Available Online: 24 October 2025 DOI: 10.54254/3029-0880/2025.28191

A review of stock price and volatility prediction based on random forest

Feifei Wang

College of Mathematics Science, Inner Mongolia Normal University, Hohhot, China

13327107103@163.com

Abstract. The intrinsic nonlinearity and non-stationarity of financial time series create major challenges for traditional forecasting approaches. By virtue of its strengths in non-parametric modeling and feature selection, Random Forest (RF) has developed into a key methodology in financial forecasting. Nevertheless, prior research has focused mainly on applications, with limited attention to methodological limitations. This paper explores the application of RF in stock price and volatility prediction, highlighting their strengths and identifying key challenges to facilitate progress in next-generation intelligent financial forecasting systems. Based on a literature review and comparative analysis, it synthesizes the application of RF in feature engineering, task definition, and model optimization, and proposes a "threefold challenge" framework encompassing theoretical memorylessness, applicational static nature, and practical complexity. The results indicate that RF is effective in addressing feature lags and integrating multi-source information like realized and implied volatility, yet it exhibits notable constraints in modeling long-term dependencies, responding to concept drift, and handling the costs of optimization and deployment. Future work may emphasize extending temporal memory through hybrid models with deep learning, enabling adaptive responses to market shifts, and improving transparency and trust with Explainable AI (XAI).

Keywords: Random Forest, stock price prediction, volatility forecasting, financial time series, feature selection

1. Introduction

The inherent nonlinearity, high noise, and non-stationarity of financial markets make them particularly challenging for traditional statistical models. Machine learning techniques, exemplified by Random Forest (RF), have emerged as a robust alternative, leveraging their strong non-parametric modeling capabilities [1]. Previous studies have thoroughly investigated its applications in feature engineering, model optimization, and the fusion of multi-source data [2]. Nonetheless, such applications often do not adequately tackle the inherent challenges that RF encounters in modeling dynamic financial time series. Therefore, this review aims to move beyond a simple enumeration of existing applications by establishing a critical analytical framework. This paper will conduct an in-depth analysis of the core advantages of RF in financial forecasting, such as its ability to handle complex feature interactions and temporal lags, and, for the first time, systematically identify the threefold methodological challenges it confronts: the theoretical "memoryless" nature, the applicational "static" nature (difficulty adapting to "concept drift"), and the practical complexity [3-5]. As such, this study explores emerging technical approaches designed to overcome these challenges, with a particular focus on hybrid models incorporating deep learning, the development of dynamic adaptive frameworks, and the integration of Explainable AI (XAI). Ultimately, it seeks to offer a comprehensive roadmap for advancing the field, contributing to the development of more accurate, robust, and reliable next-generation intelligent financial forecasting systems.

2. Principles and features of the random forest algorithm

2.1. Dual randomization and ensemble decision mechanisms

The robustness of Random Forest is not accidental but results directly from a powerful three-part mechanism comprising dual randomization and ensemble decision-making. This architecture is carefully constructed to mitigate the high variance and overfitting tendencies inherent in a single decision tree, with the first element involving randomization at the sample level. Through a technique known as bootstrap aggregating (or Bagging), the algorithm creates multiple, diverse training subsets by sampling with replacement from the original dataset. This ensures that each tree in the forest is trained on a slightly different

perspective of the data, which is crucial for decorrelating the trees and, consequently, reducing the overall variance of the final model. The second component is randomization at the feature level. When constructing each node of a decision tree, the algorithm does not consider all available features to find the optimal split. Instead, it searches for the best split within a randomly selected subset of features. This strategy is particularly effective for financial datasets, where technical indicators often exhibit high collinearity. It forces the model to explore a wider variety of predictive relationships and prevents it from repeatedly relying on a few dominant features, which is a key factor in mitigating overfitting. The final component is the ensemble decision process. After constructing a large "forest" of diverse and decorrelated decision trees, the model aggregates their individual predictions. For classification tasks, this is achieved through a majority vote, where the most frequently predicted class becomes the final output. For regression tasks, the predictions from all trees are averaged. This aggregation step smooths out the idiosyncratic errors and biases of individual trees. In synthesis, it is the synergistic effect of these three components, namely sample randomization, feature randomization, and ensemble aggregation, that endows the RF with its celebrated robustness and strong generalization performance, making it a formidable tool for noisy and high-dimensional data like financial time series [6].

