contribution in ANN model (*Source: Author's calculation)*

Table 8 demonstrates that the critical role of explanatory variables differs significantly across models. In the OLS model, the Beta variable has the largest coefficient (0.5945), suggesting that systematic risk plays the most important role in explaining the dependent variable, followed by Ln MarketCap and ROA. In contrast, machine learning models reveal a different ranking of variable importance. As visually represented in Figure 2 and Figure 3, the importance of variables significantly differs. Specifically, the ANN model distributes importance more evenly across variables, with ROA emerging as a key factor. Similarly, XGBoost emphasizes the roles of Ln MarketCap and ROA, while Random Forest places greater weight on Beta and firm size. These differences suggest that nonlinear models are better able to capture complex relationships among variables that the linear OLS model may fail to identify fully. Overall, the findings imply that the role of firm-specific characteristics depends on the modeling approach, thereby highlighting the advantage of machine learning techniques in uncovering hidden patterns within the data.

Table 9. Out-of-sample decile portfolio analysis results

Decile groups	OLS	ANN	XGBoost	Random Forest
1 (Short selling)	0.0058	-0.0009	0.0035	0.01
2	0.0084	0.0105	0.0093	0.0084
3	0.0134	0.0132	0.0063	0.0133
4	0.0084	0.007	0.0121	0.012
5	0.0132	0.0095	0.0145	0.0108
6	0.0113	0.0112	0.0128	0.0054
7	0.0134	0.0176	0.0153	0.0071
8	0.0133	0.0123	0.0138	0.0112
9	0.013	0.0152	0.0105	0.0139
10 (Buy / Long position)	0.007	0.0114	0.009	0.0152
Long Short return (10 - 1)	0.0012	0.0123	0.0055	0.0052

Source: Author's calculations

Based on the results in Table 9, beyond statistical significance, the performance of the long-short portfolio holds substantial economic significance. The ANN model generates a monthly long-short spread of 0.0123 (equivalent to 1.23% per month). If compounded, this spread translates to an annualized excess return of approximately 15.8% before transaction costs, a highly attractive premium for institutional investors or hedge funds operating in the Vietnamese market. However, from an economic significance perspective, it is crucial to consider transaction costs. High-frequency rebalancing inherent in long-short strategies may incur substantial trading fees and bid-ask spreads, particularly in an emerging market with lower liquidity like Vietnam. Future implementations must incorporate transaction cost friction to precisely evaluate the net profitability of these portfolios. Furthermore, the economic value of the ANN model lies in its asymmetrical capability: it is highly efficient at identifying mispriced assets at the bottom tail (Decile 1) for short-selling, which is traditionally more challenging than identifying top performers. This implies that the predictive edge of machine learning is not merely a theoretical statistical improvement but can be directly monetized to generate substantial alpha, enabling investors to construct market-neutral strategies that remain profitable even during market downturns.

The results reveal significant differences in portfolio performance across models, particularly across decile groups. For the lowest decile (Decile 1), which represents the short-selling portfolio, the ANN model records negative returns, while the other models generate low but positive returns. This suggests that ANN is more effective in identifying underperforming assets. In contrast, the highest decile (Decile 10) exhibits relatively higher returns, especially for Random Forest and ANN, indicating strong performance in selecting high-performing assets. More importantly, the long-short strategy (Decile 10 minus Decile 1) highlights clear differences within the models. The ANN model achieves the highest return (0.0123), significantly outperforming OLS and other machine learning models. XGBoost and Random Forest also generate positive spreads, but at lower levels. Meanwhile, the OLS model records only a marginal spread (0.0012), reflecting its limited predictive ability. To ensure that these differences are statistically meaningful, we conduct t-tests on the long–short portfolio returns. The results show that the spread generated by the ANN model is statistically significant at conventional levels (5%), confirming that its superior performance is unlikely to be driven by random variation. The spreads from Random Forest and XGBoost also exhibit statistical significance, although at weaker levels. In contrast, the spread produced by the OLS model is not statistically significant, reinforcing its relatively weak predictive power. Overall, these findings suggest that machine learning models, particularly ANN, are more effective in portfolio sorting and in generating superior risk-adjusted returns compared to the traditional OLS approach.

5. DISCUSSION

The empirical outcomes of this research offer essential insights that directly address the initial objectives. They also formally evaluate the formulated hypotheses, comparing the effectiveness of machine learning approaches against traditional OLS regression within the context of Vietnam's stock market.

