Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER
PUBLISHER

Research Article

A Novel Hybrid Monte Carlo-Random Forest Framework for Improved Financial Predictions

Steve Karam

College of Engineering and Technology, American University of the Middle East, Egaila, 54200, Kuwait
E-mail: steve.karam@aum.edu.kw

Received: 15 December 2024; **Revised:** 14 March 2025; **Accepted:** 28 March 2025

Abstract: In an era of increasing market volatility, accurate predictive tools are critical to navigate financial uncertainties. This study introduces a novel hybrid ensemble model that integrates Random Forest Regressors (RFRs) with Monte Carlo (MC) simulations to enhance predictive accuracy and quantify uncertainties. The hybrid model combines RFRs' ability to capture complex patterns with MC simulations' strength in modeling random price movements and volatility patterns. Specifically, the model generates multiple price paths using MC simulations based on historical volatility and drift, while the RFR component provides robust predictions by aggregating decision trees trained on historical price data. The integration is achieved by combining the RFR predictions with the ensemble average of MC-simulated price paths, creating a hybrid output that balances deterministic and probabilistic insights. The model effectively captures volatility clustering by reflecting historical volatility patterns in simulated paths and models stochastic trends through random sampling of future price movements. When applied to Disney stock prices and the S&P 500 index, the model demonstrated significant improvements in predictive performance over traditional methods, with reductions in the Mean Absolute Error (MAE). The results underscore its ability to capture volatility clustering and stochastic trends, providing practical advantages for portfolio optimization and risk assessment. To address the computational trade-off, the model employs parallel processing and optimized simulation parameters, ensuring scalability for real-world applications. Although computationally intensive, the hybrid approach presents a reliable framework, paving the way for advances in financial forecasting methodologies.

***Keywords*:** hybrid ensemble model, Monte Carlo simulation, financial machine learning, market prediction algorithms, stock price forecasting, volatility clustering, stochastic trends, ensemble forecasting, uncertainty quantification, portfolio optimization, risk assessment

MSC: 91G70, 91G60, 62P05, 62M10, 62M20

1. Introduction

Financial markets are inherently volatile and complex, requiring advanced predictive models that can effectively capture both underlying trends and unpredictable fluctuations. While Random Forest Regressors (RFRs) excel at capturing non-linear relationships, they struggle with stochastic market behavior and uncertainty quantification. Specifically, RFRs struggle with uncertainty quantification, as they provide point estimates without offering probabilistic insights into

DOI: https://doi.org/10.37256/cm.6220256079

potential outcomes. This limitation is particularly problematic in financial forecasting, where understanding the range of possible future scenarios is critical for decision-making. Moreover, Random Forest Regressions (RFRs) exhibit a reduced capacity to model volatility clustering, specifically the observed tendency for periods of high volatility to be followed by more high volatility, and low volatility by low. This limitation, which affects the modeling of a fundamental trait of financial time series, requires the development of hybrid strategies. These strategies would aim to leverage the advantages of machine learning alongside the statistical rigor of probabilistic models.

To address these gaps, we propose a novel Hybrid Monte Carlo-Random Forest (HMC-RF) framework tailored for financial forecasting. The model integrates the predictive capabilities of random forests with the probabilistic strengths of Monte Carlo simulations, creating a comprehensive tool to capture complex market dynamics. MC simulations excel in modeling stochastic behavior and uncertainty quantification by generating multiple possible future price paths based on historical volatility and drift. This probabilistic approach complements the deterministic predictions of RFRs, which are trained to capture non-linear relationships in historical price data. By combining the ensemble average of MC-simulated price paths with the predictions from the Random Forest model, our hybrid framework provides a balanced output that integrates both deterministic and probabilistic insights. To validate its ability to address the challenges of volatility clustering and stochastic trends in real-world scenarios, we apply the HMC-RF model to two distinct datasets: Disney stock prices over a five-year period (2019-2024) and the S&P 500 index over a ten-year period (2010-2020). Disney's stock was selected due to its recent increased volatility, making it an ideal candidate to test the model's ability to handle company-specific fluctuations. The S&P 500 index, on the other hand, represents broader market trends and captures multiple market cycles, allowing us to demonstrate the generalizability of the model in different financial contexts. Together, these datasets provide a comprehensive evaluation of the model's ability to predict both individual stock movements and broader market dynamics.

