OPTIMIZED GA-LIGHTGBM MODEL FOR FORECASTING STOCK RETURN VOLATILITY OF SAO MAI GROUP (ASM.VN) IN THE SUSTAINABLE ENERGY SECTOR

Tran Ba Thuan[1*]

[1] Faculty of Economic Information Systems, Hue University of Economics, Hue University,
99 Ho Dac Di Street, Hue City, Vietnam

[*]Corresponding author: tbthuan@hce.edu.vn

Abstract

This study proposes a Genetic Algorithm–optimized LightGBM (GA – LightGBM) model for forecasting stock volatility of ASM.VN, demonstrating superior performance compared with traditional models (GARCH, SVMRegressor, XGBoost) and deep learning architectures (GRU, TabNet, Transformer, CNN, LSTM). The model achieves MAE = 0.0470; RMSE = 0.0731; R^2 = 0.8480; STD = 0.1480, with a mean R^2 = 0.805 across 5-fold cross-validation, indicating high accuracy and stability. Predicted volatility over 06 February 2025 – 06 March 2025 ranges from 0.1419 to 0.2154, reflecting moderate adjustment risk. SHAP analysis identifies key original features in descending order: HV, BB – width, BB – lower, VWAP, ATR, MA_21, BB – upper, OBV, highlighting the roles of market fluctuations, price amplitude, liquidity, and medium-term trends in shaping forecasts. Sensitivity analysis indicates colsample – bytree and num – leaves exert the strongest influence (ΔR^2 = 2.02% and 0.81%, respectively), whereas other hyperparameters produce minor variations (<0.6%), confirming model robustness. During the period starting 09 April 2025, US retaliatory tariffs (+46%) triggered strong volatility spikes in energy stocks, particularly ASM.VN (+130%), HDG.VN (+93%), BCG.VN (+60%), with differentiated forecasts persisting through September 2025. The GA – LightGBM model demonstrates strong potential for guiding investment strategies, risk management, and policy-sensitive market analysis. While empirical results are robust, economic implications emphasize practical benefits including risk-adjusted returns, improved hedging efficiency, and reduced pricing errors in derivative markets. Caution is warranted under "black swan" events or structural market changes, necessitating periodic retraining and hyperparameter monitoring to maintain predictive reliability.

Keywords: Genetic Algorithm; HOSE; LightGBM; Machine learning; Retaliatory tariffs; Volatility forecasting; Sustainable energy; SHAP.

JEL codes: C53, G17, Q42, F13.

1. INTRODUCTION

In the context of globalization and the rapid fluctuations of financial markets, forecasting risk and stock return volatility has become an essential requirement to support investors in optimizing strategies and managing portfolios (Zhuang et al., 2023). Modern markets are simultaneously influenced by macroeconomic, geopolitical, and psychological behavioral factors, which increase nonlinearity and complexity in forecasting. Recent studies have increasingly applied advanced machine learning models such as the Light Gradient Boosting Machine (LightGBM) to capture nonlinear data structures, thanks to the advantages of leaf-wise growth and histogram binning techniques (Ke et al., 2017). However, the effectiveness of LightGBM still largely depends on hyperparameter optimization, which is often conducted manually or through local optimization methods, potentially leading to overfitting or unstable predictive performance.

This study integrates LightGBM with the Genetic Algorithm (GA), a global optimization method inspired by natural evolution (Holland, 1992), to automatically search for the optimal set of hyperparameters. In addition, Principal Component Analysis (PCA) is applied to reduce data dimensionality, while the SHAP approach is employed to interpret the contribution of input variables to predictive outcomes. The model is validated on a group of renewable energy stocks in Vietnam, including GEG.VN, HDG.VN, ASM.VN, BCG.VN, PC1, and TV2.VN, which represent the trend of green and sustainable investments. The research pursues four main objectives: (1) optimizing LightGBM through GA to enhance forecasting accuracy; (2) applying PCA to improve data quality and mitigate the risk of biased learning; (3) employing SHAP to identify key variables driving stock volatility; and (4) analyzing the impact of macroeconomic policy factors on the returns of the sustainable energy stock portfolio.

2. LITERATURE REVIEW

2.1. Related Research

Recent studies have emphasized the role of machine learning and artificial intelligence in forecasting financial markets and optimizing renewable energy systems. Ukoba et al. (2024) proposed integrating deep learning to improve the performance of wind, solar, and photovoltaic systems, while Lu et al. (2024) examined the relationship between energy consumption and stock market volatility. In the financial domain, Muminov et al. (2023) applied blockchain and social media signals to forecast Bitcoin volatility using Q-learning, whereas Kristjanpoller (2024) combined machine learning and econometric models to predict natural gas prices, demonstrating the broad applicability of advanced approaches. Traditional GARCH models have been employed by Al Najjar (2016) to analyze volatility clustering and asymmetry in emerging markets, while Li (2022) compared LSTM, LightGBM, and CNN, identifying LightGBM as superior in short-term forecasting. Accurate volatility forecasting in emerging and energy markets has garnered increasing attention. Alfeus et al. (2025) demonstrate that HAR, GARCH, RECH, and RFSV capture distinct features of realised volatility in South Africa, highlighting market roughness. Kristjanpoller (2024) shows that combining econometric and machine learning models improves natural gas volatility prediction, with exchange rates as key drivers. Fan et al. (2024) introduce EWT-S-ALOSVR for new energy stocks, revealing chaotic dynamics and nonlinear volatility. Lyócsa and Todorova (2024) apply DCCE-HAR, Ridge, and Random Forest, evidencing technology and oil sector influences on clean energy volatility.

In Vietnam, Duong (2023) applied XGBoost to forecast the VNIndex with an accuracy rate exceeding 80%. More recently, Tran Ba Thuan and Tran Thi Dieu Thuan (2025) developed deep learning models such as the Adaptive Laguerre ANN optimized with Bayesian Optimization for ICL stock, and a GARCH-NN model for First Solar stock, proving their potential to enhance predictive performance and trading strategies. Globally, Huang et al. (2025), using Chinese A-share market data (2006–2021), demonstrate that GA-LightGBM constructs networks that better capture market volatility and serves as an effective tool for forecasting firm performance, linked to financial risk and information opacity. Yang (2024) applies GA-LightGBM within a stacked ensemble to predict Bitcoin trends, capturing cryptocurrency market volatility more effectively than buy-and-hold strategies. GA optimizes hyperparameters and mitigates overfitting, while LightGBM efficiently handles nonlinear and complex time series; combined, GA-LightGBM outperforms XGBoost and CatBoost in forecasting. However, GA-LightGBM has not been applied to Vietnam's renewable energy market, presenting a significant **research gap**. Studies on sustainable energy stocks in emerging markets like Vietnam remain limited, while policy volatility, Net Zero 2050 commitments, and geopolitical events such as tariff effects strongly influence return fluctuations. The **novelty** of this study lies in applying GA-LightGBM to process nonlinear data and optimize hyperparameters, while capturing policy specifics, green transition trends, and geopolitical impacts in stock return analysis. Additionally, integrating PCA for dimensionality reduction and SHAP for model interpretability introduces methodological innovation, enhancing predictive performance and transparency. The **main contribution** is the development of a GA-LightGBM hybrid model to forecast sustainable energy stock returns, supporting investors in optimizing risk management, improving strategic effectiveness, and providing empirical evidence for sustainable development.

