

**Utilizing stochastic differential equations and random forest for
precision forecasting in stock market dynamics**

P. Vidya Sagar *

Department of Computer Science and Engineering

Koneru Lakshmaiah Education Foundation

Vaddeswaram 522302

Andhra Pradesh

India

M. Rajyalaxmi [†]

School of Business

S R University

Warangal 506371

Telangana

India

A.V.V.S. Subbalakshmi [§]

Department of Commerce

School of Social Sciences and Languages

VIT University

Vellore 632014

Tamil Nadu

India

Sudhakar Sengan [‡]

Department of Computer Science and Engineering

PSN College of Engineering and Technology

Tirunelveli 627451

Tamil Nadu

India

* E-mail: vasudeva.nishal@gmail.com (Corresponding Author)

† E-mail: rajyalaxmi.m@sru.edu.in

§ E-mail: subbusravani76@gmail.com

‡ E-mail: sudhasengan@gmail.com (Corresponding Author)

Ravi Kumar Bommiseti [@]

*P G Department of Business Administration and
Dean - Research and Development
Maris Stella College
Vijayawada 520008
India*

Pankaj Dadheech [#]

*Department of Computer Science and Engineering
Swami Keshvanand Institute of Technology,
Management & Gramothan (SKIT)
Jaipur 302017
Rajasthan
India*

Abstract

The investigation for precision in Stock Market Forecasts (SMF) developments has led financial professionals to explore numerous modeling approaches. This paper investigates a new technique aimed at advancing the precision of Support Vector Machine (SMF) by combining the application of Random Forest (RF) methods with Stochastic Differential Equations (SDEs). Market deviations can be enhanced using the Geometric Brownian Motion (GBM) scheme, but this approach has problems because it denotes that market deviations will stay identical. To solve these challenges and develop how the GBM context more precisely corresponds to the unpredictable stock market factors, Random Forest (RF) is used for variable parameter prediction. The change in migration and deviation variables in GBM procedures has been defined using RF approaches in the present study using historical data from the stock market. It examines the performance of the RF-enhanced GBM method by comparing it to static variables in GBM, the Heston Model, and standard GBM while taking unpredictable chances into account. The quantitative metrics, such as the Sharpe Ratio (SR), Cumulative Returns (CR), and Maximum Drawdown (MD), were computed for a total of five stock markets in the research. When compared to the other models, the RF-enhanced GBM generally displayed a competitive edge in terms of risk-adjusted income and was highly accurate in terms of coming up with precise projections.

Subject Classification: 34A06.

Keywords: Stochastic differential equation, Random forest, Geometric Brownian motion, Accuracy, Machine learning, Stock market forecasting.

[@] E-mail: ravi9949418650@yahoo.com

[#] E-mail: pankajdadheech777@gmail.com

1. Introduction

Accurate forecasts about the stock market have the most significance in the financial services and economic advantages since seemingly insignificant improvements can produce significant profits. Primary hypotheses, including Geometric Brownian Motion (GBM), which acknowledges the unpredictable nature of stock prices, provide a framework for forecasting economic problems. The constantly changing and unpredictable features that define the market for business-related products are frequently in contradiction to the outcomes of GBM's market behaviour study findings, which depend on the theory of unchanging predictability. The employing of the Heston Model in order to compensate for deviations that have become predicted is a refinement of the static deviation hypothesis suggested by the GBM [1]. The method of enhancing the mathematical framework proves challenging and demands the efficient use of complicated resources and procedures, which could make them appropriate for precise predictions of the market for financial services. The objective of this research is to examine the methods in which Machine Learning (ML) might improve the predictive abilities of the parameters in stochastic differential equations (SDEs), with particular focus on Random Forest (RF) methods due to their reliability and their ability to handle non-linear relationships properly. The researchers have invented a technique that employs radio frequency (RF) to predict migration and unpredictability in the Generalised Bounded Model (GBM) by utilizing historical information on shares in the company. The contrast has been conducted between this RF-improved method, standard GBM methodologies, and other mathematical models based on SDE, such as the Heston Model (HM), which utilizes random variation [2]. A complete set of three different metrics—Maximum Drawdown (MD), Sharpe Ratio (SR), and Cumulative Returns (CR)—are implemented in order to assess the results of the stock market predictions. A large dataset has been employed in the study, and the primary emphasis is on how well the mathematical models can predict changes in the price of stocks. The paper is organized as follows: The background study of research for this work is discussed in Section 2, the knowledge of the SDE is recommended in Section 3, Section 4 presents the proposed methodology, Section 5 presents the performance analysis of the proposed model, and Section 6 conclusion and future work.