Our findings indicate that the HMC-RF model outperforms traditional Random Forest models in terms of predictive accuracy and uncertainty quantification. Key performance metrics such as Mean Absolute Error (MAE) demonstrate significant improvements, with reductions in MAE observed for both Disney stock and the S&P 500 index. These improvements were confirmed to be statistically significant through rigorous hypothesis testing. The MAE metric is particularly important in financial forecasting because it directly measures the model's ability to minimize prediction errors, which is critical for practical applications like portfolio optimization and risk assessment. Furthermore, the model effectively captures volatility clustering, tail risks, and stochastic trends, offering practical benefits for investors and policy makers. To address computational trade-offs, the model employs parallel processing and optimized simulation parameters, ensuring scalability for real-world applications.

This study builds on recent advances in financial machine learning and hybrid modeling, such as those of Gu et al. [1] and Zhou et al. [2]. Although these works have explored various aspects of machine learning and probabilistic modeling, they often focus on deterministic or probabilistic approaches in isolation. In contrast, our proposed HMC-RF model integrates both methodologies, addressing a critical gap in the literature. Specifically, prior studies have not fully explored the combination of random forests with Monte Carlo simulations for stock market prediction, leaving room for methodological innovation. Our approach differs from existing hybrid models in its unique implementation: by modifying the Random Forest structure (e.g., selectively removing trees) and integrating MC simulations, we achieve superior performance while maintaining interpretability.

The remainder of this manuscript is organized as follows. Section 2 reviews the relevant literature and positions our research within the existing body of work. Section 3 provides an overview of Monte Carlo simulations and their applications in financial modeling. Section 4 describes the datasets and outlines the data preparation process, including the hybrid methodology. Section 5 presents the experimental results and discusses their implications. Section 6 concludes with practical insights, limitations, and directions for future research. Section 7 details the data availability and reproducibility guidelines.

2. Literature review

2.1 *Machine learning applications in financial markets*

Stock market prediction has long been a key focus of financial research, using methods ranging from traditional statistics to modern machine learning. Among these, random forest regressors have garnered significant attention because of their ability to capture complex relationships in high-dimensional data. As summarized in Table 1, studies such as Khaidem et al. [3] and Krauss et al. [4] highlight the strengths of random forests in financial forecasting, including their capacity to model non-linear dependencies and handle noisy data.

Table 1. Summary of random forest strengths and limitations

Category	Key points
Strengths	- Captures non-linear relationships (e.g., technical/macro variables). - Reduces overfitting via tree averaging. - Handles high-dimensional data efficiently.
Limitations	- Poor extrapolation to new market conditions. - Sensitive to noise/extreme events. - No probabilistic uncertainty quantification.

Note: Key studies supporting this summary include Khaidem et al. [3] and Krauss et al. [4]

The table highlights that random forests offer distinct advantages for financial forecasting, but face critical limitations in uncertainty quantification and adaptability to novel market conditions. These challenges motivate the hybrid approach proposed in this study.

Additional studies, have further explored the application of random forests in stock market prediction, incorporating sentiment analysis and technical indicators to improve predictive accuracy. These additional elements, such as sentiment scores derived from news articles or social networks, are typically integrated as input features within the random forest model. By enriching the feature space, these inputs help the model capture more nuanced relationships between external factors and stock price movements. For example, sentiment analysis provides insight into market psychology, which can significantly influence short-term price fluctuations. This integration demonstrates how supplementary data sources enhance the predictive performance of random forests without altering their core structure.

2.2 *Hybrid and ensemble approaches*

Despite their strengths, random forest regressors, like any model, have inherent limitations. Their deterministic nature may not fully account for the uncertainty and stochasticity of financial markets. Specifically, random forest regressors can be prone to overfitting, particularly with noisy financial data, potentially leading to poor out-of-sample performance [5]. Overfitting in financial data often arises due to the presence of noise, outliers, or extreme market events that distort the underlying patterns. For example, during periods of high volatility (measured in %), such as financial crises or flash crashes, the model may incorrectly learn spurious correlations from anomalous data points. Furthermore, random forests provide point estimates without quantifying the uncertainty associated with the prediction, which is crucial in financial decision making [6]. This limitation becomes particularly problematic when forecasting under uncertain conditions, where probabilistic insights are essential for assessing risk-reward trade-offs.

To address these challenges, researchers have explored hybrid and ensemble approaches to enhance predictive robustness. The key methodologies and their applications are summarized in Table 2 below.