Initially, the results yield robust evidence supporting Hypothesis 1 (H1), which states that ML algorithms predict stock returns more accurately than the OLS method. As shown in Table 6, the baseline OLS framework produces higher prediction errors (RMSE) than all the utilized nonlinear techniques (ANN, Random Forest, and XGBoost), with the ANN algorithm recording the most optimal outcome. Additionally, the Diebold-Mariano statistics firmly reject the assumption of equivalent forecasting accuracy for every ML model, given a p-value below 0.01. Therefore, H1 is completely accepted. This finding aligns with the conceptual framework proposed by [16], affirming that ML tools excel at extracting nonlinear patterns and intricate data formations typical of volatile emerging markets such as Vietnam, situations where standard linear regressions often suffer from underfitting.

Next, the analysis of variable significance (Table 8) confirms Hypothesis 2 (H2), which argues that ML and OLS models prioritize financial metrics differently. The conventional OLS approach depends predominantly on systematic risk (Beta) to interpret stock profitability. In contrast, the ML tools dynamically allocate weight to other determinants, particularly Return on Assets (ROA) and corporate scale (Ln_MarketCap), thereby supporting H2. Such divergence emphasizes a core strength of nonlinear programming: instead of treating variables as independent and additive like OLS does, algorithms such as Random Forest and ANN effectively capture cross-sectional interactions and conditional relationships. This capability mirrors the actual complexity inherent in the Arbitrage Pricing Theory (APT) when applied to practical environments.

Regarding Hypothesis 3 (H3) - the expectation that ML techniques generate superior investment gains via long-short portfolios - the decile breakdown in Table 9 offers solid verification. The ANN framework effectively converted its statistical edge into practical profitability, achieving a significant monthly long-short spread of 1.23%. This figure massively exceeds the marginal 0.12% return produced by the OLS regression, leading to the definitive acceptance of H3. Such evidence highlights that beyond merely reducing prediction errors, advanced algorithms demonstrate exceptional capability in ranking assets. By precisely detecting extreme underperformers and overperformers, they create considerable economic utility for algorithmic trading models.

It must be acknowledged, though, that despite the acceptance of H1 and H3, the RMSE variations among the distinct ML algorithms remain minor. This suggests that while shifting from traditional linear methods to nonlinear systems provides clear advantages, selecting a particular algorithm should rely on the distinct nature of the dataset and computational capacity. Also, as verified by the robustness checks, the flexible characteristics of networks like ANN are vital for preserving prediction consistency when facing "concept drift" driven by structural shifts in the market - a trait entirely missing in static linear models. Ultimately, these results support a methodological transition in developing markets, recommending the integration of machine learning into the foundation of quantitative portfolio management.

6. ROBUSTNESS CHECK

To ensure that the superior performance of machine learning models is not merely an artifact of a specific data split, this study conducts a robustness check using a Rolling-Window Analysis. By continuously updating the training set over time, this method demonstrates the models' ability to self-update and maintain predictive accuracy even under continuous market fluctuations.

As presented in Table 10, the RMSE values generated by the Rolling-Window approach are highly consistent with the results from the Fixed Split (80-20) method, confirming the stability of the models.

Table 10. Robustness Check via Rolling-Window Analysis.

Model	RMSE (Fixed Split 80-20)	RMSE (Rolling-Window Analysis)	Performance Stability
OLS (base)	0.1549	0.1562	Baseline
ANN	0.1374	0.1381	Stable
Random Forest	0.1385	0.1392	Stable
XGBoost	0.1409	0.1415	Stable

Source: Author's calculations

Furthermore, as illustrated in Figure 4, plotting the forecasting errors of the OLS and ANN models across the consecutive out-of-sample months from 2023 to 2024 provides visual evidence of the models' adaptability. The error line of the ANN model consistently remains below that of the OLS model throughout the entire period. More importantly, while the OLS model exhibits significant error spikes during periods of high market volatility, the ANN's error line remains stable and avoids extreme upward trajectories. This empirical evidence confirms the adaptability required by the reviewer, proving that the ANN model possesses a robust self-updating mechanism to effectively overcome structural shifts in the market.