2.2. Positioning and comparison in the modeling ecosystem

To precisely define the role of RF in financial forecasting, its positioning must be revealed through a systematic comparison with other mainstream methodologies, progressing from traditional statistical models to advanced deep learning architectures. First, when contrasted with traditional linear models such as ARIMA, the core advantage of RF becomes immediately apparent. While ARIMA is effective for series with clear linear trends and seasonality, RF excels at capturing the complex non-linear dependencies, feature interactions, and threshold effects that are pervasive in stock return data [7]. This allows it to model market dynamics that are invisible to linear specifications. Second, within the realm of conventional machine learning, RF distinguishes itself from models like Support Vector Machines (SVM) and shallow Artificial Neural Networks (ANN). RF generally offers a more straightforward training process, exhibits lower sensitivity to hyperparameter tuning, and possesses an intrinsic mechanism for resisting overfitting. Furthermore, its ability to provide intuitive feature importance metrics makes it a more transparent choice for many practical applications. Third, among its direct counterparts in ensemble tree methods, RF has a distinct profile compared to Gradient Boosting Machines such as XGBoost. RF utilizes Bagging to build independent, decorrelated trees in parallel, which grants it superior robustness to noisy data. Conversely, XGBoost employs Boosting to serially build trees that correct the errors of their predecessors, a process that often yields higher predictive accuracy on large, well-structured datasets but can be more susceptible to overfitting if not carefully tuned. Finally, the most fundamental limitation of Random Forest is highlighted when compared with deep learning models, particularly Long Short-Term Memory (LSTM) networks. RF is inherently memoryless, treating each data point as an independent sample. Consequently, it cannot explicitly model or leverage the long-range temporal dependencies that are a hallmark of financial time series. LSTM, with its recurrent architecture and memory cells, is specifically designed for this purpose [8]. The strength of Random Forest lies in its robust, non-linear modeling capabilities, which often outperform traditional linear and machine learning models in noisy, high-dimensional settings. However, its core limitation is the inability to explicitly model sequential dependencies, a domain where deep learning models like LSTM hold a definitive advantage. This positioning clearly defines RF's optimal use cases and illuminates the rationale for hybrid modeling.

3. Application of random forest in stock price prediction

The application of Random Forest in stock price prediction depends on effectively capturing the temporal information embedded in various data sources and translating it into accurate forecasts of future price movements [9]. This is primarily reflected in three aspects: feature construction, task definition, and model optimization.

3.1. Feature engineering: handling temporal correlation and multi-source information

Effective feature engineering is the cornerstone of successful stock price prediction, and its fundamental challenge lies in handling the temporal correlations, lags, and high dimensionality of financial data. The process involves constructing meaningful predictors from a vast array of potential sources, a task where Random Forest plays a pivotal role. The most basic feature source is technical indicators derived from historical price and volume. Indicators such as Moving Averages (MA) and the Relative Strength Index (RSI) are essentially different weighted smoothings of past information, designed to capture momentum, trend, and mean-reversion patterns. While valuable, these indicators can number in the hundreds, creating a high-dimensional feature space. This is where the intrinsic feature selection capability of Random Forest becomes invaluable. Faced with a vast feature set comprising numerous technical indicators, fundamental data, and other variables, RF's built-in importance evaluation mechanism (e.g., Gini impurity decrease or permutation importance) enables researchers to discern truly predictive signals from statistical noise, thus mitigating the risk of overfitting on irrelevant features [10]. A more sophisticated approach involves explicitly

constructing lagged features to allow the model to indirectly learn temporal dependencies. Although Random Forest does not directly model sequential data, it can effectively process time-series information when historical values are provided as distinct input features. For instance, by feeding the model the returns of the past 5, 10, and 20 days as separate columns, RF can non-linearly learn how this constellation of historical information collectively influences future returns. Furthermore, a cutting-edge direction involves integrating multi-source, unstructured data. Sentiment indices extracted from financial news or social media, for example, provide a proxy for market psychology. Incorporating sentiment scores with carefully chosen time lags allows models to capture information beyond price data, enhancing predictive power [11]. RF excels not by modeling time directly, but by managing the high-dimensional, noisy space created by temporal features. Its ability to select salient predictors from a wide range of technical, lagged, and alternative data makes it a robust tool for building sophisticated forecasting models.