Table 2. Summary of hybrid and ensemble approaches

Method	Description and application
Stacking ensembles	- Combines base models (e.g., random forests, gradient boosting, neural networks) with a meta-model. - Leverages diverse algorithm strengths (e.g., gradient boosting for sequential dependencies, neural networks for non-linearities). - Improves adaptability in dynamic financial markets [7].
Bayesian hybrid models	- Integrates Bayesian inference with machine learning for probabilistic forecasting. - Quantifies uncertainty in stock price predictions [1].

2.3 *Monte carlo simulations in finance*

Probabilistic modeling techniques, such as Monte Carlo simulations, are widely recognized for their ability to quantify uncertainty and model various scenarios [8]. This makes them particularly well-suited for financial applications, where understanding and managing risk are paramount. For example, Monte Carlo simulations have been successfully employed in portfolio optimization frameworks [9] and credit risk assessment [7], showcasing their utility in enhancing risk-return profiles and assessing uncertainties. However, Monte Carlo simulations alone may not capture the complex dependencies and non-linearities present in stock market data, potentially leading to inaccurate predictions [10]. One specific challenge is their reliance on assumptions about the underlying data distribution, such as normality or stationarity, which may not hold in real-world financial markets. For example, during periods of market stress, asset returns often exhibit fat tails and skewness, violating the assumptions of standard Monte Carlo models. Recent studies, such as Chen et al. [11], have highlighted these limitations, emphasizing the need for hybrid approaches that integrate machine learning techniques to better capture market dynamics.

Hybrid models specifically address the issue of uncertainty quantification by combining probabilistic modeling with machine learning approaches. For example, Bayesian methods and Monte Carlo integration have been used to incorporate prior knowledge and update predictions dynamically based on new data [1]. These techniques complement the strengths of random forests by providing probabilistic insights into potential outcomes, enabling more informed decision-making in uncertain environments.

2.4 *Integration of random forests and Monte Carlo simulations*

The combination of random forests and Monte Carlo simulations has shown promise in various domains, such as risk assessment and portfolio optimization. For example, Xiong et al. [7] employed a hybrid model leveraging random forests and Monte Carlo simulations to assess credit risk, demonstrating the potential of such ensemble approaches to enhance model performance. Similarly, Deng et al. [9] showed that a portfolio optimization framework integrating these methods generated diversified portfolios with improved risk-return profiles. Although the literature showcases the effectiveness of both random forest regressors and Monte Carlo simulations individually, there is limited research on their combined application for stock market prediction. This gap in the literature highlights the novelty of our proposed approach and underscores the need for further exploration of hybrid models that leverage the strengths of both techniques. For example, similar hybrid models have been successfully applied in fields such as healthcare and energy forecasting, where they have demonstrated superior performance in handling uncertainty and capturing nonlinear relationships. These successes suggest the broader applicability of integrating random forests with Monte Carlo simulations, paving the way for advancements in financial forecasting.

Our research aims to fill this gap by investigating the potential of integrating Monte Carlo simulations into a modified random forest regressor. This integration seeks to enhance the accuracy and reliability of stock market forecasts, addressing the stochastic nature of financial markets while maintaining robust predictive capabilities. Specifically, we propose a methodology that combines the ensemble average of Monte Carlo-simulated price paths with the predictions from the random forest model, creating a hybrid output that balances deterministic and probabilistic insights. This approach addresses existing gaps in financial forecasting by providing a comprehensive framework for uncertainty quantification and volatility modeling, beyond just enhancing accuracy and reliability.

3. Monte carlo simulations: principles and applications in finance

Monte Carlo (MC) simulations, rooted in probabilistic modeling, provide a versatile framework for simulating scenarios and quantifying uncertainties in financial markets. These simulations are particularly valuable for stock price prediction, as they enable risk management and informed decision making by capturing the stochastic nature of stock prices. Using historical data, MC simulations model key parameters such as volatility (the uncertainty in price movements) and drift (the expected return), which are critical to generate realistic forecasts.

A widely used model within MC simulations for stock price prediction is Geometric Brownian Motion (GBM). GBM is a mathematical model that describes the movement of stock prices over time. The key assumptions of GBM include:

• Normality of Returns: GBM assumes that the logarithmic returns of a stock are normally distributed. This implies that large price movements (both positive and negative) are less likely, which may not hold true during periods of market stress or extreme volatility.

• Constant Volatility: GBM assumes that volatility remains constant over time. However, in real-world financial markets, volatility is often time-varying and exhibits clustering, where periods of high volatility are followed by more high volatility, and vice versa.

• No Jumps or Discontinuities: GBM does not account for sudden jumps or discontinuities in stock prices, which can occur due to unexpected news, earnings announcements, or macroeconomic events.