2.2. Research Methodology

2.2.1. Return volatility

Return volatility (Historical return volatility-HV) reflects the degree of fluctuation in stock prices over time, calculated based on logarithmic returns (log returns). Logarithmic return is defined as: $r_t = \ln\left(P_t / P_{t-1}\right)$, where P_t and P_{t-1} represent the adjusted closing prices at time t and $t-1$, respectively. This formulation eliminates percentage scale effects, producing a continuous and statistically tractable return series (Jorion, 2010). After constructing the log-return series, historical volatility is estimated as the standard deviation of log returns over a 21-day rolling window:

$$HV = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(r_{t-i} - \bar{r})^2}$$

where N denotes the window length and $\bar{r}$ is the mean log return within the same window. To annualize volatility, the result is multiplied by the square root of 252 (the average number of trading days in a year): $HV_t = HV.\sqrt{252}$ (Jorion, 2010). This methodology is widely applied in market risk measurement, portfolio optimization, and the development of advanced financial forecasting models (Jorion, 2010).

2.2.2. LightGBM model

The predicted return volatility $\hat{HV}_t$ (Predicted Volatility) is subsequently estimated using **LightGBM** based on an input vector X_t, which includes the scaled GARCH volatility estimate (Volatility_GARCH_Scaled = $\hat{\sigma}_t$), historical volatility HV_t and selected technical indicators. The predictive relationship is formalized as: $\hat{HV}_t = f\left(X_t;\theta\right)$ where: X_t = { HV_t, technical indicators}, $f(\bullet)$ is a nonlinear function approximated by the LightGBM model, and θ denotes the hyperparameters optimized via a Genetic Algorithm (GA). The predicted volatility $\hat{HV}_t$ is then benchmarked against realized volatility HV_t to evaluate forecasting performance.

LightGBM (Light Gradient Boosting Machine), a state-of-the-art machine learning algorithm from the Gradient Boosted Decision Trees (GBDT) family, employs a leaf-wise tree growth strategy rather than the conventional level-wise approach, thereby enhancing accuracy and computational efficiency (Ke et al., 2017). At iteration m, the prediction is expressed as:

$$\hat{HV}_t^{\,m}(X_t) = \hat{HV}_t^{\,m-1}(X_t) + \eta.h_m(X_t)$$

which $\hat{HV}_t^{m}(X_t)$ is the model after m iterations, η is the learning rate, and $h_m(X_t)$ is the newly added decision tree. The objective function is defined as:

$$Obj = \sum_{t=1}^{T} L\left(HV_t, \hat{HV}_t\right) + \sum_{k=1}^{K} \Omega(h_k)$$

where $L\left(HV_t, \hat{HV}_t\right)$ denotes the loss function, and $\Omega(h_k)$ is a regularization term controlling model complexity to mitigate overfitting. LightGBM further accelerates training by employing histogram-based binning to discretize continuous variables, thereby reducing computation without sacrificing accuracy (An et al., 2023; Zhuang et al., 2023).

2.2.3. Genetic Algorithm (GA)

The Genetic Algorithm (GA) is employed to optimize the hyperparameters of the LightGBM model, leveraging the principles of natural evolution (Holland, 1992; Yang, 2014). The optimization process consists of several key steps: initializing a population of individuals, where each individual represents a random combination of LightGBM parameters; evaluating the fitness of each individual based on the forecasting loss function; selecting and performing crossover among high-performing individuals to generate the next generation; and applying random mutation to maintain genetic diversity and prevent premature convergence to local optima. This evolutionary process is repeated iteratively until convergence is reached at an optimal parameter set θ^*. With this optimal configuration, the volatility forecasting function can be expressed as: $\hat{HV}_t = f\left(X_t; \theta^*\right)$. The integration of GA enhances predictive accuracy, improves generalization capability, and reduces the risk of convergence at suboptimal local solutions (Mirjalili, 2019; Xue et al., 2021).

2.2.4. Data

Stock data for Sao Mai Group (ASM.VN) were collected from Yahoo Finance (Yahoo Finance, 2025), using the Adjusted Close price to calculate logarithmic returns (log returns) and historical return volatility - two fundamental measures in risk management and price forecasting (Hull, 2023). The dataset comprises 23 financial features designed to enhance both predictive accuracy and interpretability of the model.

Traditional technical indicators such as MA_21, RSI, MACD, ATR, and Bollinger Bands (BB_upper, BB_lower, BB_width), as well as Momentum, VWAP, CCI, Stochastic (%K, %D), OBV, and MFI, reflect market trends, oscillations, and money flows (Murphy, 1999). Trading volume measures the level of market attention, while cyclical features such as Day_of_Week, Month_of_Year, Fourier_Week, and Fourier_Month allow the model to capture seasonal patterns (Hyndman et al., 2018). In addition, statistical indicators including historical volatility (HV), skewness, kurtosis, and trend slope provide valuable insights into the distributional properties and directional movements of price series (Tsay, 2010).

The data were split into an 80%-20% ratio, with 80% used for training the model and 20% reserved for testing, enabling the evaluation of generalization performance while mitigating overfitting. This split ratio is consistent with common practices in machine learning when the dataset is sufficiently large (Géron, 2019; Raschka et al., 2019).