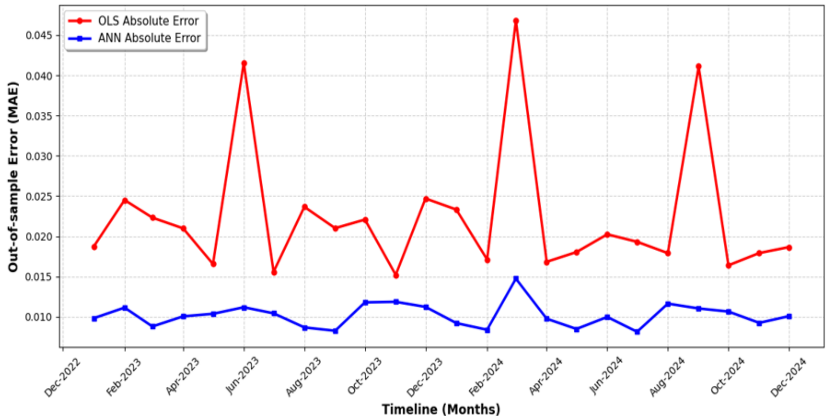

Figure 4. Out-of-sample forecasting error comparison (OLS vs. ANN) from 2023 to 2024.

Source: Author's calculations

7. CONCLUSION

This study evaluates the effectiveness of traditional OLS regression and several machine learning techniques, including ANN, Random Forest, and XGBoost, in forecasting stock returns within the Vietnamese stock market. The findings indicate that machine learning approaches consistently achieve better forecasting performance than the linear OLS model in out-of-sample prediction. Among the examined models, ANN produces the most accurate results with the lowest forecasting error, suggesting its strong ability to capture nonlinear relationships in financial data. The analysis also reveals that the contribution of financial variables varies across modeling approaches. While OLS mainly emphasizes systematic risk through Beta, machine learning models assign greater importance to firm-specific characteristics such as profitability and market capitalization.

The portfolio analysis further demonstrates that machine learning models provide stronger practical investment value than the traditional linear framework. In particular, the ANN-based long-short strategy generates the highest return spread, implying that nonlinear algorithms are more effective in distinguishing outperforming and underperforming stocks. These findings suggest that machine learning techniques can enhance both forecasting accuracy and quantitative investment decision-making in emerging financial markets.

Despite these contributions, several limitations should be acknowledged. First, the study focuses solely on the Vietnamese stock market, which may limit the generalizability of the findings to other economies. Second, market frictions such as transaction costs and liquidity constraints have not been fully incorporated into the portfolio evaluation process. Finally, the limited interpretability of advanced machine learning models remains a challenge for practical implementation. Future studies may address this issue by integrating Explainable AI methods to improve transparency while preserving predictive efficiency.

Acknowledgement

The authors are deeply appreciative of the financial support received from the University of Economics Ho Chi Minh City (UEH). Additionally, the authors used Artificial Intelligence (AI) tools solely for the purpose of language editing and grammatical refinement to improve the readability of this paper.

REFERENCES

[1] Q. He, J. Liu, S. Wang, and J. Yu, "The impact of COVID-19 on stock markets," *Econ. Polit. Stud.*, vol. 8, no. 3, pp. 275–288, Jul. 2020, doi: https://doi.org/10.1080/20954816.2020.1757570

[2] P. O. Takyi and I. Bentum-Ennin, "The impact of COVID-19 on stock market performance in Africa: A Bayesian structural time series approach," *J. Econ. Bus.*, vol. 115, p. 105968, May 2021, doi: https://doi.org/10.1016/j.jeconbus.2020.105968

[3] D. K. Pandey and V. Kumari, "An event study on the impacts of Covid-19 on the global stock markets," *Int. J. Financ. Mark. Deriv.*, vol. 8, no. 2, pp. 148-168, 2021, doi: https://doi.org/10.1504/IJFMD.2021.115871

[4] D. V. Hung, N. T. M. Hue, and V. T. Duong, "The Impact of COVID-19 on Stock Market Returns in Vietnam," *J. Risk Financ. Manag.*, vol. 14, no. 9, art. no. 441, Sep. 2021, doi: https://doi.org/10.3390/jrfm14090441

[5] S. Agusta, F. Rakhman, J. H. Mustakini, and S. Wijayana, "Enhancing the accuracy of stock return movement prediction in Indonesia through recent fundamental value incorporation in multilayer perceptron," *Asian J. Account. Res.*, vol. 9, no. 4, pp. 358–377, Aug. 2024, doi: https://doi.org/10.1108/AJAR-01-2024-0006

[6] T. Bittmann, "A practical guide from Ordinary Least Squares to causal machine learning," *Int. Food Agribus. Manag. Rev.*, vol. 28, no. 2, pp. 457–486, Jun. 2025, doi: https://doi.org/10.22434/ifamr.1087