• Independence of Returns: GBM assumes that returns are independent and identically distributed (i.i.d.), ignoring potential autocorrelations or dependencies in financial time series.

While GBM provides a simple and tractable framework for modeling stock prices, its limitations must be acknowledged:

• Fat Tails and Skewness: Real-world financial returns often exhibit fat tails and skewness, violating the assumption of normality. This can lead to underestimation of tail risks, such as market crashes or extreme gains.

• Time-Varying Volatility: The assumption of constant volatility fails to capture the dynamic nature of financial markets, where volatility can change rapidly in response to market conditions.

• Inability to Model Extreme Events: GBM struggles to account for rare but significant events, such as financial crises or geopolitical shocks, which can have profound impacts on stock prices.

• Lack of Mean Reversion: GBM assumes that prices follow a random walk without mean reversion, which may not align with observed market behavior in certain asset classes.

These limitations highlight the need for hybrid approaches that integrate GBM with machine learning techniques, such as random forests, to better capture the complexities of financial markets. By combining the strengths of both methodologies, our model addresses some of the shortcomings of GBM while leveraging its probabilistic foundation for uncertainty quantification.

To predict the daily return of a stock using GBM, the following formula can be used:

$$\text{Daily Return} = \text{Drift} - \frac{1}{2} \cdot \text{Volatility}^2$$

where:

1. Drift ($\boldsymbol{\mu}$): The expected return of the stock. It represents the historical direction of the rates of return.

2. Volatility ($\boldsymbol{\sigma}$): Historical volatility multiplied by a random standard normal variable. Volatility captures the uncertainty and randomness in the movements of stock prices.

The GBM model can be used to simulate the future price of a stock S_t at a given time t, given the initial stock price S_0, as follows:

$$S_t = S_0 \cdot e^{(\mu - \frac{1}{2}\sigma^2)\cdot t + \sigma \cdot Z},$$

where Z is a standard normal random variable.

In Monte Carlo simulations, this process is repeated numerous times with varying random values for Z. Each simulation generates a possible path for the stock price, and by analyzing the distribution of these paths, we gain insights into the stock's potential future behavior, including its expected value and associated uncertainty. The robustness of these predictions improves with a larger number of simulations, as this better captures the full range of possible outcomes and reduces sampling error, though this comes at the cost of increased computational overhead. This approach offers significant advantages, such as the ability to model extreme scenarios or tail risks, which are crucial for financial forecasting and risk management. A key consideration is striking the right balance between prediction accuracy and computational efficiency when determining the simulation scale for a given application.

4. Data and methodology

This section outlines the datasets, data preparation process, and hybrid modeling approach used in this study. We first describe the Disney stock and S&P 500 index datasets, highlighting their relevance to the research. Next, we detail the data preprocessing steps, including chronological splitting and feature engineering. Finally, we present the methodology, explaining the integration of Random Forest Regressors (RFRs) with Monte Carlo simulations and the evaluation metrics used to assess model performance.

Two separate datasets form the basis of this study. The first consists of historical prices for Disney stock, spanning a five-year period from May 28, 2019, to May 24, 2024 (measured in USD). Disney's stock was selected due to its recent increased volatility, making it an ideal candidate for testing the model's ability to handle company-specific fluctuations. The second dataset covers the S&P 500 index over a ten-year period (January 4, 2010, to December 31, 2020), representing broader market trends and capturing multiple market cycles. Together, these datasets demonstrate the generalizability of our proposed method across different financial contexts.

We selected different timeframes based on data availability and relevance. While five years provided sufficient Disney data for modeling its more recent volatility, a ten-year window for the S&P 500 better captured long-term fluctuations in the broader market.

4.1 *Dataset descriptions, preprocessing and splitting*

The Disney stock dataset includes daily records of opening, closing, high, low prices, trading volumes, and dates. Key statistics reveal moderate price fluctuations, with a mean closing price of $141.23 and a standard deviation of $15.67 over the study period. Further details are provided in Table 3.

Table 3. First 5 rows of daily Disney stock prices, including opening and closing prices, daily highs and lows, and trading volumes

Date	Open	High	Low	Volume	Close
2019-05-29	131.96	132.14	130.77	7,749,600	131.57
2019-05-30	131.88	132.67	131.33	5,274,000	132.19
2019-05-31	130.96	132.92	130.77	7,420,700	132.03
2019-06-03	132.02	132.94	131.49	7,901,400	132.47
2019-06-04	133.44	134.88	132.91	8,247,500	134.82

Similarly, the S&P 500 dataset offers daily open, high, low, close, and volume values. Summary statistics show a closing price range of \$1,073.87 to \$3,756.07, with a mean of \$2,455.72 and a standard deviation of 758.23. Daily trading volumes ranged from 1.20 billion to 5.60 billion shares, with an average of 3.20 billion shares. Further details are provided in Table 4.