2.3. Comprehensive model evaluation

The LightGBM model, optimized via Genetic Algorithm (GA), is evaluated using key predictive metrics: MAE, MSE, RMSE, MAPE, R^2, and the standard deviation of errors (STD), ensuring both accuracy and stability (Hyndman & Koehler, 2006). MAE quantifies the average absolute deviation between observed and predicted values, while MSE penalizes larger deviations through squaring. RMSE, the square root of MSE, retains the original data's units, enhancing interpretability. MAPE expresses errors as percentages, facilitating comparison across time series. R^2 indicates the proportion of variance explained by the model, serving as a standard regression performance measure (PeerJ, 2023). Finally, STD captures error stability by reflecting fluctuations across forecast iterations, providing insight into model reliability.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|HV_t - \hat{HV}_t\right|, MSE = \frac{1}{T}\sum_{t=1}^{T}\left(HV_t - \hat{HV}_t\right)^2, MAPE = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{HV_t - \hat{HV}_t}{HV_t}\right|$$

$$RMSE = \sqrt{MSE}\ ,\ R^2 = 1 - \frac{\sum_{t=1}^{T}\left(HV_t - \hat{HV}_t\right)^2}{\sum_{t=1}^{T}\left(HV_t - \overline{HV}\right)^2}\ and\ STD = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left((HV_t - \hat{HV}_t) - \overline{(HV_t - \hat{HV}_t)}\right)^2}.$$

3. EMPIRICAL STUDY

3.1. Data Visualization

The historical volatility (HV) data for the period 23 February 2010 – 05 February 2025 reflect the long-term risk characteristics of the asset, with daily observation frequency, allowing the assessment of both average volatility levels and the temporal distribution of HV. The maximum HV reached 1.4737 on 27 January 2011, indicating a period of unusually high market volatility, typically associated with abrupt price adjustments or information shocks. In contrast, the minimum HV of 0.0754 on 20 February 2017 reflects a stable market environment with low risk and significantly reduced prices fluctuations. The mean (0.4098) and median (0.3946) values are relatively close, whereas the standard deviation (0.1760) indicates a moderately wide dispersion of volatility, highlighting variability around the average.

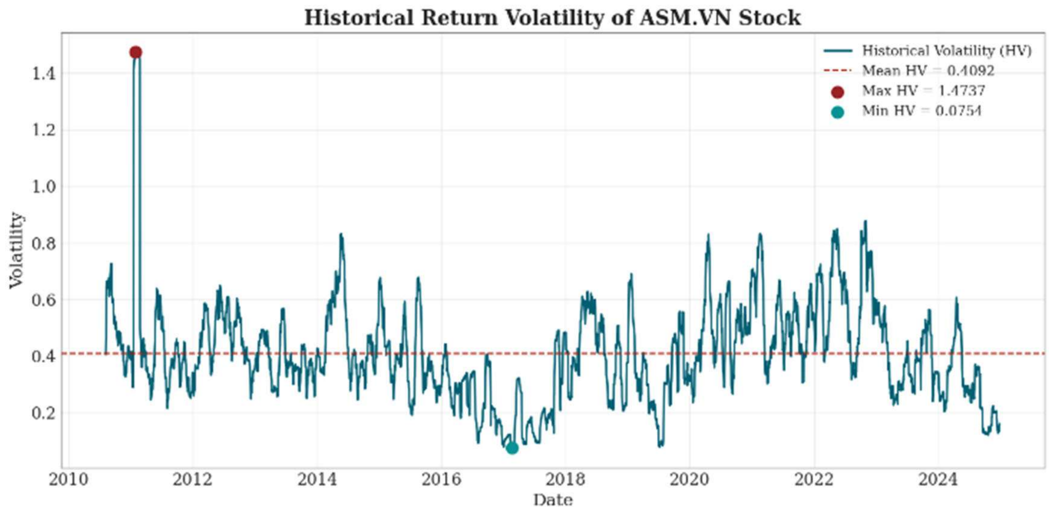

Figure 1. Historical volatility of ASM.VN stock

(Source: Compiled by the author)

A positive skewness of 1.31 and a high kurtosis of approximately 5.9 indicate that the HV distribution is strongly right-skewed with heavy tails, implying the presence of sudden spikes in volatility far exceeding normal levels. Both the Shapiro – Wilk and D'Agostino – Pearson tests yield p-values = 0.0, rejecting the null hypothesis of normality. Therefore, HV exhibits a non-normal, skewed, and heavy-tailed distribution, highlighting the existence of extreme risk events and suggesting the need for nonlinear models or fat-tailed distributions to more accurately capture market volatility.

The dataset comprises 23 technical and statistical features (X), including momentum indicators (RSI, Momentum, CCI, %K, %D), volatility measures (HV, ATR), trend indicators (MA_21, MACD, Trend_Slope), liquidity metrics (Volume, OBV, MFI, VWAP), and advanced features such as Bollinger Bands (BB_upper, BB_lower, BB_width), Fourier_Week, Fourier_Month, as well as distribution moments (Skewness, Kurtosis). Principal Component Analysis (PCA) was applied to reduce noise, eliminate multicollinearity, and extract latent structures from the highly correlated feature set.

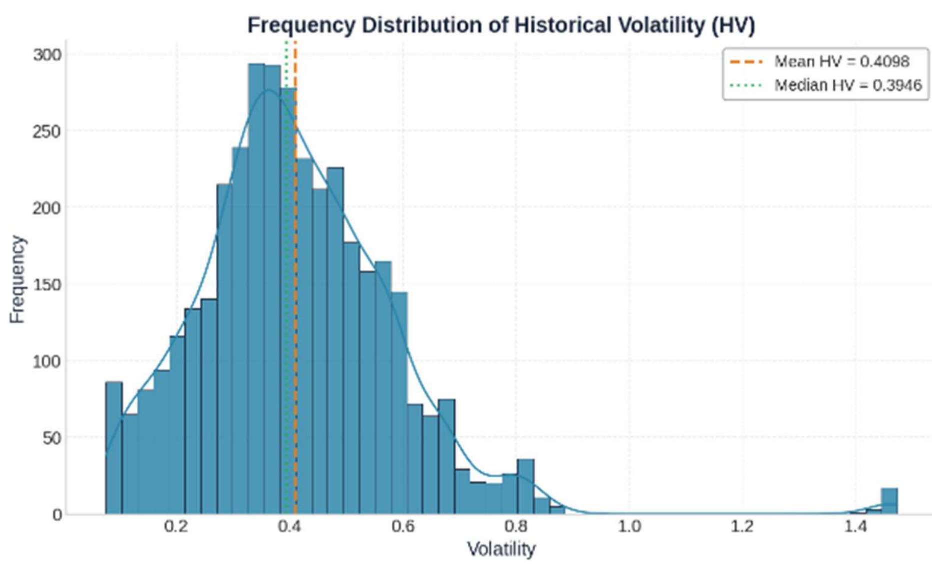

Figure 2. Frequency Distribution of Historical Volatility

(Source. Author's calculations)

