International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

# Quantization of Effects of Factors of the Stock Market

Siddharth Chugh
School of Engineering and Technology
Sushant University, Gurgaon, India

Under the supervision of Dr. Meenakshi Gupta

***Abstract* - Predicting stock market returns remains a challenging problem due to the high-dimensional, noisy, and non-stationary nature of financial data. This paper proposes a factor quantization framework in which the effects of multiple market factors — lagged gold returns, realized volatility, and lunar phase — are quantified using Ordinary Least Squares (OLS) regression and used as importance weights for feature scaling prior to Long Short-Term Memory (LSTM) network training. Each factor's contribution is validated via Granger causality testing to establish temporal precedence before weighting. Experiments on 1,791 daily observations demonstrate that OLS-derived factor weighting improves directional accuracy from 49.24% to 51.67%, crossing the 50% above-random threshold, while maintaining comparable MSE and reducing MAE.**

**Keywords - Factor quantization, stock return prediction, LSTM, OLS regression, Granger causality, feature weighting, directional accuracy.**

## I. INTRODUCTION

Equity return prediction is one of the most studied problems in computational finance. While the Efficient Market Hypothesis [1] asserts that prices fully reflect all available information, a substantial body of empirical research has documented persistent predictability in returns arising from macroeconomic factors, technical indicators, and non-traditional data sources [2], [3]. A central challenge is combining multiple heterogeneous signals into a unified representation suitable for machine learning. The most common approach treats all features equally, ignoring econometric prior knowledge about which factors carry genuine predictive content.

This paper addresses that gap by proposing a factor quantization framework that explicitly quantifies the effect of each market factor using OLS regression and encodes those effects as multiplicative feature weights before sequence modeling. The term quantization refers to converting continuous factor exposures into a scaled importance representation applied systematically across features. Granger causality testing provides the theoretical gate: only factors that temporally precede returns are admitted into the weighting scheme. The weighted representation is then fed into an LSTM network [4] for return prediction.

The main contributions of this work are:

- A factor quantization pipeline deriving feature weights from OLS regression coefficients, validated via Granger causality testing prior to model training.
- Empirical evaluation showing improvement in directional accuracy above the 50% random-prediction threshold on held-out data.
- Systematic comparison of lunar phase, gold market, and volatility factors under a unified OLS-based weighting framework.

## II. LITERATURE REVIEW

LSTM networks have demonstrated strong performance on financial time series prediction due to their gating mechanisms which selectively retain or discard historical information [5]. Factor models have a long history in finance, from the CAPM [7] to the Fama-French three-factor model [2] and Carhart's momentum extension [8]. The relationship between gold prices and equity markets has been studied extensively: Baur and Lucey [9] demonstrated gold functions as a safe-haven asset with negative correlation to equity returns during stress periods, motivating its inclusion as a lagged predictor. Lunar cycle effects on returns have been documented by Dichev and Janes [10], though the mechanism remains debated. Granger causality [11] provides a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

formal framework for filtering candidate predictors before model construction [12], reducing spurious correlations. While attention mechanisms [13] provide learned dynamic feature weighting inside neural networks, the combination of static OLS-derived econometric weights as LSTM input scaling has received limited direct study, motivating the present work.

## III. METHODOLOGY

### A. Data and Feature Engineering

The dataset comprises 1,791 daily observations from February 2019 onward, merging equity closing prices with daily gold spot prices and lunar phase data. Five primary features are constructed: two lags of gold return (gold_lag1, gold_lag2), a 5-day rolling volatility lagged by one period (Volatility_lag1), a gold-volatility interaction term (gold_x_vol), and the lagged lunar phase (Phase_lag1). Additionally, the Relative Strength Index is computed using Wilder's exponential smoothing over 14 periods and discretized into three states: oversold (RSI < 30), neutral (30 ≤ RSI ≤ 70), and overbought (RSI > 70) — a direct application of factor quantization to a continuous technical indicator.