Table 4. First 5 Rows of S&P 500 Price Dataset showcasing daily price ranges, closing prices, and trading volumes over a 10-year period

Date	Open	High	Low	Volume	Close
2010-01-04	1,116.56	1,133.87	1,116.56	3,991,400,000	1,132.99
2010-01-05	1,132.66	1,136.63	1,129.66	2,491,020,000	1,136.52
2010-01-06	1,135.71	1,139.19	1,133.95	4,972,660,000	1,137.14
2010-01-07	1,136.27	1,142.46	1,131.32	5,270,680,000	1,141.69
2010-01-08	1,140.52	1,145.39	1,136.22	4,389,590,000	1,144.98

For both datasets, the date columns were converted to datetime format, and the datasets were split into 80% training sets and 20% testing sets, with the chronological order strictly preserved. This ensures rigorous assessment of model generalization to future price predictions. Furthermore, this chronological split mirrors real-world conditions where future data cannot inform the past, ensuring that the evaluation of the models remains unbiased and realistic in their predictive capabilities.

4.2 *Evaluation metric*

Before creating our models, it's important to define the evaluation metric. To quantify predictive accuracy, we rely on the Mean Absolute Error (MAE), defined as

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|.$$

Where $\hat{y}_j$ is the predicted price, y_j is the original price, and n is the total number of samples in the test set.

The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's calculated as the average over the verification sample of the absolute differences between prediction and ac- tual observation where all individual differences have equal weight. With this in mind, we built our models and assessed their predictive accuracy using the MAE as our primary evaluation metric.

4.3 *Hybrid model construction*

We constructed a Random Forest Regressor (RFR) as our baseline model. To mitigate overfitting, we set a maximum tree depth of 3 (balancing model complexity and generalization) and optimized hyperparameters via 5-fold cross-validation (MAE metric). The choice of 50 trees (Disney) and 20 trees (S&P 500) was determined through grid search to balance performance and computational efficiency.

Model Cloning and Tree Pruning: Using Scikit-learn's clone function, we created two variants of the optimized RFR:

- RFR-hybrid-1: Retained the first 50% of trees (e.g., 25/50 trees for Disney, 10/20 for S&P 500).
- RFR-hybrid-2: Retained the latter 50% of trees (e.g., 25/50 trees for Disney, 10/20 for S&P 500).

This pruning strategy tests the impact of tree subset selection while maintaining hyperparameter consistency. The 50% removal was chosen as a starting point to evaluate performance sensitivity to tree count. Future work could explore alternative pruning strategies to determine the optimal balance between model complexity and predictive performance.

Monte Carlo Integration: We generated 25 (Disney) and 10 (S&P 500) Monte Carlo price paths, averaged into an ”Ensemble Average”. Final predictions combined:

- EHM1: Average of RFR-hybrid-1 and Ensemble Average.
- EHM2: Average of RFR-hybrid-2 and Ensemble Average.

The full algorithm is summarized below:

1. Train RFR with optimized hyperparameters (max depth $=3$, number of estimators $=50/20$).
2. Clone RFR and prune trees: Retain first 50% (RFR-hybrid-1) or latter 50% (RFR-hybrid-2).
3. Generate Monte Carlo price paths and compute Ensemble Average.
4. Combine hybrid models with Ensemble Average for final predictions.
5. Evaluate using MAE and Wilcoxon Signed-Rank Test.

5. Results

5.1 *Disney stock price prediction*

The table below presents the Mean Absolute Error (MAE) of the models: The performance of our hybrid models compared to the baseline Random Forest Regressor (RFR) for Disney stock price prediction is presented in Table 5. The original RFR achieved an MAE of 15.18, while our Ensemble Hybrid Models showed varying performance levels. EHM2 demonstrated superior performance with an MAE of 12.82, representing a significant improvement over the baseline model. Conversely, EHM1 showed slightly higher error with an MAE of 16.79.