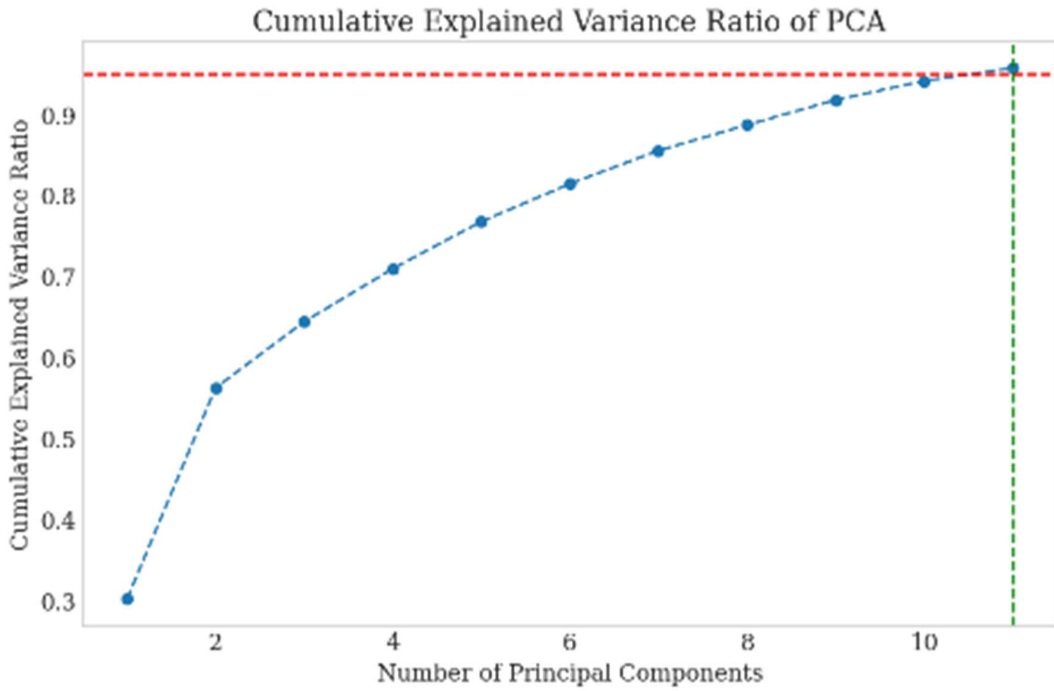

Figure 3. Illustration of the dimensionality reduction process using PCA

(Source: Compiled by the author)

Principal Component Analysis (PCA) was applied to the 23 technical and statistical variables to reduce noise and eliminate redundant information. The first 10 components retained 95% of the variance and were used for model training after normalization with StandardScaler. This dimensionality reduction, consistent with the recommendations of Jolliffe et al. (2016), enhances the training efficiency of LightGBM and mitigates the risk of overfitting.

3.2. Construction of the Optimal LightGBM Model

The LightGBM model was optimized through Genetic Algorithm (GA) with hyperparameters designed to balance accuracy and efficiency. Specifically, colsample_bytree = 0.67 regulates the proportion of features sampled, while learning_rate = 0.053 enables gradual learning to mitigate overfitting.

Table 1. Optimal hyperparameter set of the LightGBM model

Hyperparameters	**Values**
colsample_bytree	0.67
learning_rate	0.053
min_child_samples	17
min_child_weight	0.014
n_estimators	500
num_leaves	100
reg_alpha	0.05
reg_lambda	0.159
subsample	0.953

(Source: Author's calculations)

The parameters min_child_samples = 17 and min_child_weight = 0.014 restrict the minimum number of samples per leaf to control model complexity. The setting n_estimators = 500 determines the number of trees in the ensemble, whereas num_leaves = 100 adjusts the depth and complexity of each tree. To further reduce overfitting, reg_alpha = 0.05 and reg_lambda = 0.159 impose L1 and L2 regularization penalties, respectively. Finally, subsample = 0.953 controls the proportion of samples used in each tree.

3.3. Training and Evaluation of the LightGBM Model

A LightGBM regression model was constructed using optimal hyperparameters derived from the Genetic Algorithm (GA) and subsequently trained on PCA-transformed data. The model generated predictions on the test set, and multiple evaluation metrics were computed to assess accuracy. PCA effectively reduced data dimensionality, while GA optimized hyperparameters, thereby enhancing overall performance. The GA-optimized LightGBM model achieved high predictive accuracy with minimal error. The Mean Absolute Error (MAE = 0.0470) indicates an average deviation of 0.0470 units between predicted and actual values. The Mean Squared Error (MSE = 0.0053) and Root Mean Squared Error (RMSE = 0.0731) reflect a low level of squared prediction error, demonstrating model stability.

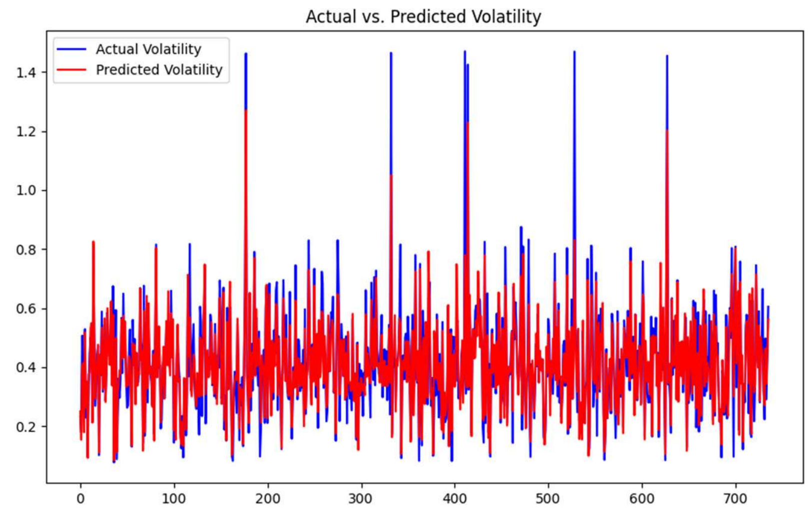

Figure 4. Comparison of Actual and Predicted Volatility

(Source: Author's calculations)

The Mean Absolute Percentage Error (MAPE = 12.71%) aligns with the threshold for good predictive accuracy suggested by Montaño Moreno et al. (2013). Furthermore, the coefficient of determination ($R^2 = 0.8481$) highlights the model's strong explanatory power for actual data, while the standard deviation of prediction errors (STD = 0.1480) illustrates moderate variability, reinforcing the model's consistency and reliability in financial forecasting.

