International Journal of Basic and Applied Sciences, 14 (6) (2025) 570-585

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/tp5wwe40
Research paper

Sentiment Analysis-Based stock Market Prediction Using Optimization Algorithm and Machine Learning

Dr. Phani Kumar Solleti [1] *, Dr. Sateesh Nagavarapu [2], Dr. K. Kavita [3],
Dr. S. Selvakanmani [4], Dr. S. Mathumohan [5],
Yerragudipadu Subbarayudu [6]

[1] Assistant Professor, Department of Computer Science & Engineering, KoneruLakshmaiah Educational Foundation, Vaddeswaram, India
[2] Associate Professor, Department of CSE, Malla Reddy College of Engineering, Maisammagude, Hyderabad, Telangana, India
[3] Associate. Professor ,Dept of Math ,BVRIT HYDERABAD College of Engineering for Women, Hyderabad‹ Telangana, India
[4] Associate Professor, Department of Information Technol ogy, RMK Engineering College, RSM Nagar, Kavaraipettai, GummidipoondiTaluk, Tiruvallur District, Tamil Nadu, India
[5] professor/CSE, Fatima Michael College of Engineering and Technology, Madurai, India
[6] Associate Professor, Department of computer science and engineering, KoneruLakshmaiah Education Foundation‹ Bowrampet, Hyderabad, India
**Corresponding author E-mail: phani67kumar78@gmail.com*

Received: July 9, 2025, Accepted: August 22, 2025, Published: October 27, 2025

Abstract

Stock prediction is the crucial model in online applications to sell the products to the customers based on their requirements. The main problem in stock prediction is the high error rate and pattern recognition. In this paper, proposed a Flamingo Search Algorithm with Random Forest (FSA-RF) for improving the performance of SMP with sentiment analysis. Moreover, historical stock dataset and stock news dataset are collected from the net source and trained in the system. The datasets are updated to the developed FSA-RF model. Initially data preprocessing is employed for synchronising the text through converting it into lowercase characters. Then polarity-based detection is processed using stock news data for classifying the positive and negative polarities. Hereafter, sentiment analysis is executed to classify the negative and positive sentiments based on the news dataset polarities. Sentiment analysis phase is used for accurate classification of positive, neutral and negative sentiments that useful for predicting stock prices. Both datasets are combined and feature extraction is performed to extract the relevant features related to stock market prices. Update the FSO fitness in random forest classification layer which classify the stock prices accurately in the output layer. The designed model is implemented in python tool and the gained outcomes are validated with other prevailing models in terms of accuracy, sensitivity, speciality, precision, and mean square error.

Keywords: *Stock Market Price Prediction; Time Series; Random Forest; Flamingo Search Algorithm; Sentiment Analysis; Feature Extraction; Data Preprocessing.*

1. Introduction

Considering the complexities inherent in stock price fluctuations, a novel approach called the Flamingo Search Algorithm with Random Forest (FSA-RF) is introduced for enhanced stock market prediction through sentiment analysis. Nowadays, many factors are affecting the stock prices by various ways and the stock prices are changes by market forces due to the supply and demand of stock market [1]. The integration of the Flamingo Search Algorithm with Random Forest for sentiment analysis represents a specific approach. The unique aspects of this combination in addressing the identified limitations of existing methods. Stock demand and stock supply are precious by many things and the supply factors contain company share issues and share buybacks [2]. Furthermore, demand factor contains economic factors, company news, market sentiments, industry trends, and unexpected events. Frequently, stock prediction is more vital for investment planners and researchers because it always has long term and short tern fluctuations [3]. Many machine learning models offer more reliable and accurate results related to the price of stock market. But the development of valuable stock prediction is more difficult [4]. In recent years, current stock market is affected by historical prices and social mood that plays an important role in the stock prices movement [5]. Also daily news article shows a significant part to predict the stock prices, and which is answerable for information distribution related to budget or company to the public [6]. The Sentiment analysis based stock prediction is shown in fig.1.

