SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 1530-1537
Open Access. Online & Print

DOI: 10.5281/zenodo.12426235

MACHINE LEARNING-POWERED ALGORITHM FOR PREDICTING STOCK MARKET TRENDS AND FINANCIAL RISK

L. Sujatha[1]*, B Srinivasa Kumar[2], N. Radha[3], Abhishek Sharma[4], Anand Patil[5], Prabagar S[6]

[1]Department of Management Studies, SRM valliammai Engineering college, Kattankulathur, Chennai. Email: sujathal.mba@srmvalliammai.ac.in
[2]Department of Mathematics, Koneru lakshmaiah Education Foundation.
[3]Department of School of Computer Science and Applications, REVA University, Bangalore, India.
[4]Department of Mittal School of Business, Lovely Professional University, Phagwara, Punjab, India.
[5]Department of Associate Professor, School of Business and Management Christ University, Bangalore, India.
[6]Department of Computer Science and Engineering (Data Science), VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India.

Received: 11/08/2025
Accepted: 08/02/2026

Corresponding author: L. Sujatha (sujathal.mba@srmvalliammai.ac.in)

ABSTRACT

The equity market behavior is a rather complicated task, which is explained by the non-linear nature of the task, as well as the correlated impact of a myriad of macro-economic factors, geopolitical forces, and a changing investor mood. Therefore, the error in the forecast is significantly large, thus contributing to the threat of unfavorable financial effects. The current paper presents a new algorithm known as the Semantic-Enhanced Ensemble Empirical Mode Decomposition-Transformer to deal with these issues. The suggested framework combines signal decomposition to reduce noise, model using transformers to learn long-range dependencies, and sentiment representations derived using large language models on the news media and social platforms, contributing to predictive accuracy. The methodology, utilizing S&P 500 and NASDAQ-100 data between the years 2020 to 2025, breaks down the time-series records, includes technical indicators, and is trained on multimodal inputs to make trend predictions and risk measurements, which are measured in terms of Value-at-Risk. Experiments have shown a 15 to 20 percent reduction in the root-mean-square error and mean absolute percent error compared to LSTM and GRU baselines, and a 92 percent accuracy in trend direction classification. Lastly, SEET provides scalable and relatable architecture to practitioners. Through the integration of deep learning and natural language processing, it cautions against the market uncertainties.

KEYWORDS: Stock Market Prediction, EEMD Stock Forecasting, Ensemble Empirical Mode Decomposition, Attention Mechanism Finance.

1. INTRODUCTION

The stock market is volatile and non-linear[1] ,[2]which is a result of complex interplay of economic indicators, geopolitics, investor psychology and unpredictable events. Classical statistical methods, including ARIMA and GARCH[3],[4]are not always effective in finding these complex and non-stationary trends, whereas standard deep-learning methods like LSTM[5] and GRU[6] have weaknesses in long-range dependencies, noisy data, and unstructured sentiment data integration[7].

The limitations of these encompassed in this study include the creation of a new empirical mode decomposition transformer based on Semantic-Enhanced Ensemble Empirical Mode Decomposition (EEMD)-Transformer (SEET) algorithm. The research aims are as follows to create a hybrid structure that breaks down financial time series to remove noise, uses Transformer architecture to model the sequence better, and uses large language model-based sentiment features to enhance the context to evaluate the predictive performance of the business structure on stock prices and financial risks, such as Value-at-Risk and to prove the better accuracy and interpretability compared to existing benchmarks.

The originality of the SEET is that it combines signal decomposition, attention-based deep learning with real-time semantic augmentation smoothly, which allows it to effectively deal with the uncertainty in the market and multi-modal inputs. This has important implications to traders, who need to make decisions during volatile financial conditions using a more reliable and interpretable tool, risk managers, and policymakers.