3.4. Application of the Optimized LightGBM Model in Forecasting

Phase 1: The forecasted volatility for the period 06 February 2025 – 06 March 2025 exhibits a narrow range, varying between 0.1419 and 0.2154, reflecting a relatively stable state of the target variable. In the initial phase (06–14 February 2025), a slight downward trend is observed, declining from 0.1895 to approximately 0.1767, indicating that the model predicts a temporary weakening of market momentum. From mid to late February 2025, values fluctuate around 0.16–0.18, representing a sideways movement with low volatility. Notably, local spikes occur on 20 February 2025, 04 March 2025, and particularly on 06 March 2025, reaching 0.2154, the highest value within the forecast horizon. These dynamics suggest a potential market rebound or the emergence of stronger signals in early March 2025. Overall, the forecasted series demonstrates stable behavior without any abnormal shocks.

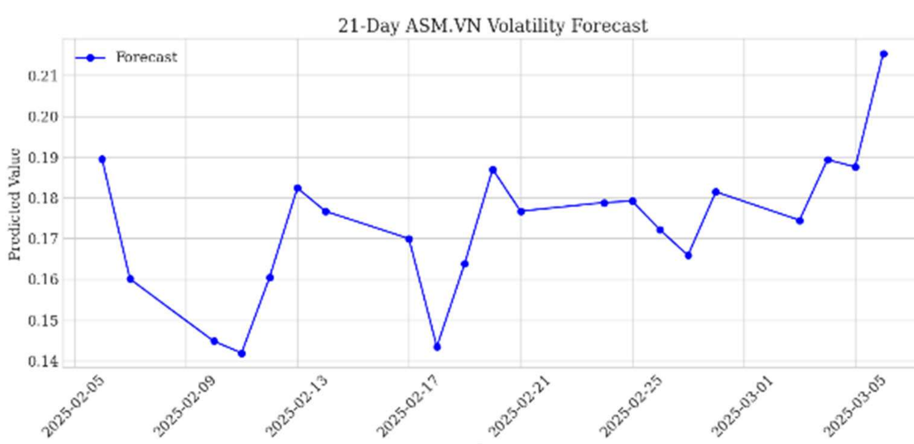

Figure 5. Forecasted Return Volatility of ASM.VN Stock over 21 Days

(Source: Author's calculations)

The SHAP analysis following PCA was conducted to recover the influence of the original features and ensure model interpretability. SHAP values were initially computed in the PCA space and then mapped back to the original feature space via a linear transformation, allowing direct assessment of the original variables rather than relying solely on abstract principal components. The results indicate that HV, BB_width, BB_lower, VWAP, ATR, MA_21, BB_upper, and OBV exert the strongest influence on the forecast. HV has the largest impact, reflecting the role of price volatility in shaping market conditions. BB_width and BB_lower describe price range dynamics and short-term buying–selling pressure. VWAP and OBV represent the driving force of cash flow and liquidity. ATR emphasizes inherent risk, while MA_21 and BB_upper capture trends and resistance levels. Collectively, these features provide a robust foundation for accurate forecasting.

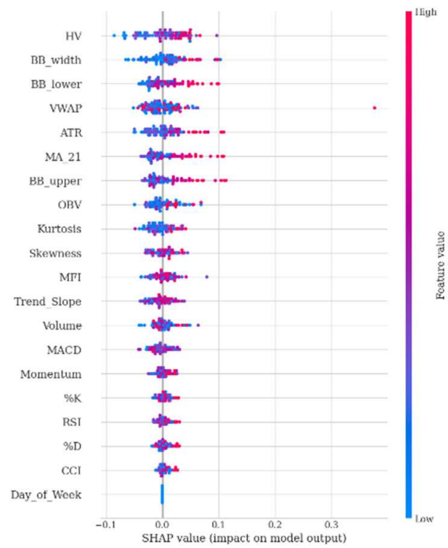

Figure 6. Mapping SHAP values from PCA space to the original features

(Source: Author's calculations)

The SHAP analysis following PCA enables the recovery of the influence of each original feature on the model's predictions through a back-projection from the PCA space to the original feature space. Across the dataset, the resulting normalized importance values are as follows: HV = 0.1549, BB_width = 0.1406, VWAP = 0.1199, ATR = 0.1128, MFI = 0.0917, BB_upper = 0.0642, MA_{21} = 0.0428, %D = 0.0351, %K = 0.0246, CCI = 0.0234, Volume = 0.0199, BB_lower = 0.0130, RSI = 0.0113, while Fourier_Week and Month_of_Year are nearly zero. At the overall level, these weights indicate that the model primarily relies on features representing volatility, price range expansion, and liquidity, which generally act as short-term leading indicators in financial markets. This comprehensive analysis demonstrates that the model integrates information on price fluctuations, market breadth, and cash flow dynamics across the dataset to generate robust predictions, highlighting the systemic role of these key features in shaping forecasted stock volatility.

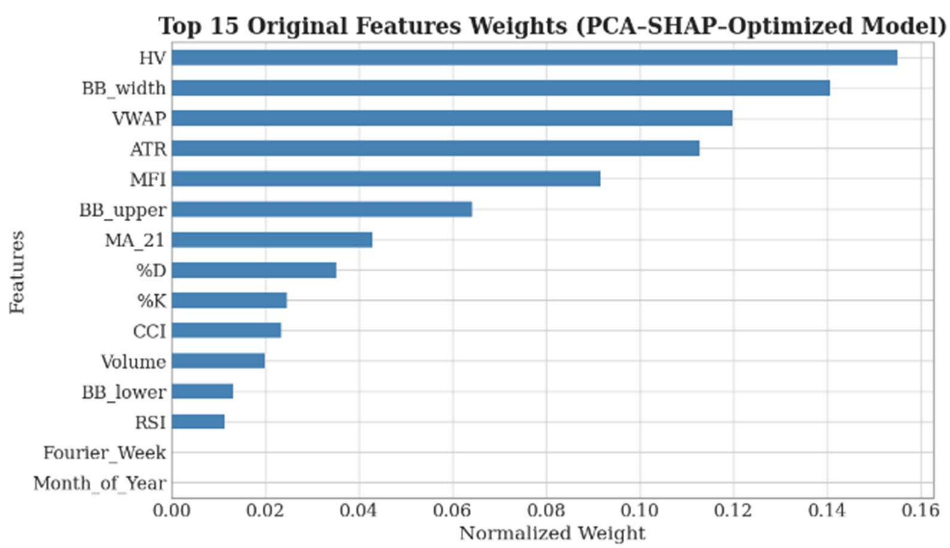

Figure 7. Top 15 Original Features with the Strongest Influence

(Source. Author's calculations)

The trading strategy derived from the model focuses on price volatility and range, reflecting the dominant role of HV and BB_width in signaling trend changes. ATR and MFI further highlight the importance of price momentum and cash flow, while seasonal features such as Fourier and Month_of_Year have minimal impact, implying that the asset is largely insensitive to cyclical effects. Variations in the forecasts primarily stem from sharp changes in HV, the width of the Bollinger Bands, and liquidity fluctuations. When price ranges expand or ATR spikes, the model registers significant adjustments in its predictions.