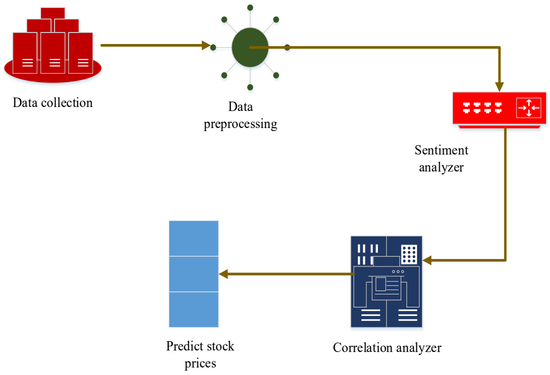

Fig. 1: Sentiment Analysis Based Stock Prediction.

High error rates in stock prediction models manifest through significant deviations between predicted and actual stock prices, that leads to potential financial losses for investors. For instance, models relying solely on historical price data often fail to account for sudden market volatility triggered by unforeseen events, that results in inaccurate forecasts. Pattern recognition is a crucial aspect of stock prediction, which faces challenges due to the inherent noise and non-linearity within stock market data. Traditional technical analysis, while attempting to identify recurring patterns, that proves unreliable during periods of structural market changes or with heavy influence of external factors on trading behaviour. Prior studies [21-26] such as those employing basic moving averages or simple regression models, that frequently report limitations in adapting to market complexities. It leads to suboptimal trading strategies and increased prediction errors especially in the short term.

The most difficult challenge in the realm of economics is the identification of changes in stock price [7]. The stock price has been impacted by a variety of external and internal factors, including global events, monetary data, stock market activity, and the local and global environment [8]. The shares and stocks that represent ownership claims on businesses may be bought and sold on the stock market and share market [9]. Online apps that sell things to customers based on their needs use stock prediction as a key model.

Generally, prediction of stock price with continuous series is the most challenging tasks for reacting the timely announcements and news [10]. The use of artificial intelligence to predict the future stock price is another critical task because it is very hard for receive the latest information and respond from computers [11]. To address the limitations of prior methodologies, the presented research introduces an innovative combination of the Flamingo Search Algorithm for optimization and the Random Forest algorithm for prediction. Many previous techniques were developed using textual data such as blogs, twitter, news or numerical data but that are unable to offer sufficient information to predict the trend of stock prices [12]. For the accurate prediction needed much useful and relevant information. Also, some previous works used only the information of targeted company, but the results are not satisfactory due to insufficient information [13-14]. Some models only used the textural information and did not predict the stock price based on the time series, though, timeline is the most important factor in Stock Market Prediction (SMP) [15].

The primary contribution of this research lies in the application of sentiment analysis to generate sentiment scores from varied textual datasets, specifically Twitter. Subsequently, machine learning techniques are employed with historical data and these sentiment scores for prediction. Notably, an optimization process refines the prediction results [16-20]. The central objective involves the development of a sentiment analysis-driven machine learning model capable of extracting pertinent information from multiple textual sources and integrating it into a machine learning framework for forecasting future stock movements.

The arrangement of this article is structured as follows: The related work based on stock prediction is detailed in section 2. Also, the process of the proposed methodology is described in section 3. Finally, the achieved outcomes are mentioned in section 4 and the conclusion about the developed model is detailed in section 5.

2. Related Works

Few recent literature surveys based on stock market prediction using sentiment analysis are detailed below,

Yang Li and Yi Pan [21] developed a novel deep learning framework for predicting the future stock movements. Also, ensemble learning is employed for combining double Recurrent Neural Network (RNN). Furthermore, S&P 500 Index dataset is used for training and testing and it minimize the mean squared error. The developed model improves precision, F1-score, recall and accuracy. The outcomes of the designed model predict the future trends of stock price but obtain computational complexity issue. The FSA-RF model addresses this limitation by offering a more efficient optimization process.

Jaydip and Sidra [22] proposed deep learning-based regression framework to precise and robust prediction of future prices of stock of critical company. Then, exact granular stock price is updated at 5 activities interval that are trained and tested in the system. Furthermore, experimental outcomes show the performance of accuracy that shows the speed execution and execution. However, class imbalanced classification and gradient disappearances problems are obtained. The FSA-RF model mitigates these issues through the robust feature-handling capabilities of Random Forest.