Table 5. Comparison of predictive performance for Disney stock. MAE represents the average magnitude of prediction errors, where lower values indicate better performance

Model	MAE
Random Forest Regressor (RFR)	15.18
Ensemble Hybrid Model 1 (EHM1)	16.79
Ensemble Hybrid Model 2 (EHM2)	12.82

From Table 5, it’s clear that the Ensemble Hybrid Model 2 (EHM2) has a lower MAE of 12.82 compared to the original Random Forest Regressor (RFR) with a MAE of 15.18. This indicates that EHM2 outperforms RFR in terms of predictive accuracy for Disney stock. These results suggest that the modifications made in EHM2, specifically retaining the latter half of the Random Forest trees, contributed to its superior performance.

To validate that the superior performance of EHM2 was not due to chance, we conducted statistical hypothesis testing. Absolute errors were calculated for both EHM2 and the original RFR predictions. A visual inspection of the histogram of absolute errors for EHM2 (Figure 1) revealed a non-normal distribution. Consequently, we employed the Wilcoxon Signed-Rank Test, a non-parametric test ideal for comparing paired samples without assuming normality. This test is particularly robust for small sample sizes or skewed data.

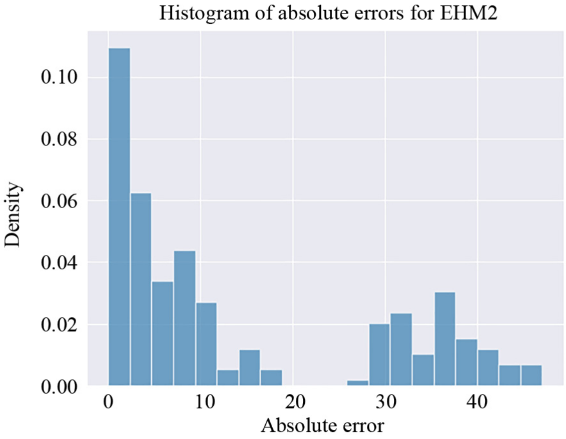

Figure 1. Histogram of absolute errors for model EHM2

We set our significance level at 0.05. The results of the Wilcoxon Signed-Rank Test yielded a test statistic of 7,022 and a p-value < 0.0001. Since the p-value was less than our significance level, we rejected the null hypothesis, indicating that the hybrid model provides significantly better prediction accuracy than the original RFR. This result reinforces our confidence in the effectiveness of our hybrid approach, as it demonstrates that the improvements are statistically significant and not due to random variation.

To visually demonstrate the predictive capabilities of our models, Figure 2 presents a comparison between actual Disney stock prices and the predictions generated by both the baseline Random Forest Regressor (RFR) and our best-performing Ensemble Hybrid Model 2 (EHM2). As illustrated in the figure, EHM2 tracks the actual price movements more closely than the baseline model, particularly during periods of price volatility. The superior performance of EHM2 is visually apparent, with predictions that more accurately capture both the trend direction and magnitude of price changes. This visualization complements our quantitative findings, providing clear evidence that our hybrid approach not only reduces error metrics but also generates predictions that better reflect the actual price behavior of Disney stock over time.

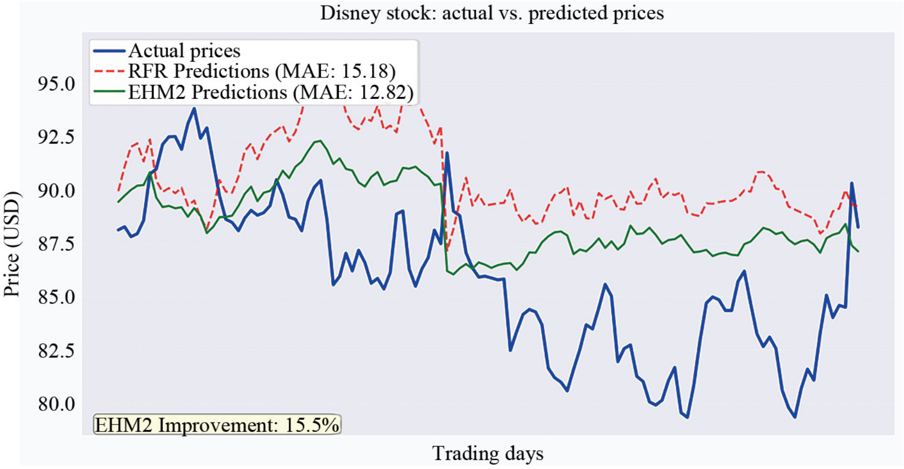

Figure 2. Comparison of actual Disney stock prices with predictions from the baseline Random Forest Regressor (RFR) and Ensemble Hybrid Model 2 (EHM2) over the test period. The plot demonstrates EHM2's superior ability to track actual price movements, especially during market fluctuations