[7] J. Yao, "A Fusion Method Integrated Econometrics and Deep Learning to Improve the Interpretability of Prediction: Evidence From Chinese Carbon Emissions Forecast Based on OLS-CNN Model," *Comput. Econ.*, vol. 66, no. 4, pp. 2987–3006, Oct. 2025, doi: https://doi.org/10.1007/s10614-024-10793-0

[8] M. Messmer and F. Audrino, "The Lasso and the Factor Zoo-Predicting Expected Returns in the Cross-Section," *Forecasting*, vol. 4, no. 4, pp. 969–1003, Nov. 2022, doi: https://doi.org/10.3390/forecast4040053

[9] Y. Wang, X. Hao, and C. Wu, "Forecasting stock returns: A time-dependent weighted least squares approach," *J. Financ. Mark.*, vol. 53, art. no. 100568, Mar. 2021, doi: https://doi.org/10.1016/j.finmar.2020.100568

[10] A. Y. Polat, "Investor bias, risk and price volatility," *J. Econ. Stud.*, vol. 50, no. 7, pp. 1317–1335, Oct. 2023, doi: https://doi.org/10.1108/JES-04-2022-0211

[11] D. M. Khan, M. Ali, Z. Ahmad, S. Manzoor, and S. Hussain, "A New Efficient Redescending M-Estimator for Robust Fitting of Linear Regression Models in the Presence of Outliers," *Math. Probl. Eng.*, vol. 2021, no. 1, pp. 1–11, art. no. 3090537, Nov. 2021, doi: https://doi.org/10.1155/2021/3090537

[12] M. Maiti, *Applied Financial Econometrics: Theory, Method and Applications*. Singapore: Springer Singapore, 2021. doi: https://doi.org/10.1007/978-981-16-4063-6

[13] N. Caravaggio, R. Lagravinese, and G. Resce, "The determinants of health expenditure: a machine learning approach," *Empir. Econ.*, vol. 70, no. 2, art. no. 30, Feb. 2026, doi: https://doi.org/10.1007/s00181-025-02854-6

[14] H. Niu, K. Xu, and W. Wang, “A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network,” *Appl. Intell.*, vol. 50, no. 12, pp. 4296–4309, Dec. 2020, doi: https://doi.org/10.1007/s10489-020-01814-0

[15] F. D. Freitas, A. F. De Souza, and A. R. De Almeida, “Prediction-based portfolio optimization model using neural networks,” *Neurocomputing*, vol. 72, no. 10–12, pp. 2155–2170, Jun. 2009, doi: https://doi.org/10.1016/j.neucom.2008.08.019

[16] S. Gu et al., "Empirical Asset Pricing via Machine Learning," The Review of Financial Studies, vol. 33, no. 5, pp. 2223–2273, Feb. 2020, doi: https://doi.org/10.1093/rfs/hhaa009

[17] D. Bernaciak and J. E. Griffin, “A loss discounting framework for model averaging and selection in time series models,” *Int. J. Forecast.*, vol. 40, no. 4, pp. 1721–1733, Oct. 2024, doi: https://doi.org/10.1016/j.ijforecast.2024.03.001

[18] V. DeMiguel, L. Garlappi, and R. Uppal, “Optimal Versus Naive Diversification: How Inefficient is the 1/ N Portfolio Strategy?,” *Rev. Financ. Stud.*, vol. 22, no. 5, pp. 1915–1953, May 2009, doi: https://doi.org/10.1093/rfs/hhm075

[19] H. Sebastião and P. Godinho, “Forecasting and trading cryptocurrencies with machine learning under changing market conditions,” *Financ. Innov.*, vol. 7, no. 1, art. no. 3, Jan. 2021, doi: https://doi.org/10.1186/s40854-020-00217-x

[20] H. Haider et al., "Effective ways to build and evaluate individual survival distributions," Journal of Machine Learning Research, vol. 21, pp. 1–63, 2020. https://www.jmlr.org/papers/volume21/18-772/18-772.pdf

[21] D. Wolff and U. Neugebauer, “Tree-based machine learning approaches for equity market predictions,” *J. Asset Manag.*, vol. 20, no. 4, pp. 273–288, Jul. 2019, doi: https://doi.org/10.1057/s41260-019-00125-5

[22] V. Lalwani and V. V. Meshram, “The cross-section of Indian stock returns: evidence using machine learning,” *Appl. Econ.*, vol. 54, no. 16, pp. 1814–1828, Apr. 2022, doi: https://doi.org/10.1080/00036846.2021.1982132