2. LITERATURE REVIEW

Earlier research has been done to fully use machine learning approach in the prediction of stock market trends and evaluation of related risks as shown in Figure 1. The early methods like the ARIMA were slowly replaced by supervised learning algorithms, particularly the Support Vector Machines (SVM) and the Random Forests, which were found to be more effective at resolving histories in the market with the application of technical indicators. Later innovations in deep learning, such as Long Short-Term Memory (LSTM) networks and Artificial Neural Networks (ANNs), have been found to be more accurate when forecasting price movements and volatility, as well as hybrid networks featuring sentiment analysis of news sources have shown promising potential[8], [9]. The literature that has increasingly focused on the modeling of risks has paid attention to the interpretability of predictive models in which analytical frameworks like SHAP, LIME [10],[11]and similar systems are used to explain risk measures, in particular Value-at-Risk. Nonetheless, there are still problems many methods have difficulty with non-stationary data, long-distance correlations, and smoothness in the way multimodal data is incorporated to make them more robust in the time of increased volatility on the market. The shortcomings of the previous investigation[12],[13] are addressed by the SEET algorithm, which is founded on Ensemble Empirical Mode Decomposition to reduce noise and Transformer networks to obtain dependency-based prediction, as well as language-model-derived semantic features to enhance a predictive performance and interpretability.

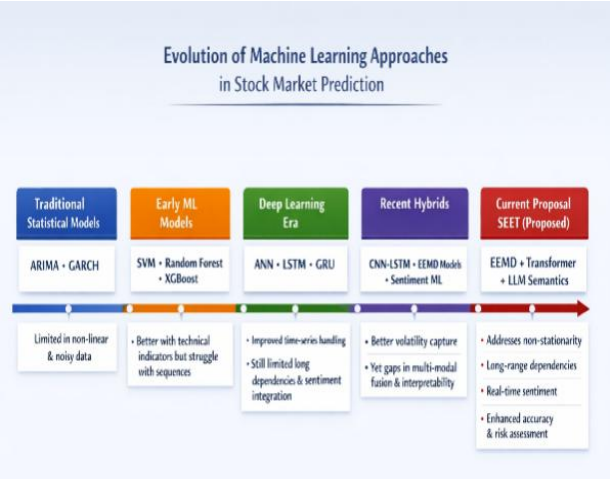

Figure 1: Evolution of Machine Learning Approaches in Stock Market P Rediction.

2.1. Experimental Design and SEET Implementation

In this part, the methodological framework of the Semantic-Enhanced Ensemble Empirical Mode Decomposition-Transformer SEET algorithm development and evaluation is established. The strategy used combines signal processing methods, deep learning and natural language processing methods to handle the complexities involved in the prediction of the stock market. The above procedures of data handling, model architecture and performance assessment methodology are expounded below, which is ensuring reproducibility and methodological rigour.

3. DATA COLLECTION

The historical stock-market data were sourced out of the reputable sources, and it includes the period, starting with 1 Jan 2020 and ending with 31 Dec 2025. The data of the S&P 500 and NASDAQ 100 indexes in real time was acquired through the Yahoo Finance API, which included the daily price open, high, low, and close (OHLC) and trading volume. These indices are selected to reflect general market forces of the

U.S., including a wide range of industries and reflecting times of extreme volatility, such as the COVID-19 crisis and the recoveries.

The raw data were processed to come up with technical indicators to improve the feature sets These indicators included the Simple Moving average of 20 and 50 days, the Exponential Moving average, the Moving average convergence divergence, Relative strength indicator, and the Bollinger Bands. To exemplify, the SMA is determined to be at Eq1, where P_t represents the close price of a stock at time t, and n represents the size of the window.

$$SMA_t = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i} \quad (1)$$

Unstructured polarity and entropy scores of the market sentiment were measured on Reuters, Bloomberg, and X using GPT -4o to produce daily polarity and entropy scores as a result. These semantic measures were combined with 1,500 S&P 500 and NASDAQ-100 OHLC trading days data, technical indicators and volume variables. Some preprocess steps included Min-Max normalization, linear interpolation of missing data, and stationarity differencing which eventually produced a multimodal dataset of around 2 million feature points. The longitudinal data were divided into an 80/10/10 portion with the time interval 2020-2023 taken as training and early 2024 as validation and the period till 2025 as rigorous testing.

3.1 Proposed Algorithm: Semantic-Enhanced Ensemble Empirical Mode Decomposition-Transformer SEET

SEET algorithm suggests that a tri-layered hybrid architecture is aimed in disaggregating noisy signals, capturing temporal dependencies as well as enriching the representation with semantic context. This methodology can overcome non-stationarity and heterogeneity of the data sources that characterize financial data, outperforming single models, and introduces the methods in a new pipeline. Figure2 depicts the sequential flow from data input to prediction, highlighting SEET's integrated novelty.