Phase 2: The optimized LightGBM model forecasts 21-day volatility from April 9, 2025, capturing the significant impact of U.S. announcements regarding a potential 46% retaliatory tariff on Vietnamese goods, later postponed by 30 days. The results reveal sharp short-term fluctuations: GEG.VN rose from 0.498 (April 24) to a peak of 0.572 (April 29), an increase of approximately 14.7%; HDG.VN surged from 0.294 (April 24) to 0.567 (May 1), nearly 93%; ASM.VN climbed from 0.256 to 0.590 (April 28), an increase of about 130% - the highest among the group. Meanwhile, BCG.VN and TV2.VN recorded respective gains of ~60% and ~50%. These findings underscore the heightened short-term risks and uncertainties, highlighting the critical importance of closely monitoring international policy developments to optimize portfolio risk management.

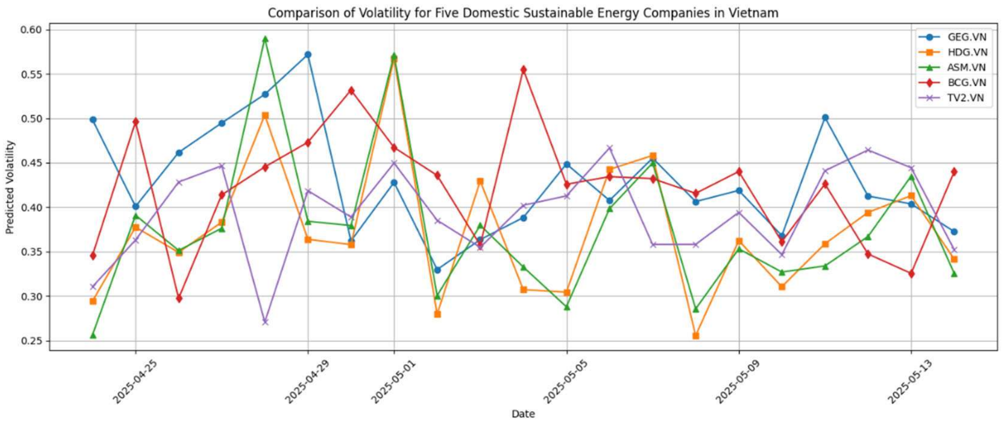

Figure 8. Forecasted Volatility of Five Domestic Sustainable Energy Stocks

(Source: Author’s calculations)

Phase 3: The current tariff rate stands at 20%, with higher levels applied in special cases such as trans-shipment (40%), while certain renewable energy equipment is subject to anti-dumping duties of up to 271.28% (U.S. Department of Commerce, 2024; Financial Times, 2025). During the period following the expiration of the postponement and the reinstatement of the 20% tariff, the volatility of domestic renewable energy stocks listed on HOSE (GEG.VN, HDG.VN, ASM.VN, BCG.VN, TV2.VN) is projected to diverge significantly.

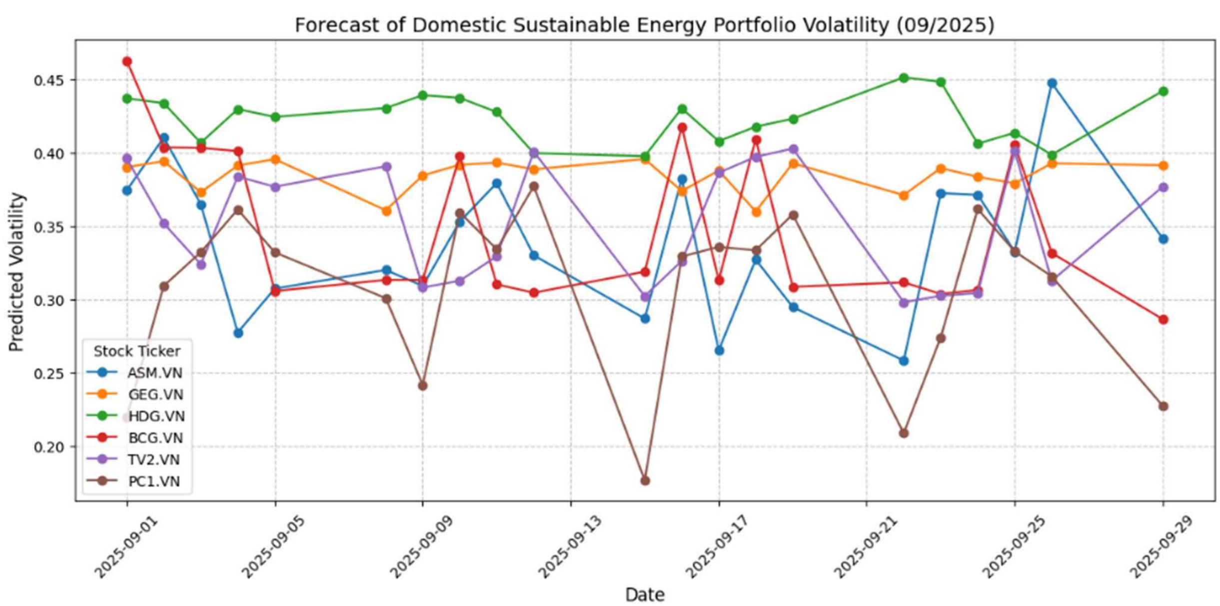

Figure 9. Forecasted Volatility of the Domestic Sustainable Energy Stock Portfolio

(Source: Author’s calculations)

Forecast data reveal a clear differentiation within the group: HDG.VN exhibits the highest volatility (0.40 - 0.45), reflecting greater sensitivity to market fluctuations. BCG.VN and ASM.VN remain in the medium range, while TV2.VN demonstrates comparatively more stability (0.30 - 0.40). GEG.VN records steady volatility around 0.38 - 0.40, suggesting lower risk. Notably, PC1.VN shows the lowest volatility (0.22 - 0.36), indicating limited short-term risk exposure. These variations are largely attributable to differences in financial structure, operational performance, and dependence on imported equipment - factors that are particularly relevant in the context of the reinstated 20% import tariff, which imposes added pressure on firms with higher import reliance.