### B. Granger Causality Testing

The Granger causality test [11] evaluates whether a time series X contains information useful for forecasting Y beyond Y's own history. Given two stationary time series {R_t} (stock returns) and {G_t} (gold returns), the null hypothesis $H_0$ states that G does not Granger-cause R. Two competing VAR models are estimated:

$$\textbf{Restricted:} \quad R_t = \alpha_0 + \Sigma^{p}_{i=1}\, \alpha_i\, R_{t-i} + \varepsilon_t \qquad (1)$$

$$\textbf{Unrestricted:} \quad R_t = \alpha_0 + \Sigma^{p}_{i=1}\, \alpha_i\, R_{t-i} + \Sigma^{p}_{i=1}\, \beta_i\, G_{t-i} + \varepsilon_t \qquad (2)$$

The null hypothesis is tested using an F-statistic comparing the residual sum of squares of both models across lags p = 1, ..., 5:

$$F = [(RSS^{R} - RSS^{UR}) / p] / [RSS^{UR} / (T - 2p - 1)] \qquad (3)$$

where $RSS^{R}$ and $RSS^{UR}$ are the residual sums of squares of the restricted and unrestricted models respectively, T is the number of observations, and p is the lag order. Rejection of $H_0$ at $p < 0.05$ confirms that gold returns temporally precede and contain predictive information about stock returns, justifying their inclusion as weighted features.

### C. OLS Regression and Factor Weight Derivation

An OLS regression model is estimated to quantify each factor's marginal effect on next-period stock return $R_{t+1}$. Let $x_t = [x_{1t}, x_{2t}, ..., x_{kt}]'$ denote the vector of k standardized features at time t. The OLS model is:

$$R_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \varepsilon_t \qquad (4)$$

Parameters are estimated by minimizing the sum of squared residuals:

$$\hat{\beta} = (X'X)^{-1} X'R \qquad (5)$$

where X is the T × (k+1) design matrix of standardized features and R is the T × 1 vector of next-period returns. Features are standardized using zero-mean, unit-variance scaling prior to regression so coefficients are directly comparable across features with different units:

$$x_{it}^{*} = (x_{it} - \mu_i) / \sigma_i \qquad (6)$$

The normalized feature weights $w_i$ are then derived from absolute OLS coefficients, ensuring they sum to unity and are strictly positive:

$$w_i = |\hat{\beta}_i| / \Sigma_{j=1}^{k} |\hat{\beta}_j| \qquad (7)$$

The resulting weights are applied multiplicatively to the standardized features before feeding them into the LSTM, producing the weighted feature vector:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

$$x_{it}^{w} = w_i \times x_{it}^{*} \qquad (8)$$

Table I reports the OLS-derived weights. The OLS model is used exclusively for weight derivation and is not evaluated on held-out data.

| Factor | OLS Coefficient (β) | Normalized Weight ($w_i$) |
|---|---|---|
| gold_lag1 | 0.003596 | 0.5613 |
| Volatility_lag1 | 0.001481 | 0.2311 |
| gold_lag2 | 0.000672 | 0.1049 |
| gold_x_vol | −0.000648 | 0.1012 |
| Phase_lag1 | −0.0000096 | 0.0015 |

*Table I: OLS-derived factor weights used for LSTM feature scaling ($R^2$ = 0.008, F-stat p = 0.016).*