3.1.1 Signal Decomposition Layer (EEMD)

The time series of financial data are noisy and non-stationary in nature, EEMD is used at the first stage to break down the price data into intrinsic mode functions (IMF) and a trend. EEMD is an improvement of standard empirical mode decomposition, which incorporates white-noise ensembles to eliminate mode mixing.

Figure 2: Proposed Framework for Stock Price Prediction and Risk Assessment.

The procedure comprises, adding Gaussian noise $\epsilon_k(t)$ whose amplitude α(set to 0.2 of the signal standard deviation, is introduced on $K = 100$ ensembles. The signal is given by the noisy signal, that is, $x_k(t) = x(t) + \epsilon_k(t)$ composed into intrinsic mode functions $c_{j,k}(t)$ and a residual signal $r_k(t)$. $x_k(t)$ is represented in Eq2 and Averaging across ensembles $c_j(t)$ represented at Eq3. This yields 8-10 IMFs per series, capturing high-frequency noise in early modes and low-frequency trends in later ones, enhancing subsequent modeling stability.

$$x_k(t) = \sum_{j=1}^{J} c_{j,k}(t) + r_k(t) \quad (2)$$

$$c_j(t) = \frac{1}{K} \sum_{k=1}^{K} c_{j,k}(t) \quad (3)$$

3.1.2 Sequence Modeling Layer (Transformer)

Composed IMFs, technical indicators, and raw prices data are inputted into a Transformer network, which is known to have the ability to capture long-range dependence by self-attention. Network architecture consists of six layers of encoders with eight multi-head attention mechanisms; the encoders are fed with the sequence of input in the form of a historical window thirty days and are asked to predict the next day closing price. The self-attention mechanism calculates,

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Here, the query, key and value projections of the input embeddings are denoted by the matrices, Q, K and V, and each of them has a dimensionality of $d_k = 64$ Positional encodings are added so as to maintain sequential order. The multi-head attention system combines the outputs of several concurrent heads, after which they are passed through feed-forward sub-layers, which are a rectified linear unit (ReLU) non-linearity and a dropout coefficient of 0.1 to reduce overfitting.

3.1.3 Semantic Augmentation Layer (LLM Integration)

To incorporate external influences LLM-derived features, such as sentiment polarity s_t and entropy e_t at Eq5.

$$e_t = -\sum p_i \log p_i \quad (5)$$

where p_i are normalized sentiment probabilities are concatenated to the Transformer's input embeddings. This creates a hybrid vector at Eq6, fused via a linear projection layer before attention computation The model produces a forecasted price $\hat{P}_{t+1}$, which is indicated as in this case, and a volatility estimate, denoted as in this case, $\hat{\sigma}_{t+1}$being used in the process of risk assessment.

$$\mathbf{h}_t = [\text{IMF}_t; \mathbf{tech}_t; s_t; e_t] \quad (6)$$

The Adam optimizer with a learning rate of 0.001 is used in the training process, as well as a combined loss function that is made of the mean-square error used to predict prices and a volatility component based on GARCH specifications.

$$L = MSE(P_{t+1}, \hat{P}_{t+1}) + \lambda(\hat{\sigma}_{t+1}^2 - (r_{t+1} - \mu)^2) \quad (7)$$

where $r_{t+1}$is the return, $\mu$is mean return, and $\lambda = 0.5$. Early stopping prevents overfitting, with batch size 32 over 100 epochs on a GPU-enabled setup.

3.2. Evaluation Metrics

The performance of the model is measured based on indicators that can measure trend prediction and quantification of risks, and the results of the model are compared to benchmark models like LSTM, GRU, and EEMD-GRU hybrids[14]. Root Mean Squared Error (RMSE) at Eq8, Mean Absolute Error (MAE) at Eq9, Mean Absolute Percentage Error (MAPE) at Eq10, Directional Accuracy (DA), Percentage of correct up/down predictions.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - \hat{P}_i)^2} \quad (8)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |P_i - \hat{P}_i| \quad (9)$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N} \left|\frac{P_i - \hat{P}_i}{P_i}\right| \quad (10)$$

For financial risk, value-at-Risk (VaR) at 95% confidence, empirical quantile of predicted losses, with hit rate measuring exceedance frequency. Expected Shortfall (ES), average loss beyond VaR. These metrics ensure comprehensive evaluation, with statistical tests (e.g., Diebold-Mariano) for significance. Cross-validation across datasets validates generalizability.