3.2. Prediction tasks: from price regression to trend classification

In the context of stock price prediction, RF is primarily applied to two distinct but related tasks: the regression of continuous price values and the classification of directional trends. The first task is price regression, which aims to directly forecast a future stock price or, more commonly, its return. This is a highly challenging objective that demands exceptional precision from the model. The performance of regression models is typically evaluated using metrics such as the Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE), which quantify the average magnitude of the prediction errors. The second, and often more practical and robust task, is trend classification. This approach reframes the problem by simplifying the prediction to a directional outcome, such as binary (up/down) or ternary (up/down/flat) classification. The goal here is to maximize metrics like prediction accuracy or the F1-score. A critical aspect of this task is the model's sensitivity in capturing not just the continuation of trends, but also their potential reversals. It is in identifying these subtle trend inflection points that Random Forest's non-linear capabilities truly shine. Financial markets are characterized by a complex interplay of trending and mean-reverting behaviors. A superior forecasting model must therefore be adept at recognizing the weakening of an existing trend and providing early signals of a potential reversal. By virtue of its ability to define complex, non-linear decision boundaries in the feature space, Random Forest exhibits great potential in detecting the nuanced patterns that often precede these critical market turning points. In conclusion, irrespective of whether the task is regression or classification, a rigorous evaluation methodology is a non-negotiable prerequisite for valid results. This includes not only standard procedures like data cleaning and feature normalization but, most critically, the mandatory use of strict time-series cross-validation techniques (e.g., rolling window or Purged K-Fold). Such methods are essential to properly evaluate the model and to avoid the lookahead bias that can arise from the temporal dependencies inherent in financial data [12].

3.3. Performance enhancement: combining external optimization and hybrid modeling

To push the performance ceiling of RF in financial forecasting, researchers typically employ two major strategies: external optimization and hybrid modeling. The first strategy involves external optimization algorithms. This approach treats the selection of the best feature subset and the tuning of model hyperparameters as a complex search problem. Instead of manual or grid-based searches, metaheuristic algorithms such as Particle Swarm Optimization (PSO) or Genetic Algorithms (GA) are used to intelligently and efficiently navigate this vast strategy space. These optimizers systematically search for the combination of features and parameters that maximizes the model's predictive performance on validation data [2]. The second strategy focuses on constructing hybrid models, aiming to compensate for the inherent limitations of Random Forest by combining its strengths with those of other models. A prominent example is the vertical integration with deep learning models like LSTM, which fuses RF's powerful non-linear decision-making with LSTM's ability to capture long-range temporal dependencies. Another approach is horizontal integration, such as using RF as a strong base-learner within a Stacking ensemble framework, where its predictions are combined with those of other diverse models (e.g., SVM, XGBoost) to produce a more robust final forecast [13]. In essence, these two classes of strategies provide distinct but complementary avenues for improvement. External optimization enhances the model’s performance by improving the local search for the best possible configuration within the RF framework itself, while hybrid modeling expands the model's expressive power by integrating capabilities that RF inherently lacks. Together, they offer a powerful toolkit for overcoming the performance bottlenecks of Random Forest in complex financial applications.

4. The deepening role of random forest in stock volatility prediction

The essence of stock volatility prediction lies in quantifying and anticipating market risk, which differs substantially from price prediction in both task characteristics and feature engineering. In this domain, the application of RF has evolved from a simple predictive tool to an advanced analytical framework capable of deeply characterizing and leveraging volatility dynamics.

4.1. Volatility feature engineering: fusing multi-source information and characterizing non-linearity

Unlike price prediction, feature engineering for volatility forecasting focuses on quantifying market uncertainty. The main challenge is to integrate historical data with forward-looking market expectations to construct a comprehensive set of predictors. Historical information is primarily derived from high-frequency price data, from which Realized Volatility (RV) and its variants are computed. Advanced metrics, such as realized semivariances, are often included to capture the leverage effect, which refers to the tendency of volatility to increase following negative returns. The fundamental value of RV-based features lies in their ability to provide a precise, model-free, ex-post measure of realized volatility. This historical perspective is complemented by forward-looking information from the options market, primarily in the form of Implied Volatility (IV), which reflects the market's collective, ex-ante expectation of future volatility over a specific horizon. Incorporating both RV and IV as input features allows a model to synthesize past volatility patterns with current market sentiment [14]. Volatility series exhibit complex statistical properties, including heteroskedasticity and non-linearity, which pose challenges for traditional econometric models. Models such as those in the GARCH family must rely on pre-specified functional forms to capture these characteristics. In contrast, RF, as a data-driven, non-parametric method, can directly learn these complex dependency structures from the features. Its non-linear modeling capability allows it to effectively process rich, multi-source feature sets, seamlessly integrating historical and forward-looking information. This flexibility enables RF to inherently capture the non-linear dynamics of volatility, offering a significant advantage over models constrained by rigid parametric assumptions.