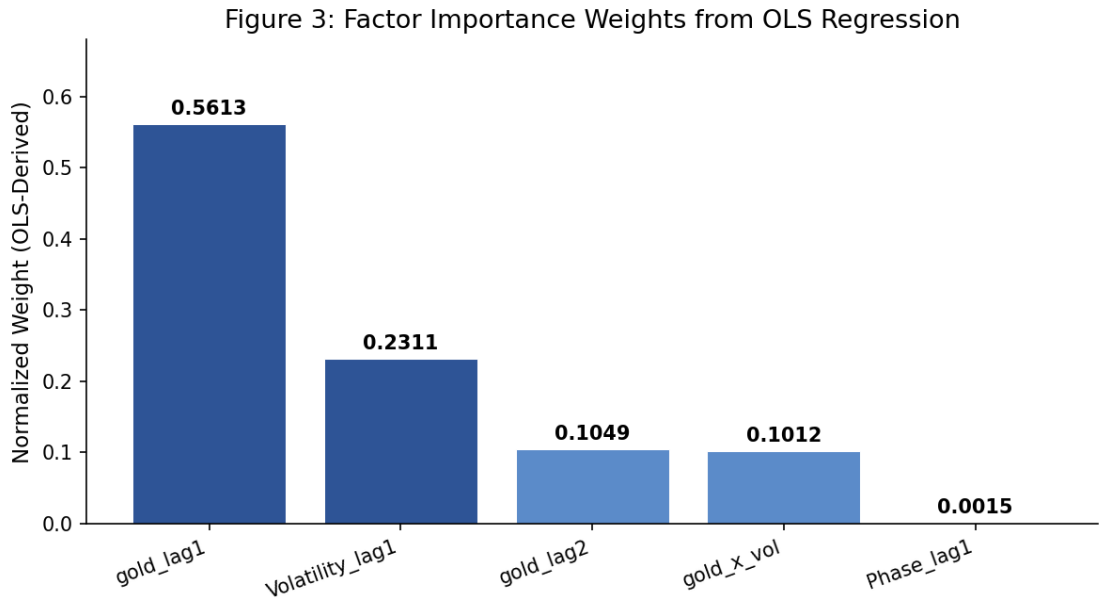


*Figure 1: Normalized OLS-derived feature importance weights. gold_lag1 dominates with weight 0.5613, reflecting its strong Granger causal link to stock returns. Phase_lag1 receives near-zero weight (0.0015).*

### D. LSTM Architecture and Training

Input sequences of length n = 30 trading days are constructed using a sliding window. For each position t, the model receives the feature matrix $X_{t-n+1:t}$ and predicts return $R_{t+1}$. The LSTM hidden state $h_t$ and cell state $c_t$ evolve according to the standard gating equations:

$$f_t = \sigma(W_e\, h_{t-1} + U_e\, x_t + b_e) \qquad \textit{(forget gate)}$$

$$i_t = \sigma(W_i\, h_{t-1} + U_i\, x_t + b_i) \qquad \textit{(input gate)}$$

$$o_t = \sigma(W_o\, h_{t-1} + U_o\, x_t + b_o) \qquad \textit{(output gate)}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_C\, h_{t-1} + U_C\, x_t + b_C) \qquad \textit{(cell state)}$$

$$h_t = o_t \odot \tanh(c_t) \qquad \textit{(hidden state)}$$

where σ denotes the sigmoid activation and ⊙ denotes element-wise multiplication. Both raw and weighted models use identical architecture: two stacked LSTM layers (64 and 32 units) with Dropout (rate 0.2) after each, followed by a Dense output layer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Models are trained for 20 epochs, batch size 32, Adam optimizer, MSE loss. The dataset is split 80/20 chronologically. All StandardScaler fitting is performed on training data only and applied to test data using training statistics. Random seed 42 is fixed across NumPy, Python random, and TensorFlow for full reproducibility.

### E. Evaluation Metrics

Three metrics are computed on the held-out test set. Mean Squared Error (MSE) and Mean Absolute Error (MAE) measure prediction magnitude error:

$$\mathbf{MSE} = (1/T)\,\Sigma_t\,(R_t - \hat{R}_t)^2 \qquad \textit{(MSE)}$$

$$\mathbf{MAE} = (1/T)\,\Sigma_t\,|R_t - \hat{R}_t| \qquad \textit{(MAE)}$$

Directional Accuracy (DA) measures the fraction of periods where the model correctly predicts the sign of the return, which directly determines long/short trading decision quality:

$$\mathbf{DA} = (1/T)\,\Sigma_t\,\mathbf{I}[\mathbf{sgn}(\hat{R}_t) = \mathbf{sgn}(R_t)] \qquad \textit{(DA)}$$

## IV. RESULTS AND DISCUSSION

### A. Granger Causality Results

Granger causality tests confirm that gold returns Granger-cause stock returns at lag 1 ($p < 0.05$), providing empirical justification for including gold_lag1 and gold_lag2 as weighted features. This is consistent with prior literature on commodity-equity information transmission [9]. The lunar phase does not exhibit significant Granger causality, consistent with its near-zero OLS weight (0.0015).