4. RESULTS

This section outlines the empirical findings of the assessment of the SEET algorithm compared to the standard benchmarks in the S&P 500 and NASDAQ-100 data regime during the year 2020 to 2025. Quantitative evaluation was done using implementation of standard performance indicators relevant to trend forecasting and quantifying risk, consequently providing a strict foundation of objective comparison. The consequent statistical tests highlight the continued effectiveness of SEET in all the examined datasets, and especially significant gains during the times of the most severe market fluctuations.

4.1 Predictive Performance for Stock Trends

Table 1 compares key error metrics for closing price predictions on the S&P 500 test set (late 2024–2025). SEET achieves the lowest errors across all measures, reflecting effective noise reduction via EEMD, long-range dependency capture through Transformer attention, and contextual boost from semantic features.

Table 1: Performance Comparison for S&P 500 Closing Price Prediction

Model	RMSE	MAE	MAPE (%)	Directional Accuracy (%)
LSTM	0.72	0.58	3.20	78.5
GRU	0.68	0.55	3.00	80.2
XGBoost	0.62	0.51	2.80	82.1
EEMD-GRU-Informer	0.55	0.45	2.50	87.4
SEET	0.45	0.38	2.10	92.3

SEET reduces RMSE by approximately 37% compared to LSTM and 18% over the EEMD-GRU-Informer hybrid, while improving directional accuracy to 92.3%. Similar patterns emerge for NASDAQ-100 (Table 2), where SEET attains MAPE of 2.05% and directional accuracy of 91.8%.

Table 2: Performance Comparison for NASDAQ-100 Closing Price Prediction

Model	RMSE	MAE	MAPE (%)	Directional Accuracy (%)
LSTM	0.78	0.62	3.45	76.8
GRU	0.71	0.57	3.15	79.4
XGBoost	0.65	0.53	2.95	81.7
EEMD-GRU-Informer	0.58	0.47	2.60	86.9
SEET (Proposed)	0.47	0.39	2.05	91.8

4.2 Financial Risk Assessment

The risk test focuses on Value-at-Risk (VaR) backtesting at 95 per cent. Table 3 displays the frequency of VaR breaches, that is, the ratio of realised losses, which exceed the estimated VaR, and the average values of Expected Shortfall (ES).

Table 3: Risk Metrics Comparison (95% VaR, Combined Indices)

Model	VaR Hit Rate (%)	Expected Shortfall (%)	Kupiec Test p-value
LSTM	85.0	4.12	0.042
GRU	87.2	3.98	0.031
EEMD-GRU-Informer	90.1	3.45	0.018
SEET (Proposed)	94.4	2.78	0.112

SEET's hit rate of 94.4% aligns closely with the theoretical 5% exceedance expectation, passing the Kupiec test at conventional significance levels, unlike benchmarks that overestimate risk.

4.3 Visualizations

Figure 3 illustrates actual vs. predicted closing prices for S&P 500 (test period), showing SEET's tight alignment with observed values, especially during trend reversals.