4.2. Dynamic modeling: capturing time variation and price jumps

Financial market volatility is inherently non-stationary and exhibits time-varying dynamics, with price jumps being a major source of sudden changes. Advanced modeling approaches often decompose volatility into continuous and jump components, forecasting them separately to capture these distinct patterns. For instance, jump components can be extracted from high-frequency data using techniques such as bipower variation and used as key input features for RF [15]. By distinguishing between smooth volatility from regular trading and abrupt jumps caused by major information shocks, RF can leverage these heterogeneous features to enhance predictive accuracy, identify the most informative signals, and account for complex, non-linear interactions that traditional models may overlook.This fine-grained decomposition makes RF particularly effective within the Heterogeneous Autoregressive (HAR) framework [16]. By incorporating both jump components and multi-period realized volatility into an RF-HAR model, RF not only captures the long-memory behavior of volatility but also responds sensitively to sudden price jumps. Its non-linear, data-driven structure enables the integration of diverse temporal and cross-sectional features, automatic selection of salient predictors, and adaptive modeling of intricate relationships between continuous and jump components. This combination of flexibility, robustness, and sensitivity gives RF a decisive advantage over traditional parametric models, allowing it to deliver more accurate, responsive, and reliable volatility forecasts across a wide range of market conditions.

4.3. Financial applications: from risk management to option pricing

The superior performance of RF in volatility forecasting has direct practical value in two key areas of quantitative finance: risk management and option pricing. In risk management, accurate volatility forecasts form the foundation for calculating critical measures such as Value at Risk (VaR) and Expected Shortfall (ES). Because RF can better capture the complex dynamics of volatility, particularly during periods of market stress, its forecasts yield more precise and responsive estimates of a portfolio’s risk exposure. This enhanced accuracy is crucial for financial institutions seeking to maintain adequate capital reserves and avoid underestimating potential losses in extreme market conditions [17]. In option pricing, volatility is the most critical and notoriously difficult parameter to estimate within classic frameworks such as the Black-Scholes model. By forecasting future realized volatility, RF provides a data-driven estimate that complements the market’s implied volatility. Comparing the model’s forecasts with implied volatility allows traders to identify potential mispricings, which can support sophisticated volatility arbitrage or more effective hedging strategies. This dual perspective offers a richer basis for decision-making than relying on a single source of information. As such, the application of RF in these domains reflects a broader methodological shift. By providing a robust, data-driven alternative to traditionally model-dependent parameters, RF is helping to move financial practice from a purely theory-driven paradigm toward a hybrid, data-centric approach to risk management and derivatives analysis.

5. Methodological bottlenecks and technical outlook

5.1. Methodological bottlenecks: the threefold challenges of RF in financial applications

Despite its strong performance, RF encounters three fundamental challenges in complex financial forecasting that limit its broader applicability. The first challenge lies in its theoretical assumption of independent and identically distributed observations, which prevents RF from explicitly capturing long-range dependencies or sequential patterns in time series. As a result, its ability to perform accurate long-horizon forecasting is inherently constrained compared with models designed for temporal dependencies, such as recurrent neural networks. The second challenge concerns its static nature. Once trained, RF cannot adapt to concept drift, referring to dynamic changes in the underlying data distribution that are common in highly non-stationary financial markets [3]. This rigidity can lead to significant declines in predictive performance, particularly when structural breaks or regime shifts occur. The third challenge involves computational complexity. Training and tuning RF models over large hyperparameter spaces can be resource-intensive, and inference on large-scale datasets may incur latency that limits applicability in time-sensitive contexts like high-frequency trading. Thus, these limitations highlight areas where RF's practical implementation in dynamic financial environments requires careful consideration or complementary strategies.