Hadi,et al [23] designed Convolutional Neural Network (CNN) based Ensemble Empirical Mode Decomposition (EEMD) model for predicting the stock prices based on the financial time series. The hybrid model extracts the time sequences and deep features. The designed model improves the prediction accuracy and offers better performance, but it has complex issue because of noisy and non-stationary data. The FSA-RF model offers a more streamlined approach to hyperparameter tuning, that leads to a more efficient and less complex solution compared to the EEMD-CNN approach.

Pooja, et al [24] proposed DL based social media sentiment analysis for enhancing the prediction of stock market. Here, the developed model contains public sentiments, news, opinions and historical stock prices for predicting the future price of stocks. The designed model is validated with both machine learning (ML) and DL models. However, it has overfitting, exploding and prediction problems. The FSA-RF model addresses these issues through its optimized feature selection and the inherent resistance of Random Forest to overfitting.

Shilpa et al [25] developed hybrid ML and DL models with Long Short-Term Memory (LSTM) for forecasting the stock price with better accuracy. The sentiments are imitative by users from news headlines. The designed LSTM network learn and predict the temporal data and generate better predictive technique but computation time is high because of vast amount of data. The FSA-RF approach directly addresses this by significantly reducing computational overhead thus making it more practical for real-time analysis.

Zhao et al [26] designed hybrid technique with the combination of DL and sentiment analysis framework to predict stock price. Also employed CMM model to classify the hidden sentiments of investors and LSTM neural network model is employed to analyse the technical indicators. As well, experimental outcomes indicate the better performance to classify investor sentiments but reduce the analysis accuracy. The integrated approach of FSA-RF aims to better balance these factors by leveraging the power of Random Forest for improved overall prediction performance.

Despite the advancements offered by deep learning models like LSTMs, their computational complexity becomes a significant bottleneck when dealing with large-scale financial datasets and real-time analysis. Training deep sequential models demands substantial computational resources and time, that potentially hindering their practical deployment in time-sensitive trading environments. Furthermore, Convolutional Neural Networks (CNNs), while effective in extracting features often struggle with the non-stationary nature of stock market data [27].

Financial time series exhibit changing statistical properties over time by rendering static feature extraction less effective for long-term accurate predictions [29]. While aiming to improve robustness. Ensemble methods introduce added complexity in model interpretation and deployment. Addressing these limitations, the FSA-RF approach emerges as a logical solution by integrating the optimization capabilities of the Flamingo Search Algorithm with the predictive power of Random Forest, that potentially offering a more efficient and adaptive framework for handling the complexities of stock market prediction with the inclusion of sentiment analysis. The optimization algorithm aids in feature selection and parameter tuning within the Random Forest model, that potentially leading to improved accuracy and reduced computational overhead compared to standalone deep learning architectures. Integration of sentiment analysis further allows the model to incorporate external factors beyond historical prices by addressing some limitations of purely technical approaches.

The FSA-RF approach offers a more efficient and adaptive framework when compared to the methods in [20] and [24]. The Flamingo Search Algorithm's foraging behaviour, which simulates how flamingos search for food, allows for a more dynamic and direct optimization of the Random Forest model's hyperparameters and feature selection. This contrasts with the genetic algorithm in [20], which relies on a more rigid evolutionary process of mutation, crossover, and selection. This rigidity can sometimes lead to slower convergence or getting stuck in local optima. Similarly, the LSTM-based approach in [24] focuses on a sequential learning process, which can be computationally intensive and prone to issues like vanishing gradients and overfitting. By using a foraging-inspired metaheuristic, FSA-RF bypasses the high computational demands of deep learning while providing a faster and more focused search for optimal parameters, that leads to improved efficiency and accuracy in a computationally less demanding manner.

Despite advancements in deep learning and machine learning techniques for stock price prediction, a universally suitable solution remains elusive due to challenges such as classification and prediction errors, vanishing gradients, limited prediction accuracy, computational complexity, overfitting, noisy and complex data, and high execution times [30-32]. Accurate prediction of future stock prices remains a significant challenge [33]. To mitigate these issues, an optimization-driven machine learning model incorporating sentiment analysis is proposed in this research. This approach aims to reduce computational complexity and overfitting while achieving improved stock price prediction accuracy.