2. Literature Review

Recent financial modeling and prediction literature presents a range of innovative approaches, each addressing specific complexities inherent in financial markets. Hybrid stochastic differential games with Stackelberg differential subgames and non-zero-sum games are used to study reinsurance and investment strategies. In competitive markets, delay factors affect investment strategies [3]. A dynamic grey time-delay forecasting model is used to study energy supply, demand, and price variations. Time-delayed market responses show the model's crude oil futures price prediction skills [4]. [5] tackle the difficulties of parameter identification in chaotic financial and economic models. Their novel method estimates and predicts complex financial system time-varying parameters using Differential Evolution, Gaussian process regression, and neural networks. [6] respond to the surge in individual investments during the global pandemic by developing a model that combines GBM and Neural Networks. Their research results on SPF and inventory improvement are beneficial in the present state of the economy. The GBM system has been amended in [7] as a concluded approach. The study they conducted distinguishes out because it analyses SP predictions under different distributional hypotheses, demonstrating minor differences in prediction accuracy and how these presumptions affect market line predictions.

3. Introduction to Stochastic Processes

Random variables that are referenced by time are symbolized by a stochastic method, which is typically expressed as $\{X(t): t \geq 0\}$. The following steps have significance when attempting to define systems that evolve in an environment of uncertainty. For the market for financial goods, Brownian motion—sometimes termed as $B(t)$. The most significant characteristics of Brownian motion, which are additionally referred to as a Wiener process, are:

- It starts at zero, $B(0) = 0$.
- This method has autonomous iterations that were so motion in one time period has no impact of motion in any other non-overlapping time limit.
- The changes have a zero-sum average and a time-varying deviation; if not, they have a typical distribution. That is stating the distinction

$B(t) - B(s)$ adheres to a normal distribution $N(0, t-s)$ for any $0 \leq s < t$.

3.1 Stochastic Calculus and Ito's Lemma

The Stochastic Calculus broadens the scope of calculus to include random or unpredictable processes [8]. Central to this is the Wiener process, $W(t)$, a fundamental example of standard Brownian motion characterized by the condition $W(0)=0$. Ito's Lemma, a basic theory of stochastic calculus, is a model for random variations in various fields, particularly financial stock markets, describing the evolution of a differentiable function. Formally, for a twice continuously differentiable function $f(t, X)$, where $X(t)$ is an I to process, the differential df is given by EQU (1)

$$df(t, X) = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial X} dX + \frac{1}{2} \frac{\partial^2 f}{\partial X^2} (dX)^2 \quad (1)$$

This lemma is essential for deriving the dynamics of complex financial instruments.

3.2 Stochastic Differential Equations (SDEs)

SDEs are used to model systems that exhibit random behavior. The general form of an SDE is EQU (2)

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t \quad (2)$$

In this expression, X_t is the stochastic process of interest, $\mu(t, X_t)$ is the drift term representing systematic changes, and $\sigma(t, X_t)$ is the diffusion term representing random variations with W_t . As the Wiener process, a key application of SDEs in finance is modeling SP through GBM. The SDE represents the GBM, EQU (3)

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (3)$$

This model assumes that the percentage change in SP is usually distributed, with a drift term μ and a volatility term σ . The SDE solution for the SP at the time t' is EQU (4)

$$S_t = S_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t} \quad (4)$$

where S_0 is the initial SP. This model underpins the famous Black-Scholes formula for option pricing.

4. Methodology

4.1 Description of the Dataset

The study uses Quandl's [9] "Daily Historical Stock Prices" dataset for publicly traded companies on multiple global stock exchanges. The dataset shows opening and closing prices, highest and lowest prices, and total shares traded over several years [10]. In detail, it covers bull and bear markets, financial crises, and market stability.