5.2. Future technical paths: strategies to address the threefold challenges

In response to the three major challenges identified for RF in financial forecasting, future technical development can be structured around three strategic directions: hybrid modeling, dynamic adaptation, and enhanced interpretability.To address the model's inherent limitation in capturing long-range dependencies, hybrid modeling has emerged as a key strategy. This approach combines RF's ability to handle complex, non-linear feature interactions with the sequential memory capabilities of deep learning models such as LSTM. By constructing integrated architectures such as RF-LSTM, the model can leverage RF's strength in processing heterogeneous features while simultaneously capturing long-term temporal dependencies, resulting in predictive performance that surpasses that of either component individually [4]. The second strategy addresses RF's static nature through dynamic adaptive frameworks. These frameworks move beyond the train-once, predict-forever approach and use incremental or online learning, enabling RF to update its structure continuously in response to changing market conditions and concept drift. This adaptability is key to maintaining accurate forecasts in highly non-stationary markets [5]. The third strategy addresses computational complexity and enhances model credibility through XAI and multi-modal data integration. Techniques such as SHAP and LIME provide transparency into RF's decision-making, while the incorporation of diverse data sources, including news sentiment, macroeconomic indicators, and alternative datasets, enriches the feature space and clarifies how different information jointly influences predictions. This combined approach improves both the robustness and interpretability of RF-based forecasting systems [18]. Collectively, these strategies illustrate a compelling trajectory for RF in finance. Rather than being supplanted, RF is evolving into a core component of next-generation intelligent financial analysis systems, strategically augmented with deep learning capabilities and enhanced interpretability to meet the demands of increasingly complex and dynamic markets.

6. Conclusion

This paper has examined the principles, applications, and limitations of RF in financial forecasting, with particular emphasis on stock price and volatility prediction. The analysis shows that RF's robustness arises from its dual randomization, which effectively handles high-dimensional and collinear features. Moreover, RF excels in feature engineering, assimilates multi-source data, and captures non-linear dynamics in prices and volatility, enabling applications ranging from trend classification to risk management and option pricing. However, its effectiveness in volatile markets is constrained by several factors: the memoryless nature of the model limits its ability to capture long-term dependencies; its fixed structure reduces responsiveness to concept drift; and its computational intensity poses challenges for large-scale optimization and real-time use. Future developments are likely to focus on integrated approaches rather than single-model optimization, where hybrid architectures incorporating deep learning capture temporal dependencies, adaptive and online learning frameworks address market non-stationarity, and multimodal data combined with explainable AI enhance interpretability and robustness. Together, these advances point toward a new generation of financial forecasting systems that are more precise, adaptive, transparent, and resilient to structural changes in financial markets.

References

[1] Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems with Applications, 156*, 113464.

[2] Kong, J. (2025). A Novel Hybrid Model for Accurate Stock Market Forecasting Based on PSO and RF. *Applied and Computational Engineering, 146*(1): 23-32.
[3] You, X., Zhang, M., Ding, D., Feng, F., & Huang, Y. (2021). Learning to learn the future: Modeling concept drifts in time series prediction. In *Proceedings of the Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2434-2443.
[4] Fozap, F. M. P. (2025). Hybrid Machine Learning Models for Long-Term Stock Market Forecasting: Integrating Technical Indicators. *Journal of Risk and Financial Management, 18*(4), 201.
[5] Qian, Y. (2025). An enhanced Transformer framework with incremental learning for online stock price prediction. *PLOS ONE, 20*(1), e0316955.
[6] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1301.
[7] Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access, 8*, 150199-150212.
[8] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005-2019. *Applied Soft Computing, 90*, 106181.
[9] Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies, 7*(2), 26.
[10] Daoud, M. B., Hamdi, M., Younes, R., & Oueldoubey, D. (2025). Optimized feature selection based on machine learning models for robust stock market prediction. *International Journal of Innovative Research and Scientific Studies, 8*(3), 5086-5099.
[11] Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing, 13*(7), 3433-3456.
[12] de Prado, M. L. (2020). *Machine learning for asset managers*. Cambridge University Press.
[13] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689-702.
[14] Kambouroudis, D. S., McMillan, D. G., & Tsakou, K. (2021). Forecasting realized volatility: The role of implied volatility, leverage effect, overnight returns, and volatility of realized volatility. *Journal of Futures Markets, 41*(10), 1618-1639.
[15] Liu, W., & Wang, M. (2019). Volatility estimation and jump testing via realized information variation. *Journal of Time Series Analysis, 40*(5), 753-787.
[16] Yuyan, G., Di, H., Yan, M., & Hongmin, Z. (2023). Realised volatility prediction of high-frequency data with jumps based on machine learning. *Connection Science, 35*(1), 2210265.
[17] Chen, H., Didisheim, A., & Scheidegger, S. (2021). *Deep structural estimation: With an application to option pricing*. arXiv. https://arxiv.org/abs/2102.09209
[18] Kumar, S. N. A. (2025). Explainable AI in financial forecasting using time series analysis. *International Journal for Research in Applied Science and Engineering Technology, 13*(4), 7155–7159.