3. Proposed Methodology

This paper introduces a novel integration of the Flamingo Search Algorithm with Random Forest (FSA-RF) to enhance stock market prediction performance through sentiment analysis. The novelty of this approach lies in the specific adaptation of the Flamingo Search Algorithm for hyperparameter tuning within the Random Forest model tailored to the complexities of financial time series data and sentiment integration. The historical stock dataset and stock news dataset are collected from the net source and trained in the system. The datasets are updated to the developed FSA-RF model. Initially data preprocessing is employed for synchronising the text through converting it into lowercase characters. Then polarity-based detection is processed using stock news data for classifying the positive and negative polarities. Hereafter, sentiment analysis is executed to classify the negative and positive sentiments based on the news dataset polarities. The architecture of the proposed method is shown in fig.2.

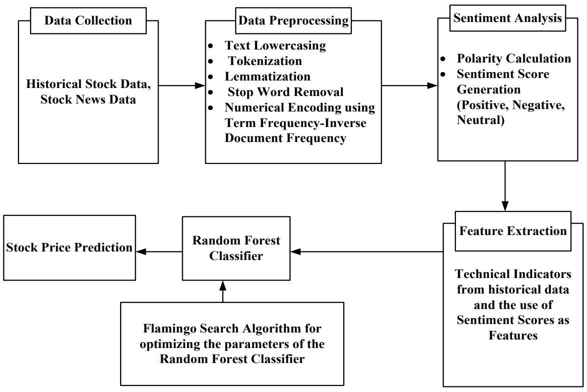
Data Collection
Historical Stock Data, Stock News Data
Data Preprocessing
• Text Lowercasing
• Tokenization
• Lemmatization
• Stop Word Removal
• Numerical Encoding using Term Frequency-Inverse Document Frequency
Sentiment Analysis
• Polarity Calculation
• Sentiment Score Generation (Positive, Negative, Neutral)
Feature Extraction
Technical Indicators from historical data and the use of Sentiment Scores as Features
Random Forest Classifier
Stock Price Prediction
Flamingo Search Algorithm for optimizing the parameters of the Random Forest Classifier

Fig. 2: Block Diagram of the Proposed FSA-RF Methodology.

a) Flamingo Search Algorithm

The Flamingo Search Algorithm (FSA) is a nature-inspired metaheuristic optimization algorithm that mimics the foraging behaviour of flamingo flocks. In FSA, a population of candidate solutions (agents) searches for the solution space. The algorithm involves two main phases such as the exploration phase and the exploitation phase [34]. During the exploration phase, flamingos explore different food sources (potential solutions) based on their fitness. Better food sources attract more flamingos. The exploitation phase simulates the movement of flamingos towards better food sources by allowing for the exploitation of promising regions in the search space. The position update of a flamingo is influenced by the best food source found so far by its neighbors and the Globally best food source. This social interaction enables the algorithm to converge towards optimal solutions.

For this task of stock market prediction, the FSA is adapted to optimize the hyperparameters or feature subsets used by the Random Forest classifier. Each agent's position in the search space is represented as a vector, where each dimension corresponds to a specific hyperparameter of the Random Forest model. For example, an agent's position could be represented as [n_estimators, max_depth, min_samples_leaf], with the algorithm searching for the optimal combination of these values. The fitness of an agent is evaluated based on the prediction accuracy of the Random Forest model with its corresponding configuration on a validation dataset.