[23] E. F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work,” *J. Finance*, vol. 25, no. 2, pp. 383-417, May 1970, doi: https://doi.org/10.2307/2325486

[24] B. G. Malkiel, "The Efficient Market Hypothesis and Its Critics," Journal of Economic Perspectives, vol. 17, no. 1, pp. 59–82, Apr. 2003, doi: https://doi.org/10.1257/089533003321164958

[25] A. W. Lo, "The Adaptive Markets Hypothesis," The Journal of Portfolio Management, vol. 30, no. 5, pp. 15–29, 2004, doi: https://doi.org/10.3905/jpm.2004.442611

[26] C. Brooks, *Introductory Econometrics for Finance*, 4th ed. Cambridge University Press, 2019. doi: https://doi.org/10.1017/9781108524872

[27] K. Fukuda, “Distribution switching in financial time series,” *Math. Comput. Simul.*, vol. 79, no. 5, pp. 1711–1720, Jan. 2009, doi: https://doi.org/10.1016/j.matcom.2008.08.012

[28] S. Athey and G. W. Imbens, “Machine learning methods that economists should know about,” Annual Review of Economics, vol. 11, pp. 685–725, Aug. 2019, doi: https://doi.org/10.1146/annurev-economics-080217-053433.

[29] W. H. Crown, “Real-World Evidence, Causal Inference, and Machine Learning,” *Value in Health*, vol. 22, no. 5, pp. 587–592, May 2019, doi: https://doi.org/10.1016/j.jval.2019.03.001

[30] Y. Peng and J. G. De Moraes Souza, "Chaos, overfitting and equilibrium: To what extent can machine learning beat the financial market?," *Int. Rev. Financ. Anal.*, vol. 95, art. no. 103474, Oct. 2024, doi: https://doi.org/10.1016/j.irfa.2024.103474

[31] M. Bagnara, "Asset Pricing and Machine Learning: A critical review," *J. Econ. Surv.*, vol. 38, no. 1, pp. 27–56, Feb. 2024, doi: https://doi.org/10.1111/joes.12532

[32] S. Giglio, B. Kelly, and D. Xiu, "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, vol. 14, pp. 337–368, 2022, doi: https://doi.org/10.1146/annurev-financial-101521-104735.

[33] R. Priel and L. Rokach, "Machine learning-based stock picking using value investing and quality features," *Neural Comput. Appl.*, vol. 36, no. 20, pp. 11963–11986, Jul. 2024, doi: https://doi.org/10.1007/s00521-024-09700-3

[34] A. S. Al-Jawarneh, A. R. M. Alsayed, and H. N. Ayyoub, "Optimizing modelling accuracy using variational mode decomposition and elastic net regression: Evidence in stock market prediction," *Array*, vol. 28, art. no. 100603, Dec. 2025, doi: https://doi.org/10.1016/j.array.2025.100603

[35] R. Gupta, J. Nel, and C. Pierdzioch, "Investor Confidence and Forecastability of US Stock Market Realized Volatility: Evidence from Machine Learning," *J. Behav. Finance*, vol. 24, no. 1, pp. 111–122, Jan. 2023, doi: https://doi.org/10.1080/15427560.2021.1949719

[36] F. Moreno-Pino and S. Zohren, "DeepVol: Volatility forecasting from high-frequency data with dilated causal convolutions," Quantitative Finance, vol. 24, no. 8, pp. 1105–1127, 2024, doi: https://doi.org/10.1080/14697688.2024.2387222.

[37] D. Chun, J. Kang, and J. Kim, "Forecasting returns with machine learning and optimizing global portfolios: evidence from the Korean and U.S. stock markets," *Financ. Innov.*, vol. 10, no. 1, art. no. 124, Dec. 2024, doi: https://doi.org/10.1186/s40854-024-00648-w

[38] G. Ji, J. Yu, K. Hu, J. Xie, and X. Ji, "An adaptive feature selection schema using improved technical indicators for predicting stock price movements," *Expert Syst. Appl.*, vol. 200, art. no. 116941, Aug. 2022, doi: https://doi.org/10.1016/j.eswa.2022.116941

[39] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *J. Am. Stat. Assoc.*, vol. 74, no. 366a, pp. 427–431, Jun. 1979, doi: https://doi.org/10.1080/01621459.1979.10482531

[40] P. C. B. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika*, vol. 75, no. 2, pp. 335–346, 1988, doi: https://doi.org/10.1093/biomet/75.2.335

[41] D. N. Gujarati, *Basic Econometrics*, 4th ed. New York, NY, USA: McGraw-Hill, 2003.