4.2 Data Preparation and Preprocessing

A. Data Cleaning: Python's Pandas library is used to identify missing or anomalous data points using statistical methods like interpolation or exclusion [11]. The *isnull()* function combined with *sum()* helps quantify missing data. For sporadic missing values, the Pandas *interpolate()* method is employed, using time-series-based linear interpolation [12-14]. In cases where a massive portion of data is missing or if the data points are erroneous (e.g., SP that are non-positive), such entries are removed using the *dropna()* or conditional filtering methods in Pandas.

B. Data Normalization: This work uses *MinMax* scaling from Scikit-learn's *MinMaxScaler* to standardize data on varying scales of SP, ensuring it is suitable for comparative analysis without distorting differences in value ranges. This typically scales each feature to a range [0, 1]. For highly skewed data, a log transformation approach is adopted. This is particularly useful for SP, as they often follow a log-normal distribution. This paper uses NumPy's '*Log1p*' function, which applies $\text{Log}(1+x)$ to avoid issues with zero values. This study consistently applies normalization techniques across the entire dataset, ensuring comparability across the column-wise application.

4.3 Proposed SDE Model

The study uses the GBM model, a general financial modeling technique, to describe the behavior of SP, focusing on its mathematical basis that captures vital features of SP movements, including continuous evolution and random variations, and its mathematical EQU (5) is provided.

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (5)$$

Where, S_t denotes the stock's value at time t . The term μ is the drift coefficient, indicating the predicted rate of return for the stock. The term

σ is the volatility coefficient, meaning the standard deviation of the stock's returns. The dW_t term is the Wiener process, a mathematical representation of Brownian motion, which models the randomness inherent in market dynamics.

4.3.1 Estimation of Drift (μ)

Drift μ in the GBM model represents the expected rate of return of the stock. This parameter captures the average rate at which the SP's has historically grown (or declined). Given SP, S_t at time t and S_{t-1} at time $t-1$, the logarithmic return R_t is defined as, EQU (6)

$$R_t = \ln\left(\frac{S_t}{S_{t-1}}\right) \quad (6)$$

Logarithmic returns are preferred over simple returns because they are time-additive and more appropriate for analysis over multiple periods. To approximate μ , this work calculates the mean of these logarithmic returns. The mean represents the central tendency or predicted value of the returns. The EQU (7) for calculating the mean (drift) is:

$$\mu = \frac{1}{N} \sum_{t=1}^N R_t \quad (7)$$

Here, N signifies the aggregate count of computed returns. This predicts the probable rate of return on the stock, presuming that historical patterns remain uniform.

4.3.2 Estimation of Volatility (σ)

The parameter ' σ ' represents the unpredictability, measuring the standard deviation of the stock's returns, and is vital for assessing the risk involved in the SP's fluctuations. The determination of ' σ ' is based on computing the standard deviation of the logarithmic returns. The EQU (8) for computing volatility is:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{t=1}^N (R_t - \mu)^2} \quad (8)$$

In this equation, R_t are the logarithmic returns, μ is the mean of these returns (drift), $N-1$ in the denominator is used for an unbiased estimator of the population standard deviation when sampling from a population.

4.3.3 Monte Carlo Simulation (MCS) for Stock Price Estimation

The MCS is an effective method for predicting future SP within the context of the GBM model. It generates numerous potential future market price tracks for a stock using the statistical properties defined by the GBM model.

From Continuous to Discrete Model: To apply the GBM model in a MCS, it is necessary to convert the continuous model, as presented in equation [1], into a discrete form that can be used for numerical simulation. This transformation involves the following steps:

1. *Integration of the SDE:* Integrate the SDE over a small time interval from t to $t + \Delta t$, EQU (9)

$$\int_t^{t+\Delta t} dS_u = \int_t^{t+\Delta t} \mu S_u du + \int_t^{t+\Delta t} \sigma S_u dW_u \quad (9)$$

2. *Applying Ito's Lemma:* The solution to the SDE involves Ito's lemma, a key result in stochastic calculus. It states that if a process X_t follows a SDE, then a function $f(X_t, t)$ also follows a specific kind of SDE. For $f(S_t) = \ln(S_t)$, Ito's lemma yields EQU (10)

$$d\ln(S_t) = \left(\mu - \frac{1}{2}\sigma^2 \right) dt + \sigma dW_t \quad (10)$$

3. *Integrating over Δt :* Integrate both sides from t to $t + \Delta t$, EQU (11)

$$\ln(S_{t+\Delta t}) - \ln(S_t) = \left(\mu - \frac{1}{2}\sigma^2 \right) \Delta t + \sigma(W_{t+\Delta t} - W_t) \quad (11)$$

given that $W_{t+\Delta t} - W_t$ follows a normal distribution with a mean of 0 and a variance of Δt , it can be substituted with $\sqrt{\Delta t}Z_t$, where Z_t represents a standard normal random variable.