5.2 *S&P500 index prediction*

The table below presents the Mean Absolute Error (MAE) of the models:

Table 6. Comparison of predictive performance for S&P 500 index. MAE represents the average magnitude of prediction errors, where lower values indicate better performance

Model	MAE
RFR	933.25
EHM1	1,134.06
EHM2	753.73

The results, presented in Table 6, show that EHM2 outperformed the baseline RFR. While the original RFR model recorded an MAE of 933.25, EHM2 achieved a significantly lower MAE of 753.73, reflecting its superior predictive accuracy. A Wilcoxon Signed-Rank Test validated the statistical significance of EHM2's improvement over RFR, underscoring the efficacy of the hybrid approach for modeling the S&P 500 index.

Figure 3 provides a visual representation of the actual S&P 500 index values compared to predictions from both the baseline RFR and EHM2 models. The visualization clearly illustrates the substantial improvement in predictive accuracy achieved by our hybrid approach. EHM2 predictions follow the actual index movements with remarkable precision, capturing both major trends and subtle price adjustments. This visual evidence reinforces our quantitative findings, demonstrating that the 19.2% reduction in MAE translates to practically valuable improvements in prediction quality that could benefit real-world financial decision-making processes.

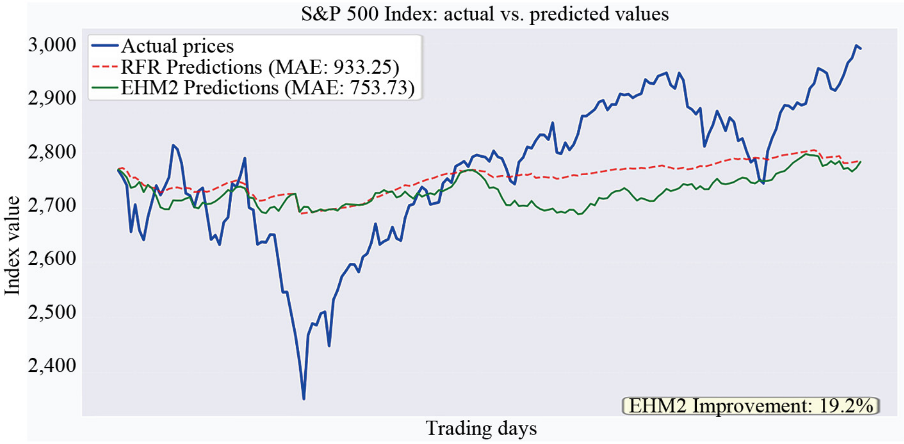

Figure 3. Comparison of actual S&P 500 index values with predictions from the baseline Random Forest Regressor (RFR) and Ensemble Hybrid Model 2 (EHM2) over the test period. The visualization highlights EHM2's significantly improved accuracy in tracking the market index compared to the baseline model

The choice of hyperparameters, such as limiting tree depth and optimizing the number of trees, was critical in achieving this performance. Monte Carlo simulations played a key role in enhancing the model's ability to capture volatility clustering and stochastic trends. By generating multiple possible future price paths based on historical volatility and drift, Monte Carlo simulations provided probabilistic insights that complemented the deterministic predictions of the Random Forest model. This integration allowed the hybrid model to better handle uncertainty and extreme scenarios, improving overall predictive accuracy.

Furthermore, EHM1 underperformed relative to both RFR and EHM2 because it retained only the first half of the Random Forest trees, which may have captured less relevant patterns compared to the latter half. This highlights the importance of selective tree retention in ensemble modeling.

The superior performance of EHM2 across both individual stocks and market indices suggests its robustness and adaptability to different market conditions and trading volumes. The model's ability to capture both short-term fluctuations and longer-term trends underscores the potential of hybrid ensemble models in advancing financial forecasting. These results validate our hybrid approach and demonstrate the value of integrating probabilistic modeling with machine learning techniques.

6. Conclusion, limitations, and future research

This study introduces a novel Hybrid Monte Carlo-Random Forest (HMC-RF) framework for financial forecasting, demonstrating significant improvements in predictive accuracy and uncertainty quantification over traditional models. By integrating random forests' non-linear pattern recognition with Monte Carlo simulations' probabilistic rigor, the framework addresses critical challenges in volatility clustering, tail risk estimation, and stochastic trend modeling. These advancements empower financial institutions to quantify risks in high-stakes scenarios such as market crashes, liquidity crises, and black swan events-with unprecedented precision. The model's interpretability and scalability make it a transformative tool for modern finance, enabling applications like dynamic portfolio rebalancing, real-time stress-testing aligned with Basel III requirements, and hedging strategies that adapt to evolving market regimes [7, 9].