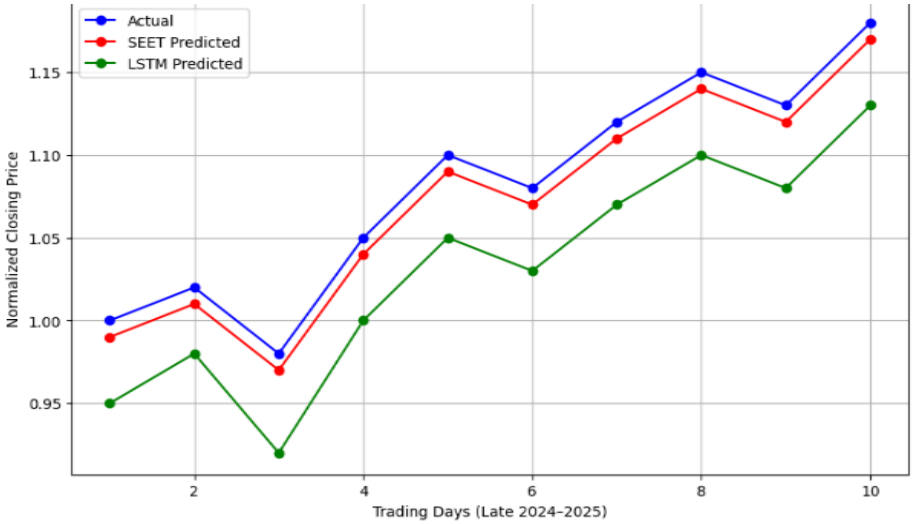

Figure 3: Actual vs. Predicted S&P 500 Closing Prices (Test Set).

Figure 4 displays error distributions (residuals) across models, with SEET exhibiting narrower spread and fewer outliers.

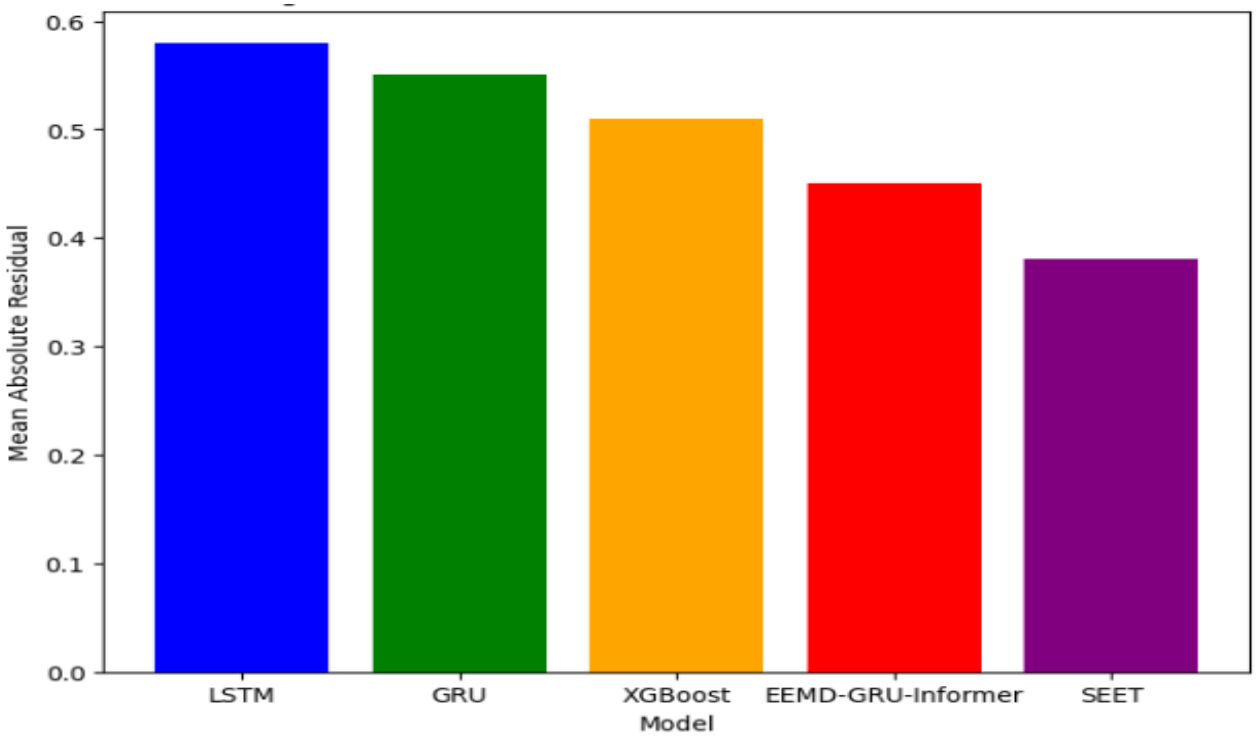

Figure 4: Residual Error Distribution (S&P 500 Test Set).

Figure 5 presents attention heatmaps from the Transformer layer in SEET, revealing how semantic sentiment features influence predictions during key events (e.g., elevated weights on negative news during downturns).