4. *Exponentiating Both Sides:* To revert to the SP, S_t , exponentiate both sides, EQU (12)

$$S_{t+\Delta t} = S_t \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}Z_t\right) \quad (12)$$

This final equation provides the discrete approximation of the GBM model for a small time step Δt . It combines the expected growth rate μ , the volatility σ , and a random shock represented by Z_t , thus encapsulating the stochastic nature of SP movements over time.

Algorithm: MCS Process for Stock Price Estimation

Input: Current SP, S_0 , drift μ , volatility σ , time step Δt , total number of steps N , number of simulations M

Output: Set of simulated SP at time $N \cdot \Delta t$

Procedure:

1 Initialize: Set the initial SP: $S_{current} = S_0$

2 For Each Simulation $i = 1$ to M :

Set $S_{start} = S_0$

For Each Time Step $j = 1$ to :

- Generate a random sample Z_j from a standard normal distribution.
- Calculate the SP for the next time step using the discrete GBM EQU (13):

$$S_{next} = S_{current} \cdot \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}Z_j\right) \quad (13)$$

- **Update** $S_{current} = S_{next}$
- **Store** the last price $S_{current}$ from each simulation.

3 Analysis: Analyze the set of final prices from all simulations to understand the SP's distribution and potential future scenarios.

4 End Procedure

4.4 Integration of RF for Enhancing GBM's Prediction Capability

This study enhances the capability of the GBM model in SMF dynamics by integrating a RF algorithm. This integration primarily focuses on optimizing the estimation of the GBM model's key parameters, drift (μ) and unpredictability (σ), utilizing the comprehensive "Daily Historical Stock Prices" dataset. Two distinct RF models, each set to predict a separate GBM variable, are implemented in the present investigation. One model is given to approximate the drift (μ), while the other aims at the unpredictability (σ). This study identified market developments and features through the examination of features of historical stock data, such as statistical measurements spanning rolling time limits and logarithmic return differentials. Standard approaches can fail to identify complex, non-linear phenomena in data pertaining to finances, but RF algorithms trained with these features could locate problems. For the objective of

accurately describing the unpredictable characteristics of SP, this degree of precision is absolutely essential. The standard parameters, μ , and σ , are replaced by their RF-predicted counterparts, denoted as $\hat{\mu}$ and $\hat{\sigma}$. This leads to a redefined GBM model, EQU (14)

$$dS_t = \hat{\mu}S_t dt + \hat{\sigma}S_t dW_t \quad (14)$$

Next, the simulation EQU (15) is adjusted to incorporate the RF-derived parameters,

$$S_{t+\Delta t} = S_t \exp\left(\left(\hat{\mu} - \frac{1}{2}\hat{\sigma}^2\right)\Delta t + \hat{\sigma}\sqrt{\Delta t}Z_t\right) \quad (15)$$

This modification improves the GBM model with ideas derived from modern machine learning methodologies while the fundamental concept of the GBM algorithm has been preserved. For the goal of predicting the drift (μ) and unpredictability (σ) variables of the GBM framework, the initial training of the RF networks was performed with an emphasis on accuracy and reliability. Training and validation datasets have been produced from the same dataset in order to promote comprehensive learning and the assessment of models in the research paper. The mathematical models have been trained to employ features such as logarithmic returns and rolling metrics from historical stock market information. The implementation of *k-fold cross-validation* has been implemented in order to avoid overfitting from occurring. Optimizing the accuracy of the models was successfully accomplished using hyperparameter tuning. The recommended RF algorithms were systematically tested and trained to ensure their precision and reliability.