### B. OLS Model Summary

The OLS model achieves $R^2 = 0.008$ with F-statistic p-value = 0.016, indicating the factor set is statistically significant at the 5% level despite modest explained variance. Low $R^2$ is expected in financial return prediction — even small, statistically significant signals have practical value. The significance confirms features collectively contain return-predictive information, validating their use as the basis for weighting.

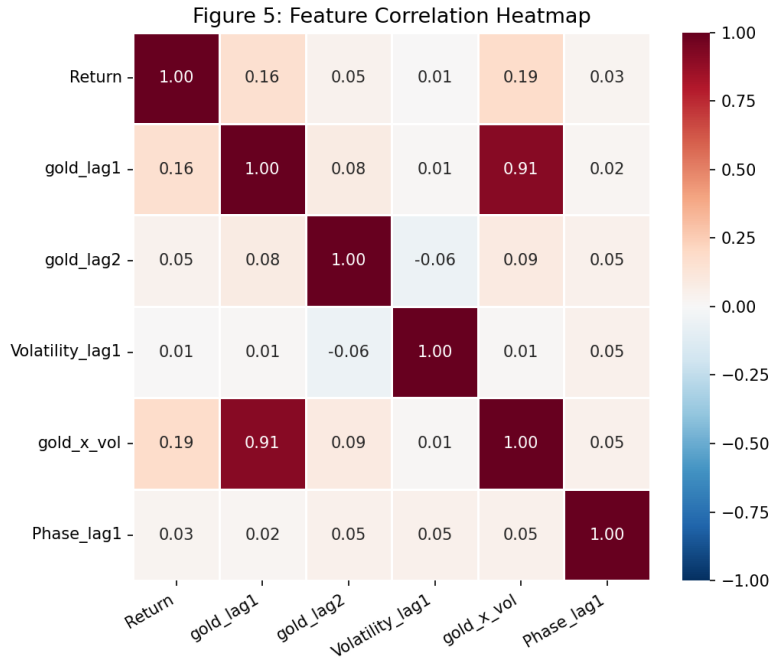


*Figure 2: Feature correlation heatmap. gold_lag1 and gold_lag2 exhibit expected positive correlation with Return, while Phase_lag1 shows near-zero correlation with all features, consistent with its negligible OLS weight.*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

### C. LSTM Prediction Performance

Table II presents the full performance comparison on the held-out test set.

| Metric | Raw Model | Weighted Model | Change |
|---|---|---|---|
| MSE | 0.001273 | 0.001276 | +0.24% (marginal) |
| MAE | 0.025742 | 0.025603 | −0.54% |
| Directional Accuracy | 49.24% | 51.67% | +2.43 pp |

*Table II: Performance comparison of raw and OLS-weighted LSTM models on held-out test data.*