Despite its strengths, the HMC-RF model is not without limitations. A primary challenge lies in its computational complexity, particularly due to the Monte Carlo simulations. As noted by Kelly and Xiu [12], the number of simulations and prediction horizon significantly impact computational overhead, potentially limiting real-time applications. Future work could address this issue by exploring variance reduction techniques or parallel computing implementations to enhance scalability. Additionally, the model's reliance on historical data patterns may hinder its ability to predict unprecedented market events or structural breaks, a limitation exacerbated by the increasing influence of social media sentiment and high-frequency trading in today's dynamic financial landscape [11]. To address this, integrating real-time data streams and adaptive learning mechanisms could enhance responsiveness to regime shifts.

To maximize the framework's impact, three focused avenues merit exploration. Hybrid deep learning architectures, such as transformer networks or graph neural networks, could be integrated with HMC-RF to better capture cross-asset dependencies and latent market states, enhancing multi-asset portfolio optimization [9]. Unstructured data integration through leveraging sentiment analysis from news, social media, and earnings calls via NLP techniques like BERT or GPT [13] thereby enriching feature sets and improving predictions during sentiment-driven market shifts. Finally, systemic risk modeling extensions could assess interconnected risks across financial institutions, supporting macroprudential policymaking and stress-testing of global banking systems, as emphasized in [7].

The HMC-RF framework represents a paradigm shift in financial risk management, bridging the gap between theoretical rigor and actionable insights. By enabling institutions to simulate complex risk scenarios with granular precision, it empowers stakeholders to preemptively mitigate systemic vulnerabilities, allocate capital efficiently, and comply with evolving regulatory demands. As financial markets become increasingly interconnected and data-intensive, hybrid models such as HMC-RF will be essential for managing uncertainty and enhancing the resilience of economic ecosystems. This study highlights the transformative potential of integrating machine learning and probabilistic methods, which serve as a cornerstone of next-generation financial engineering and align with the advancements of Chen et al. [11] and Li et al. [13].

7. Data availability

All datasets used in this research, including historical stock prices for Disney and the S&P 500 index, are publicly available on Yahoo Finance and are freely accessible.

Conflict of interest

The author declares no competing financial interest.

References

[1] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. *The Review of Financial Studies*. 2020; 33(5): 2223-2273.

[2] Zhou X, Zhou H, Long H. Forecasting the equity premium: Do deep neural network models work? *Modern Finance*. 2023; 1(1): 1-11. Available from: https://doi.org/10.61351/mf.v1i1.2.

[3] Khaidem L, Saha S, Dey SR. Predicting the direction of stock market prices using random forest. *arXiv:1605.00003*. 2016. Available from: https://doi.org/10.48550/arXiv.1605.00003.

[4] Krauss C, Do XA, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*. 2017; 259(2): 689-702. Available from: https://doi.org/10.1016/j.ejor.2016.10.031.

[5] Kroese DP, Taimre T, Botev ZI. *Handbook of Monte Carlo Methods*. USA: John Wiley and Sons; 2011.

[6] Hull JC. *Options, Futures, and Other Derivatives*. 10th ed. United Kingdom: Pearson Education; 2017.

[7] Xiong T, Shen C, Tan F. Credit risk evaluation based on random forests and Monte Carlo simulation. *Expert Systems with Applications*. 2019; 125: 44-54.

[8] Glasserman P. *Monte Carlo Methods in Financial Engineering, vol. 53*. Germany: Springer Science and Business Media; 2003.

[9] Deng X, Li Y, Wang S. Portfolio optimization via integrated random forest and monte carlo simulation. *Journal of Risk and Financial Management*. 2021; 14(11): 523.

[10] Wang G, Xie C, Chen Y. Stock price prediction based on a stacking ensemble learning model. *Journal of Computational Science*. 2022; 58: 101564.

[11] Chen L, Pelger M, Zhu J. Deep learning in asset pricing. *Management Science*. 2024; 70(2): 714-750. Available from: https://doi.org/10.1287/mnsc.2023.4695.

[12] Kelly BT, Xiu D. (2023). *Financial Machine Learning*. NBER Working Paper 31502. 2023. Available from: http://dx.doi.org/10.2139/ssrn.4501707.

[13] Li J, Wang Y, Wu C, Wang G. Stock price prediction based on sentiment analysis and random forest. *Expert Systems with Applications*. 2023; 212: 118789.