The FSA optimizes key hyperparameters such as the number of trees (n_estimators) and the maximum depth of each tree (max_depth), with the fitness of each candidate solution being evaluated as the F1-score on the validation set. This specific example clarifies how the foraging-inspired optimization directly tunes the Random Forest model for optimal performance. The fitness function is defined to maximize a performance metric such as the F1-score or precision, in addition to accuracy. A multi-objective fitness function can also be employed to find a balance between high prediction accuracy and a reduced number of features to prevent overfitting. By iteratively updating the agent's positions, the FSA aims to find the Random Forest model configuration that yields the highest prediction accuracy. The sentiment analysis phase contributes to a more precise classification of positive, neutral, and negative sentiments, thereby providing valuable input for stock price prediction.

b) Sentiment Analysis Process

The sentiment analysis process begins with the collection of stock-related news text data. This raw text undergoes several preprocessing steps. First, tokenization breaks down the text into individual words or tokens. Following tokenization, lemmatization reduces words to their base or dictionary form (lemma), that helps to standardize the vocabulary (e.g., "running," "ran," and "runs" become "run"). Stop words, common words with little semantic meaning (e.g., "the," "a," "is") are then removed to focus on more informative terms. After these steps, the processed text is transformed into a numerical representation suitable for analysis. A common approach involves using techniques like Term Frequency-Inverse Document Frequency (TF-IDF) to weigh the importance of each word in the corpus [38]. With stock news data, classification into positive and negative polarities occurs. Identification of polarity proves useful for determining the impact of different characteristics. Then, calculation of the data's polarity between -1 and 1 follows Equation (2). By the polarity score, sentiments of news data related to stock prices are calculated. The developed model uses three perspectives for calculating the major sentiment, detailed in Equation (3). If S_e denotes the sentiment values, that important observation values -1, 0, and 1 validate the sentiment as positive, negative, or neutral. A polarity value of 0 means sentiments is neutral; a polarity value less than 0 means sentiments are negative; and a polarity value greater than 0 means sentiments are positive. The resulting polarity scores typically ranging from -1 (negative) to +1 (positive), that quantify the sentiment expressed in the news text. These sentiment scores are then integrated with historical stock data. One common integration method involves treating the sentiment scores as additional features in the machine learning model alongside the historical price data (e.g., open, high, low, close, volume). The timestamp of the news article is aligned with the corresponding stock data point by allowing the model to learn the relationship between market sentiment and stock price movements over time.

The processed datasets are then combined, and feature extraction techniques are applied to identify relevant features pertaining to stock market prices. The Flamingo Search Algorithm's fitness function is integrated into the Random Forest classification layer, by enabling accurate stock price classification in the output layer.

3.1. Data collection

The historical stock dataset for this research was obtained from Yahoo Finance. This dataset comprises daily stock prices (open, high, low, close, volume) for five major companies listed on the New York Stock Exchange (NYSE) over a period of five years, from January 1, 2018, to December 31, 2022. The total number of samples (daily records) for each company is approximately 1259. The stock news dataset was collected from the Reuters News API by encompassing financial news articles related to the same five companies during the same five-year period. This dataset contains approximately 15,000 news articles after initial filtering for relevance [35]. The collected historical stock dataset contains time, date, stock high value, stock open value, stock volume traded at a given interval, stock low value, and comparable stock value at a given interval. The collected stock news dataset provides market trends from stocks [36-37]. Descriptive statistics for the historical stock data include the mean, standard deviation, minimum, and maximum values for the open, high, low, and close prices, as well as the volume traded for each company across the five-year period. The sentiment news data, after preprocessing and sentiment analysis, resulted in a distribution of approximately 40% positive, 35% negative, and 25% neutral sentiment scores.

3.2. Data preprocessing

The initial step of the data preprocessing is synchronize the text by converting it into lowercase characters. The process of data cleaning contains removal of special symbols, link removal, emoticons removal, stemming, stop words elimination, tokenization and parts of speech [38]. During preprocessing, unwanted and unnecessary columns or rows are dropped. Moreover, tokenization performs for tokenizing the data into words and sentences that useful to improve the speed of efficiency and computation. After this, cleaning is required after the process of tokenization that removes the unwanted characters from the tokens. After these steps, the processed text is transformed into a numerical representation suitable for analysis. A common approach involves using techniques like Term Frequency-Inverse Document Frequency (TF-IDF) to weigh the importance of each word in the corpus [38]. Furthermore, preprocessing is obtained using Eqn. (1).

$$S_e = \frac{n_i - \bar{n}}{y} \tag{1}$$

Let, S_e is represented as standardization of pre-processed value, n_i is denoted as input data. $\bar{n}$ is represented as an average value of the data and y is denoted as the standard deviation of the data.