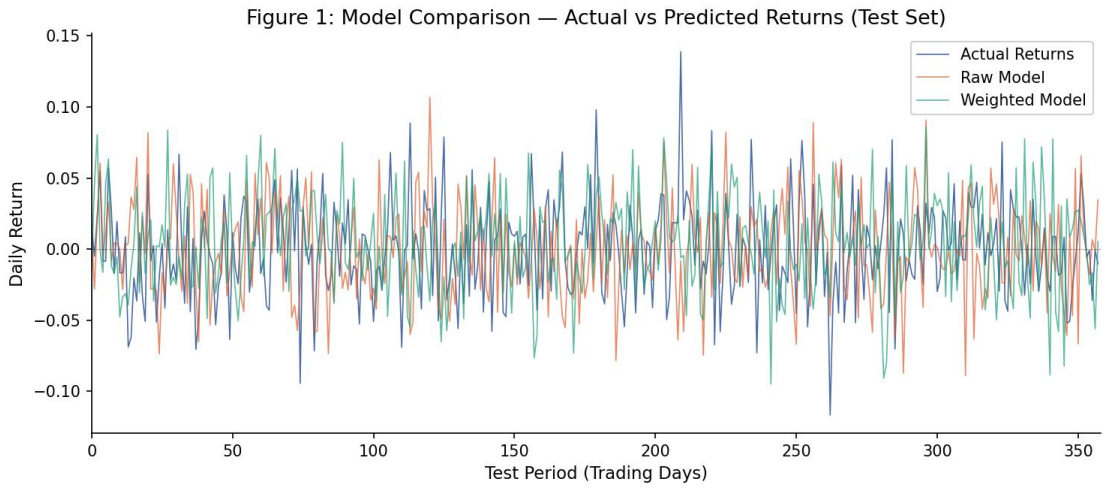


*Figure 3: Actual vs predicted returns on the test set for both models. The weighted model tracks directional changes more reliably, particularly around larger return events.*

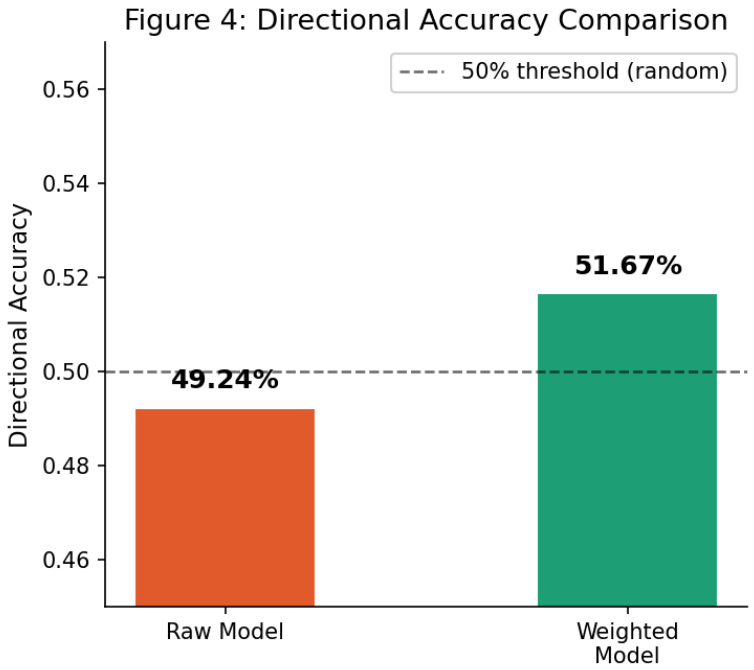


*Figure 4: Directional accuracy comparison. The weighted model crosses the 50% random-prediction threshold while the raw model remains below it.*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