5. Analysis

Researchers investigated how well the recommended RF-enhanced GBM system could predict SP. In this research, five stocks labeled A to E have been selected from the working dataset, evaluating significant metrics and similarities with various other models. The results of the regression model's precision have been verified with Mean Absolute Error (MAE) as well as Root Mean Squared Error (RMSE). The regression model's predictive accuracy was assessed with R-squared (R^2), and the investigation's feasibility was evaluated through the application of the values of the parameters from a review testing. The present review included comparing the performance of the recommended model with several distinct models, making it possible for a thorough evaluation.

Initially, the Standard Generalised Bayesian Method with Practical Parameter Calculation was employed as the standard method.

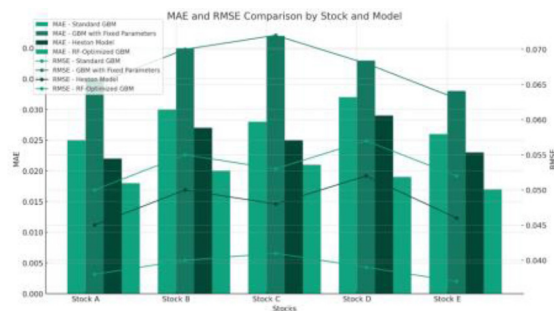


Figure 1
MAE and RMSE performance analysis

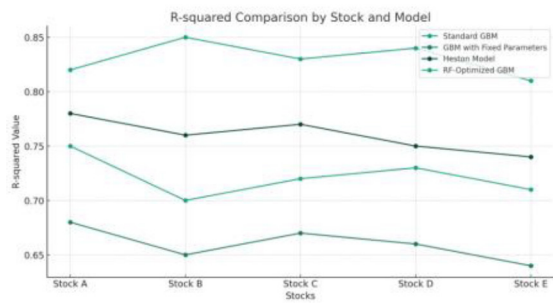


Figure 2
R² Analysis

To give users a point for assessing a system with dynamic parameters, the model that follows, the GBM with Fixed Parameters in MCS, has been implemented. An additional factor deemed was the Heston Model for Bayesian Uncertainty, a mathematical framework attempting to provide knowledge from an understanding of its unpredictable risk dimension. This comprehensive review of the RF-optimized GBM framework’s ability to analyze complex stock market phenomena using substantial datasets and unique models was reliable. The MAE and RMSE for each of the models, when performed in the stock market, are displayed in Figure 1, which enables an analysis of their predictive performance. As is the scenario with MAE, RMSE coefficients that are negative indicate an increased degree of precision. With the highest possible RMSE scores throughout all parameters, the RF-Optimized GBM model is significantly

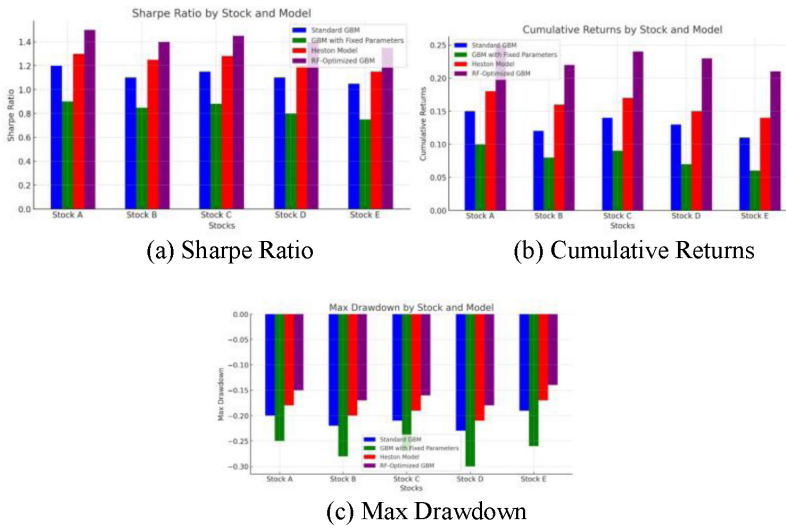


Figure 3

The performance measures of (a) CR, (b) SR, and (c) MD

superior to the other models when it comes to stock market predictions. The R^2 values, indicating how well each of the models represented the SP deviation, are presented in Figure 2. The more positive the R^2 value, which shows how powerful the model is at predicting the results that were obtained. The RF-Optimized GBM approach outperformed the other models when it comes to identifying and decoding SP variations, as demonstrated by the best R^2 value for all the stock markets. There is a comparison between the two versions of the three crucial back-testing measures in Figures 3 (a) to (c): CR, SR, and MD. Improved performance and risk avoidance have been shown by the RF-Optimized GBM model's improved CR and SR. The accuracy of the algorithms has been demonstrated successfully across an extensive number of previous economic conditions as an outcome of the investigation conducted in this study and decreased MD; the risks or losses caused during the back-testing are reduced.