4. RESULTS AND DISCUSSION

The forecasted volatility of ASM.VN using the GA–LightGBM model indicates a relatively stable sequence over the period 06 February 2025 – 06 March 2025, ranging from 0.1419 to 0.2154, reflecting moderate adjustment risk. The model demonstrates high performance with MAE = 0.0470; RMSE = 0.0731; MAPE = 12.71%; and R^2 = 0.8480, indicating low prediction errors and strong explanatory power. Post-PCA SHAP analysis identifies the most influential original features in descending order: HV, BB_width, BB_lower, VWAP, ATR, MA_21, BB_upper, and OBV, highlighting the importance of volatility, price range, liquidity, and trend in forecasting. The optimal configuration (num_leaves = 100; learning_rate = 0.053; n_estimators = 500) balances complexity and predictive performance, while also informing trading strategies based on short-term volatility and cash flow, whereas seasonal factors such as Fourier or Month_of_Year exert minimal impact.

From 9 April 2025, the imposition of a 46% retaliatory tariff by the United States on Vietnamese exports triggered sharp volatility in renewable energy stocks: ASM.VN surged by 130%, HDG.VN by 93%, BCG.VN by 60%, TV2.VN by 50%, and GEG.VN by 14.7% between 24 April and 14 May 2025. By August 2025, the baseline tariff remained at 20%, with higher rates applied in trans-shipment cases (40%) or anti-dumping duties reaching up to 271.28% (U.S. Department of Commerce, 2024; Financial Times, 2025). In September 2025, the forecasted volatility of renewable energy stocks on HOSE exhibited continued divergence: HDG.VN sustained the highest volatility (0.40–0.45), BCG.VN and ASM.VN remained within a medium range, TV2.VN displayed greater stability (0.30–0.40), GEG.VN fluctuated steadily around 0.38–0.40, while PC1.VN recorded the lowest volatility (0.22–0.36), suggesting limited short-term risk. These differences are largely attributable to variations in financial structure, operational efficiency, and import dependency-factors that exert amplified pressure under the reinstated 20% tariff, particularly on firms with a high reliance on imported equipment.

In this study, LightGBM achieved the lowest error and the highest goodness-of-fit, demonstrating both accurate and stable predictive performance. Compared to other nonlinear models such as XGBoost and SVMRegressor, LightGBM delivered substantially better results across all error metrics and accuracy measures. Although CNN and LSTM approached the performance of LightGBM (with R^2 values of 0.8391 and 0.8239, respectively), LightGBM still outperformed them in terms of generalization and stability (STD = 0.1480).

Table 2. Comparison of LightGBM Model Performance with Other Models

Models/Metrics	MAE	MSE	RMSE	MAPE	R^2	STD
LightGBM	0.0470	0.0053	0.0731	12.71%	0.8480	0.1480
Garch (4,5)	1.3488	1.8538	1.3615	439.79%	-52.8508	0.1855
SVR	0.0845	0.0142	0.1193	25.72%	0.5376	0.1066
Xgboost	0.0737	0.0010	0.0983	21.58%	0.7248	0.1271
Tabnet	0.0898	0.0136	0.1170	26.26%	0.6109	0.1573
CNN	0.0534	0.0056	0.0752	0.1393%	0.8391	0.0738
Transformer	0,0942	0.0146	0.1209	0.2559	0.5848	0.1617
LSTM	0.0536	0.0061	0.0779	15.18%	0.8239	0.0778
GRU	0.4181	0.2355	0.4852	1.102%	-5.6945	0.3631

(*Source. Author's calculations)*

Notably, the GARCH model proved entirely unsuitable for the dataset, yielding a negative R^2 (-52.85%) and a MAPE exceeding 400%, indicating extremely poor forecasting capability. While Transformer and TabNet represent modern approaches, they were not optimized in this context. Overall, LightGBM not only delivered superior predictive accuracy but also maintained high stability, reinforcing its position as the preferred choice for forecasting. The GA-LightGBM model demonstrates stable forecasting performance, with a mean R^2 of 0.805 and low variation across folds (std R^2 = 0.034), confirming robustness through 5-fold cross-validation. Computational efficiency assessment indicates moderate training time (~1.74 seconds).

Table 3. Sensitivity Analysis of GA–LightGBM Hyperparameters

Hyperparameter	Low R^2	Optimal R^2	High R^2	ΔR^2 (%)
colsample_bytree	0.841541	0.841541	0.858674	2.022
learning_rate	0.846328	0.841541	0.846533	0.591
min_child_samples	0.845142	0.841541	0.841541	0.427
min_child_weight	0.841541	0.841541	0.841541	0.000
n_estimators	0.841522	0.841541	0.841557	0.004
num_leaves	0.848384	0.841541	0.842034	0.811
reg_alpha	0.842512	0.841541	0.844008	0.293
reg_lambda	0.839190	0.841541	0.841984	0.332
subsample	0.841541	0.841541	0.841541	0.000

(Source: Author's calculations)

The sensitivity analysis shows that colsample_bytree and num_leaves exert the strongest influence on model performance, with R^2 shifts of 2.02% and 0.81%, respectively. Other hyperparameters - such as learning_rate (0.59%) and reg_lambda (0.33%) - produce only minor variations. Overall, the ΔR^2 remains within 0 - 2.02%, demonstrating that the GA-LightGBM model is highly stable, with R^2 consistently centered around 0.84, confirming minimal sensitivity to hyperparameter perturbations. This stable performance suggests potential applications for portfolio rebalancing, hedging optimization against policy shocks, and supporting ESG-focused investors aligned with Net Zero 2050. Future research will extend to FDI-backed sustainable energy stocks in Vietnam (e.g., LONGi – China) and develop hybrid GARCH-LightGBM and EGARCH-LightGBM models to enhance predictive performance. Additionally, the GA-LightGBM approach could generalize to other emerging markets Indonesia, India, Brazil with similar policy environments and Net Zero commitments, as well as policy-sensitive sectors such as green finance, sustainable agriculture, and clean technology.