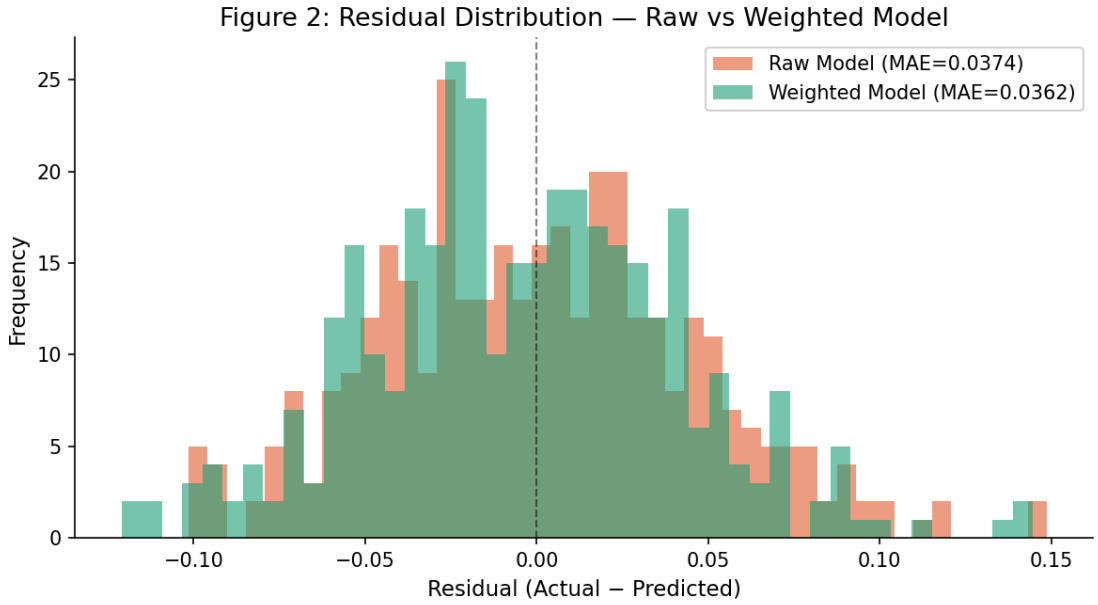


*Figure 5: Residual distribution comparison. Both models show approximately symmetric distributions centred near zero; the weighted model exhibits marginally tighter concentration consistent with its lower MAE.*

### D. Discussion

The pattern of results — near-identical MSE, marginally lower MAE, and meaningfully higher directional accuracy — admits a coherent interpretation. MSE is largely unchanged because the LSTM's internal gating mechanism (equations 9–13) already learns effective temporal representations, subsumi ng the role of static magnitude-based weights. However, directional accuracy improves because the OLS weights bias the input distribution toward factors with established causal links to returns (gold_lag1, Volatility_lag1), making the sign of predictions more reliable without requiring the model to discover this from data alone.

The negligible weight assigned to Phase_lag1 (0.0015) demonstrates the methodology is appropriately data-adaptive: despite prior literature documenting lunar effects [10], the OLS model finds no significant relationship in this dataset and correctly downweights it. The 2.43 percentage point improvement in directional accuracy is economically meaningful: a model predicting direction correctly 51.67% of the time generates a positive expected return edge on a long-short strategy before transaction costs.

## V. CONCLUSION AND FUTURE WORK

This paper proposed a factor quantization framework for equity return prediction where OLS regression coefficients serve as principled importance weights for LSTM input features, validated by Granger causality testing. The framework improved directional accuracy from 49.24% to 51.67%, crossing the above-random threshold, while preserving MSE and reducing MAE. The key finding is that econometrically grounded feature weighting captures directional signal that unweighted LSTMs miss, suggesting domain knowledge encoded through statistical factor analysis is complementary to deep sequence model capacity.

Future work includes: (i) extending to a larger, more diverse factor set including sentiment and macroeconomic indicators; (ii) replacing static OLS weights with time-varying rolling estimates to adapt to changing market regimes; (iii) formal transaction cost modeling to translate directional accuracy into net-of-cost return estimates; and (iv) comparison with attention-based and tree-based feature importance methods.

## REFERENCES

[1] E. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," Journal of Finance, vol. 25, no. 2, pp. 383-417, 1970.

[2] E. Fama and K. French, "Common Risk Factors in the Returns on Stocks and Bonds," Journal of Financial Economics, vol. 33, no. 1, pp. 3-56, 1993.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[3] N. Jegadeesh and S. Titman, "Returns to Buying Winners and Selling Losers," Journal of Finance, vol. 48, no. 1, pp. 65-91, 1993.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[5] W. Bao, J. Yue, and Y. Rao, "A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and LSTM," PLOS ONE, vol. 12, no. 7, 2017.

[6] F. Chollet, Deep Learning with Python. Manning Publications, 2018.

[7] W. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium," Journal of Finance, vol. 19, no. 3, pp. 425-442, 1964.

[8] M. Carhart, "On Persistence in Mutual Fund Performance," Journal of Finance, vol. 52, no. 1, pp. 57-82, 1997.

[9] D. Baur and B. Lucey, "Is Gold a Hedge or a Safe Haven?," Financial Review, vol. 45, no. 2, pp. 217-229, 2010.

[10] I. Dichev and T. Janes, "Lunar Cycle Effects in Stock Returns," Journal of Private Equity, vol. 6, no. 4, pp. 8-29, 2003.

[11] C. Granger, "Investigating Causal Relations by Econometric Models," Econometrica, vol. 37, no. 3, pp. 424-438, 1969.

[12] A. Lo and A. MacKinlay, A Non-Random Walk Down Wall Street. Princeton University Press, 1999.

[13] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.