6. Conclusion and Future Work

The research assessment indicates significant theory for improvement when integrating Machine Learning (ML) approaches, especially when using RF algorithms, with SDE predictions such as Geometric Brownian Motion (GBM) for SMF. After executing the current study, it has been

discovered that the RF-optimized GBM model is higher than the conventional SDE models in terms of the accuracy of forecasting and overall assessment of risk. The framework outperformed the Heston Model (HM), fixed-parameter models, and standard GBM when it comes to predicting complicated and unpredictable financial markets for stocks. The justification for this is that it continues to change its primary factors, which have been linked with migration and unpredictability, based on data from previous periods. The conclusions associated with the research indicate that the Random Forest (RF)-optimized algorithm enhances the accuracy of prediction across several types of stock parameters, such as Maximum Drawdown (MD), Cumulative Returns (CR), and Sharpe Ratio (SR). Additionally, the optimized algorithm generates a boost in profits balanced for risk. The significance of employing new data analytics for enhancing financial modeling has been demonstrated as the outcome of these findings from empirical studies, which has opened the door to the development of forecasting methods that are both more accurate and more dynamic. The research presented here contributes to the expanding repository of knowledge in quantitative finance and indicates how we are approaching an exciting development period concerning economic prediction methods.

References

- [1] Bai, Y., et al., A hybrid stochastic differential reinsurance and investment game with bounded memory. *European Journal of Operational Research*, 296(2), 717-737 (2022).
- [2] Duan, H., et al., A novel dynamic time-delay grey model of energy prices and its application in crude oil price forecasting. *Energy*, 251, 123968 (2022).
- [3] Wang, B., et al., Intelligent parameter identification and prediction of variable time fractional derivative and application in a symmetric chaotic financial system. *Chaos, Solitons & Fractals*, 154, 111590 (2022).
- [4] E. Kolářová and L. Brančík, Noise Influenced Transmission Line Model via Partial Stochastic Differential Equations, 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, pp. 492-495 (2019).
- [5] R. Jia et al., Stochastic Recursive Zero-Sum Differential Game and Mixed Zero-Sum Differential Game Problem with Payoff Functional

- in BDSDES, IEEE 3rd International Conference of Safe Production and Informatization, Chongqing City, China, pp. 337-342 (2020).
- [6] P. H. A. Ngoc, A New Approach to Mean Square Exponential Stability of Stochastic Functional Differential Equations, *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1645-1650 (2021).
 - [7] Aheer, A. K., et al., Application of Feedforward Neural Network in Portfolio Optimization and Geometric Brownian Motion in Stock Price Prediction. 4th International Conference on Electronics and Sustainable Communication Systems, pp. 1494-1503 (2023).
 - [8] E.T. Mensah et al., Simulating stock prices using the geometric Brownian motion model under normal and convoluted distributional assumptions. *Scientific African*, 19, e01556 (2023).
 - [9] W. Lu et al., A CNN-BiLSTM-AM method for stock price prediction, *Neural Computing and Applications*, vol. 33, pp. 4741-4753 (2021).
 - [10] Y. Ma, et al., Portfolio optimization with return prediction using deep learning and machine learning, *Expert Systems with Applications*, vol. 165, pp. 113973 (2021).
 - [11] S. Mukherjee et al., Stock market prediction using deep learning algorithms, *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 82-94 (2023).
 - [12] G. Sui and Y. Yu, Bayesian Contextual Bandits for Hyper Parameter Optimization, *IEEE Access*, vol. 8, pp. 42971-42979 (2020).
 - [13] S. Huckemann, T. Hotz and A. Munk, Intrinsic MANOVA for Riemannian Manifolds with an Application to Kendall's Space of Planar Shapes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 593-603 (2010).
 - [14] F. B. Mahmud et al., A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection, 7th International Conference on Data Science and Machine Learning Applications, Riyadh, Saudi Arabia, pp. 97-102 (2022).