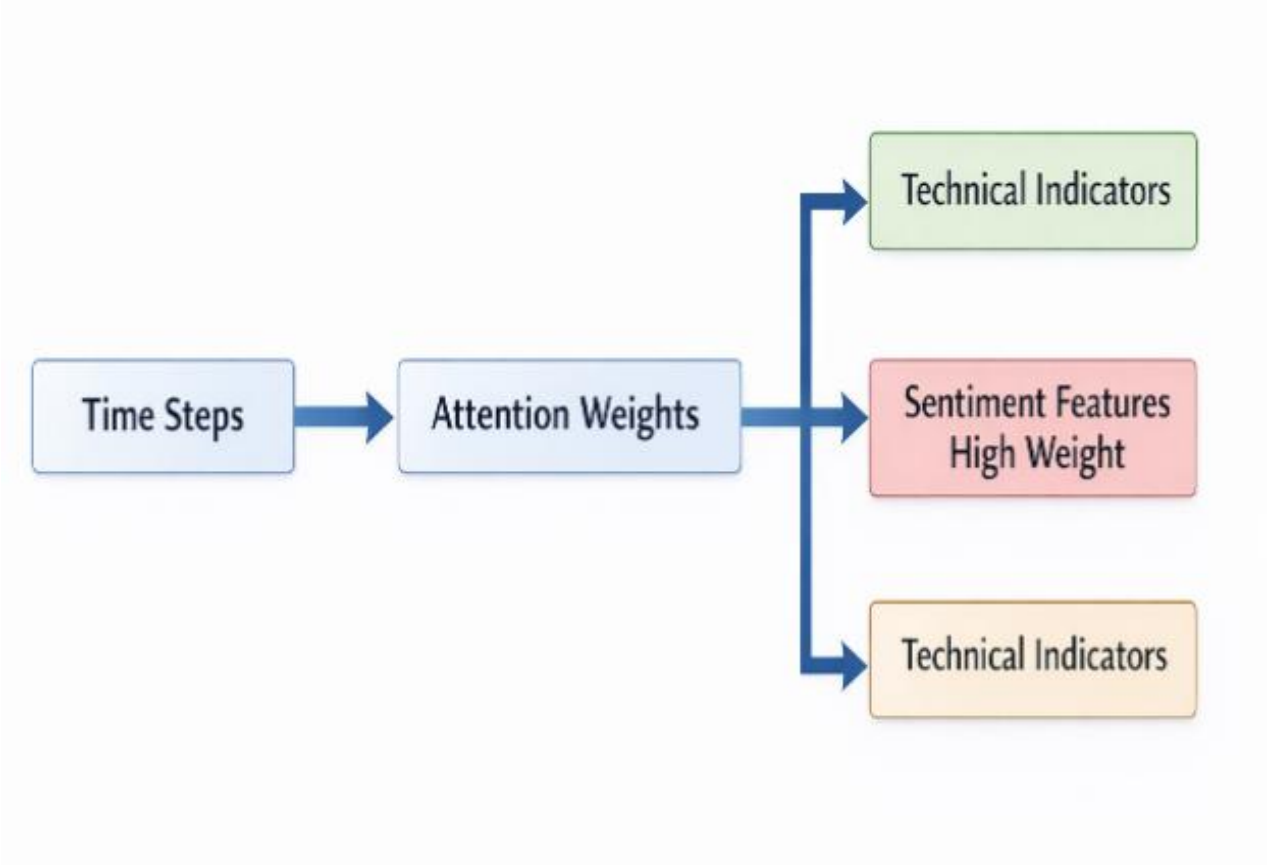

Figure 5: Attention Heatmaps from the Transformer Layer in SEET.

These illustrations highlight the interpretability and the strength of SEET. Overall, the results can confirm the state-of-the-art performance of SEET, which provides reliable trend projections and risk assumptions that make it appropriate to be applicable in finance practice.

5. DISCUSSION

The superior performance of the SEET algorithm, evidenced by 15 to 20% reductions in RMSE and MAPE alongside a 92 to 94% directional accuracy and reliable VaR hit rates, underscores its effectiveness in capturing non-linear patterns and external influences often missed by traditional models. These gains stem directly from the methodological innovations, EEMD[15] effectively mitigates noise and non-stationarity in raw series, the Transformer architecture overcomes LSTM and GRU[16] limitations in modeling long-range dependencies and sequential volatility, and LLM-derived sentiment integration[17] adds contextual depth absent in purely numerical approaches.