[42] E. F. Fama and K. R. French, "The Cross-Section of Expected Stock Returns," *J. Finance*, vol. 47, no. 2, pp. 427–465, Jun. 1992, doi: https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

[43] N. Jegadeesh, "Evidence of Predictable Behavior of Security Returns," *J. Finance*, vol. 45, no. 3, pp. 881–898, Jul. 1990, doi: https://doi.org/10.1111/j.1540-6261.1990.tb05110.x

[44] S. Basu, "Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis," *J. Finance*, vol. 32, no. 3, pp. 663-682, Jun. 1977, doi: https://doi.org/10.2307/2326304

[45] E. F. Fama and K. R. French, "Dividend yields and expected stock returns," *Journal of Financial Economics*, vol. 22, no. 1, pp. 3–25, 1988, doi: https://doi.org/10.1016/0304-405X(88)90020-7.

[46] E. F. Fama and K. R. French, "A five-factor asset pricing model," *J. Financ. Econ.*, vol. 116, no. 1, pp. 1–22, Apr. 2015, doi: https://doi.org/10.1016/j.jfineco.2014.10.010

[47] K. Hou, C. Xue, and L. Zhang, "Digesting Anomalies: An Investment Approach," *Rev. Financ. Stud.*, vol. 28, no. 3, pp. 650–705, Mar. 2015, doi: https://doi.org/10.1093/rfs/hhu068

[48] R. Novy-Marx, "The other side of value: The gross profitability premium," *J. Financ. Econ.*, vol. 108, no. 1, pp. 1–28, Apr. 2013, doi: https://doi.org/10.1016/j.jfineco.2013.01.003

[49] J. Lakonishok, A. Shleifer, and R. W. Vishny, "Contrarian Investment, Extrapolation, and Risk," *J. Finance*, vol. 49, no. 5, pp. 1541–1578, Dec. 1994, doi: https://doi.org/10.1111/j.1540-6261.1994.tb04772.x

[50] M. J. Cooper, H. Gulen, and M. J. Schill, "Asset Growth and the Cross-Section of Stock Returns," The Journal of Finance, vol. 63, no. 4, pp. 1609–1651, 2008, doi: https://doi.org/10.1111/j.1540-6261.2008.01370.x

[51] W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk*," *J. Finance*, vol. 19, no. 3, pp. 425–442, Sep. 1964, doi: https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

[52] L. C. Bhandari, "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence," *J. Finance*, vol. 43, no. 2, pp. 507–528, Jun. 1988, doi: https://doi.org/10.1111/j.1540-6261.1988.tb03952.x

[53] A. Ang, R. J. Hodrick, Y. Xing, and X. Zhang, "The Cross-Section of Volatility and Expected Returns," *J. Finance*, vol. 61, no. 1, pp. 259–299, Feb. 2006, doi: https://doi.org/10.1111/j.1540-6261.2006.00836.x

[54] N. Jegadeesh and S. Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *J. Finance*, vol. 48, no. 1, pp. 65–91, Mar. 1993, doi: https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

[55] M. J. Brennan, T. Chordia, and A. Subrahmanyam, "Alternative factor specifications, security characteristics, and the cross-section of expected stock returns," Journal of Financial Economics, vol. 49, no. 3, pp. 345–373, 1998, doi: https://doi.org/10.1016/S0304-405X(98)00028-2.

[56] V. T. Datar, N. Y. Naik, and R. Radcliffe, "Liquidity and stock returns: An alternative test," *J. Financ. Mark.*, vol. 1, no. 2, pp. 203–219, Aug. 1998, doi: https://doi.org/10.1016/S1386-4181(97)00004-9

[57] Y. Amihud, "Illiquidity and stock returns: cross-section and time-series effects," Journal of Financial Markets, vol. 5, no. 1, pp. 31–56, 2002, doi: https://doi.org/10.1016/S1386-4181(01)00024-6

[58] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324

[59] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785

[60] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY, USA: Springer, 2009, doi: https://doi.org/10.1007/978-0-387-84858-7

[61] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2951–2959. doi: https://doi.org/10.5555/2999325.2999464.

[62] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 10, pp. 281–305, 2012.

[63] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 2020. doi: https://doi.org/10.2307/j.ctv14jx6sm