5. CONCLUSION

This study proposes a Genetic Algorithm–optimized LightGBM (GA–LightGBM) model to forecast the volatility of ASM.VN stocks, outperforming traditional models (GARCH, SVMRegressor, XGBoost) as well as deep learning models (GRU, TabNet, Transformer, CNN, LSTM). The model achieves MAE = 0.0470; RMSE = 0.0731; R^2 = 0.8480; STD = 0.1480, demonstrating high accuracy and stability, with a mean R^2 = 0.805 across 5-fold cross-validation. Forecasted volatility over 06 February 2025 – 06 March 2025 ranges from 0.1419 to 0.2154, reflecting moderate adjustment risk. Post-PCA SHAP analysis identifies the most influential original features in descending order: HV, BB_width, BB_lower, VWAP, ATR, MA_21, BB_upper, and OBV, emphasizing the roles of volatility, price range, liquidity, and medium-term trend in prediction. Sensitivity analysis shows that colsample_bytree and num_leaves are the most impactful hyperparameters, with ΔR^2 of 2.02% and 0.81%, while others such as learning_rate and reg_lambda induce minor variations (<0.6%). Overall, ΔR^2 ranges from 0 – 2.02%, indicating that GA–LightGBM is highly stable and minimally sensitive to hyperparameter changes. Compared with GARCH (MAPE > 400%), the model significantly improves predictive accuracy and adaptability while providing guidance for trading strategies based on volatility and market liquidity. From 09 April 2025, when the US planned a 46% retaliatory tariff on Vietnamese goods, energy stock volatility surged, notably ASM.VN (+130%), HDG.VN (+93%), BCG.VN (+60%), TV2.VN (+50%), and GEG.VN (+14.7%). By September 2025, forecasts show clear differentiation: HDG.VN remains high (0.40–0.45), ASM.VN and BCG.VN moderate, TV2.VN stable (0.30–0.40), GEG.VN around 0.38–0.40, and PC1.VN lowest (0.22–0.36), implying low short-term risk. These differences reflect financial structure, operational performance, and import dependency, particularly under a reinstated 20% import tariff. Overall, GA–LightGBM demonstrates strong potential in forecasting stock return volatility, supporting investment strategy formulation and risk management. Although the conclusions are empirically robust, economic implications should be further emphasized, as improved forecast accuracy can enhance risk-adjusted returns, hedgings efficiency, and reduce derivative market mispricing. Volatility linked to tariff shocks also highlights the causal role of trade policy in emerging markets. However, model reliability may diminish under "black swan" events or structural market changes, necessitating cautious application, periodic retraining with updated data, and hyperparameter sensitivity monitoring.

6. ACKNOWLEDGEMENTS

This research was supported by the university-level scientific research project (Code: DHKT2025-04-20) at the University of Economics, Hue University. The author gratefully acknowledges the financial and academic support provided, which contributed significantly to the completion of this study.

7. REFERENCES

Alfeus, M., Harvey, J. & Maphatsoe, P. (2025) 'Improving realised volatility forecast for emerging markets', *Journal of Economics and Finance*, 49, pp. 299-342. doi:10.1007/s12197-024-09701-x.

Al Najjar, D. (2016). Modelling and estimation of volatility using ARCH/GARCH models in Jordan's stock market. *Asian Journal of Finance & Accounting*, 8(1), 152-167.

An, S., Kim, J., Kim, H., & Lee, D. (2023). An empirical comparison of gradient boosting algorithms for tabular data: XGBoost, LightGBM, and CatBoost.

Azamjon, M., Sattarov, O., & Cho, J. (2023). Forecasting Bitcoin volatility through on-chain and whale-alert tweet analysis using the Q-learning algorithm. *IEEE Access*, PP (99), 1-1.

Dương, T. T. T. (2023). Dự báo chiều biến động của chỉ số chứng khoán bằng thuật toán tăng cường. *Tạp chí Khoa học & Đào tạo Ngân hàng*, (252).

Fan, G.F., Zhang, R.T., Cao, C.C. et al. (2024) 'The volatility mechanism and intelligent fusion forecast of new energy stock prices', *Financial Innovation*, 10(84). doi:10.1186/s40854-024-00621-7.

Financial Times. (2025). US-Vietnam trade tensions ease with revised tariff agreement.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media. Holland, J. H. (1992). *Adaptation in natural and artificial systems* (2nd ed.). MIT Press.

Huang, C., Cai, Y., Cao, J. & Deng, Y. (2025) 'Stock complex networks based on the GA-LightGBM model: The prediction of firm performance', *Information Sciences*, 700, Article 121824. doi:10.1016/j.ins.2024.121824.

Hull, J.C. (2023). *Risk management and financial institutions* (6th ed.). Wiley.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

Jorion, P. (2010). *Financial risk manager handbook* (6th ed.). Wiley.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3149-3157.

Kristjanpoller, W. (2024). A hybrid econometrics and machine learning based modeling of realized volatility of natural gas. *Financial Innovation*, 10(45).

Lambert, D. (1980). Commodity Channel Index: Tool for trading cyclic trends. *Commodities*, 1(5), 29-32.

Li, J. (2022). The comparison of LSTM, LGBM, and CNN in stock volatility prediction. In *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 905-909). Atlantis Press.

Lu, F., Ma, F., & Bouri, E. (2024). Stock market volatility predictability: New evidence from energy consumption. *Humanities and Social Sciences Communications*, 11(1624).

Lyócsa, Š. & Todorova, N. (2024) 'Forecasting of clean energy market volatility: The role of oil and the technology sector', *Energy Economics*.

Mirjalili, S. (2019). Genetic algorithm. In *Evolutionary algorithms and neural networks* (pp. 43-55). Springer. Montaño Moreno, J. J., Palmer Pol, A., Sesé Abad, A., & Cajal Blasco, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, 25(4), 500-506.

Murphy, J. J. (1999). *Technical analysis of the financial markets*. New York Institute of Finance. PeerJ Computer Science. (2023). Forecasting models: Evaluation metrics and statistical considerations. *PeerJ Computer Science*, 9, e623.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning* (3rd ed.). Packt Publishing.

Reuters. (2025, February 6). Asian equities see sharp foreign outflows in January.

Trần, B. T., & Trần, T. D. T. (2025). Forecasting ICL stock volatility and trading strategies using adaptive Laguerre artificial neural networks optimized via Bayesian optimization. In *Kỷ yếu Hội thảo Khoa học Quốc gia: Tác động của hạn hán và xâm nhập mặn tới phát triển kinh tế vùng Đồng bằng sông Cửu Long*. Nhà xuất bản Đại học Kinh tế Quốc dân.

Trần, B. T., & Trần, T. D. T. (2025). Optimized GARCH neural network model for forecasting volatility and trading strategies for First Solar stock. In *Proceedings of the International Conference of Kalasin University*.

Tsay, R. S. (2010). *Analysis of financial time series* (3rd ed.). Wiley.

U.S. Department of Commerce. (2025). Final affirmative determinations in the antidumping and countervailing duty investigations of crystalline silicon photovoltaic cells from Cambodia, Malaysia, Thailand, and Vietnam.

Ukoba, K., Onisuru, O. R., & Jen, T. C. (2024). Harnessing machine learning for sustainable futures: Advancements in renewable energy and climate change mitigation. *Bulletin of the National Research Centre*. Springer Open.

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

Yahoo Finance. (2025). ASM.VN historical data.

Yang, Q. (2024) 'Blending Ensemble for Classification with Genetic-algorithm generated Alpha factors and Sentiments (GAS)', *arXiv preprint*, arXiv:2411.03035.