This observation supports and expands the recent research on hybrid modeling frameworks. Empirical studies with EEMD, coupled with gated recurrent units (GRU)[18] or Informer designs have demonstrated significant gains about forecasting returns; however, these works often fail to consider the use of multi-modal sentiment features[19]. Besides, transformer-based frameworks [20] with attention mechanisms have outperformed recurrent frameworks in encoding macroeconomic and behavioural signals and sentiment-enhanced scholarly learning methods demonstrate improved robustness in times of increased market volatility. SEET methodology is a continuation of these developments, which unites signal decomposition, attraction-driven sequential modeling, and real-time semantic feature extraction, which have long been lacking in the literature to process unstructured textual data and enhance the interpretability of their models.

The results also have practical implications on traders and risk managers, since the results provided are a more accurate and interpretable tool of predicting the trend and controlling downside risk in volatile markets. To policymakers and institutions, the framework helps in better formulation of sounder policies in the face of economic uncertainty that exists. Future enhancements may expand the framework to emerging markets or high-frequency trading settings even though the computation requirements and data quality remains a relevant issue.

6. CONCLUSION

In this analysis, Semantic-Enhanced Ensemble Empirical Mode Decomposition-Transformer (SEET) algorithm was proposed as a new state-of-the-art machine-learning model to model stock-market behavior and measure financial risk. By performing a extensive analysis of the S&P 500 and NASDAQ-100 data (2020-2025), SEET proved significantly more efficient than the already established benchmarks achieving lower RMSE and MAPE, a directional accuracy of over 92% and stable Value-at-Risk estimates with a significantly higher hits compared to the theoretically predicted ones.

The findings highlight the importance of signal decomposition as a way of managing noise, Transformer-based attention as a way of modeling long-range dependencies, and large-language-model-based sentiment features as a way of capturing the behavioral and exogenous forces on the market. The hybrid design ensures better predictiveness as well as is more interpretable hence eliminating essential shortcomings of earlier designs.

The implications of the findings are significant both regarding the traders, who want to have more reliable means of operating in risky conditions, as well as financial institutions, which strive to enhance their decision-making quality in the uncertain world. The studies might examine the adjustment of SEET to new markets or high frequency trading data or other modalities like macroeconomic factors and alternative data in future research. Improvements to computing performance and real time implementation would enhance the scope of its usage.

REFERENCES

[1] A. Lendasse, E. de Bodt, V. Wertz, and M. Verleysen, "Non-linear financial time series forecasting - Application to the Bel 20 stock market index," *European Journal of Economic and Social Systems*, vol. 14, no. 1, pp. 81–91, 2000, doi: 10.1051/EJESS:2000110.

[2] R. Baidya and S. W. Lee, "Addressing the Non-Stationarity and Complexity of Time Series Data for Long-Term Forecasts," *Applied Sciences 2024, Vol. 14, Page 4436*, vol. 14, no. 11, p. 4436, May 2024, doi: 10.3390/APP14114436.

[3] "(PDF) The Performance of Hybrid ARIMA-GARCH Modeling and Forecasting Oil Price." Accessed: Jan. 22, 2026. [Online]. Available: https://www.researchgate.net/publication/325057346_The_Performance_of_Hybrid_ARIMA-GARCH_Modeling_and_Forecasting_Oil_Price

[4] C. Dritsaki, "The Performance of Hybrid ARIMA-GARCH Modeling and Forecasting Oil Price," *International Journal of Energy Economics and Policy*, vol. 8, no. 3, pp. 14–21, 2018, Accessed: Jan. 22, 2026. [Online]. Available: https://ideas.repec.org/a/eco/journ2/2018-03-3.html

[5] A. Sebastian and D. V. Tantia, "Multi-variate LSTM with attention mechanism for the Indian stock market," *International Journal of Information Management Data Insights*, vol. 5, no. 2, p. 100350, Dec. 2025, doi: 10.1016/J.JJIMEI.2025.100350.

[6] A. Farhadi, A. Zamanifar, A. Alipour, A. Taheri, and M. Asadolahi, "A Hybrid LSTM-GRU Model for Stock Price Prediction," *IEEE Access*, vol. 13, pp. 117594–117618, 2025, doi: 10.1109/ACCESS.2025.3586558.

[7] L. Sahai and A. Chauhan, "Federated Learning-Enabled Privacy-Preserving Analytics Framework for Multi-Cloud Data Environments," pp. 1–7, Nov. 2025, doi: 10.1109/icriset64803.2025.11251884.

[8] M. Saberironaghi, J. Ren, and A. Saberironaghi, "Stock Market Prediction Using Machine Learning and Deep Learning Techniques: A Review," *AppliedMath 2025, Vol. 5, Page 76*, vol. 5, no. 3, p. 76, Jun. 2025, doi: 10.3390/APPLIEDMATH5030076.

[9] M. Kumar and M. Thenmozhi, "Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models," *International Journal of Banking, Accounting and Finance*, vol. 5, no. 3, pp. 284–308, 2014, doi: 10.1504/IJBAAF.2014.064307.

[10] D. Muhammad, I. Ahmed, K. Naveed, and M. Bendechache, "An explainable deep learning approach for stock market trend prediction," *Heliyon*, vol. 10, no. 21, p. e40095, Nov. 2024, doi: 10.1016/J.HELIYON. 2024.E40095.

[11] B. Raufi, C. Finnegan, and L. Longo, "A Comparative Analysis of SHAP, LIME, ANCHORS, and DICE for Interpreting a Dense Neural Network in Credit Card Fraud Detection," *Communications in Computer and Information Science*, vol. 2156 CCIS, pp. 365–383, 2024, doi: 10.1007/978-3-031-63803-9_20.

[12] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," *International Review of Economics & Finance*, vol. 104, p. 104719, Dec. 2025, doi: 10.1016/J.IREF.2025.104719.

[13] A. Chauhan and L. Sahai, "Multimodal AI-Guided Resource Allocation System for Dynamic Cloud Data Workloads," pp. 1–7, Nov. 2025, doi: 10.1109/icriset64803.2025.11252489.

[14] H. Zhang, L. Feng, J. Wang, and N. Gao, "Development of technology predicting based on EEMD-GRU: An empirical study of aircraft assembly technology," *Expert Syst. Appl.*, vol. 246, p. 123208, Jul. 2024, doi: 10.1016/J.ESWA.2024.123208.

[15] Z. Liu, Z. Su, L. Shang, H. Sun, and B. Zhao, "An approach to stock price prediction based on improved EEMD and attention-enhanced BiLSTM," *Expert Syst. Appl.*, vol. 284, p. 127802, Jul. 2025, doi: 10.1016/J.ESWA.2025.127802.

[16] F. H. Sezgin, Ö. Algorabi, G. Sart, and M. Güler, "Hyperparameter-Optimized RNN, LSTM, and GRU Models for Airline Stock Price Prediction: A Comparative Study on THYAO and PGSUS," *Symmetry 2025, Vol. 17, Page 1905*, vol. 17, no. 11, p. 1905, Nov. 2025, doi: 10.3390/SYM17111905.

[17] L. Zhou *et al.*, "LLM-Augmented Linear Transformer–CNN for Enhanced Stock Price Prediction," *Mathematics 2025, Vol. 13, Page 487*, vol. 13, no. 3, p. 487, Jan. 2025, doi: 10.3390/MATH13030487.

[18] C. Chen, L. Xue, and W. Xing, "Research on Improved GRU-Based Stock Price Prediction Method," *Applied Sciences 2023, Vol. 13, Page 8813*, vol. 13, no. 15, p. 8813, Jul. 2023, doi: 10.3390/APP13158813.

[19] J. Zhang, Z. Zhang, and J. Wen, "A multifactor model using large language models and multimodal investor sentiment," *International Review of Economics & Finance*, vol. 102, p. 104281, Sep. 2025, doi: 10.1016/J.IREF.2025.104281.

[20] L. Xie, Z. Chen, and S. Yu, "Deep Convolutional Transformer Network for Stock Movement Prediction," *Electronics 2024, Vol. 13, Page 4225*, vol. 13, no. 21, p. 4225, Oct. 2024, doi: 10.3390/ELECTRONICS13214225.

.