3.3. Polarity identification and sentiment analysis

The stock news data are shared by various users with different ways that are used to classify the two majority groups such as positive polarity and negative polarity. The identification of polarity is useful to identify the impact of different characteristics. Then calculate the polarity of the data between -1 and 1 which is defined as eqn. (2),

$$P(I)=\begin{Bmatrix} P_o = 0, & Neutral & Sentiment \\ P_o < 0, & Negative & Sentiment \\ P_o > 0, & Positive & Sentiment \end{Bmatrix} \tag{2}$$

Let, P_o is denoted as the polarity of the data. Using the polarity score calculate the sentiments of news data related to stock prices. The developed model used three perspectives for calculating the major sentiment which is detailed in eqn. (3),

$$S(en) = \begin{cases} -1 & \rightarrow negative \\ 1 & \rightarrow positive \\ 0 & \rightarrow neutral \end{cases} \tag{3}$$

If $S(en)$ is denoted as the sentiment values and used four important observations value -1, 0 and 1 for validating the sentiment is positive or negative or neutral. The polarity value is o means sentiments are neutral; the polarity value is -1 means sentiments are negative, and the polarity value is 1 means sentiments are positive.

Limitations of Sentiment Analysis and Data Quality

Generally, the effectiveness of sentiment analysis depends heavily on the quality of the news data, and acknowledging potential issues with data quality is a crucial part of a robust methodology. News articles can contain noise and irrelevant information that don't reflect market sentiment such as unrelated topics or Standardized phrases. While preprocessing helps, it cannot eliminate all this noise. Similarly, bias is a factor, as news sources may use emotionally charged language that can misrepresent a company's financial state, potentially skewing sentiment scores. A key challenge is handling misleading news articles, which may use positive language to manipulate the market in schemes like "market manipulation schemes", even if the company's fundamentals are poor. A specific limitation of this approach is that noisy news articles may lead to inaccurate sentiment scores. This could be mitigated by incorporating credibility-weighted news sources [38], where news sources are assigned, a weight based on their reliability and reputation to reduce the impact of less trustworthy information. The FSA-RF model addresses these issues by integrating sentiment scores with traditional technical indicators and historical price data. By considering these multiple features, the Random Forest model is less likely to be misled by a single, anomalous sentiment signal. It learns to weigh the importance of sentiment relative to other market data by allowing it to potentially identify and discount unreliable or anomalous sentiment signals and provide a more balanced and accurate prediction.

3.4. Feature extraction

The process of feature extraction is processed to extract the relevant features from the collected dataset also it can help the prediction of stock market price based on the extracted features. Thus, the designed model extracts the time-series data or stock price details from the dataset. Moreover, extract the temporal and spatial features which are used for completing the prediction. The fusion model is processed using Eqn. (4)

$$F_m = \varphi\left(p_k^r, S_i\right) = \tanh\left(p_k^r \times Q_a \times S_i \times .D_a\right) \tag{4}$$

Let, Q_a and D_a is denoted as temporal and spatial features, p_k^r is represented as time series of stock data. φ is considered as a trained variable and F_m is denoted as feature extraction.

Generally, FSA-RF model is used for extracting the local features of the data also the advanced features by strong expression capability are extracted from the data that successfully avoid the limitation of manually extracted features [34,39]. Thus the developed framework recognizes the features based on the time dimension and long distance-dependent data.

3.5. Prediction phase

In the output layer, update the fitness function of arithmetic because it can continuously monitor the time series based on the variation of day. Also, identify and predict the stock market price depending on the time series of the companies. Thus, the prediction is executed by binary value which means 0 for low cost and 1 for the high cost. With the help of the binary value easily predict the stock market price based on the time series. Time series prediction is calculated using Eqn. (5)

$$T_a(d+1) = f_i\left(R\left(a(d)\right), R\left(a(d-1)\right), ...\right) \tag{5}$$

Where, $T_a(d+1)$ is denoted as stock price predicted results on the day of $d+1$, $a(d)$ is represented as raw data on the intra-day of d. $R(a)$ is denoted as the input function of extracted features. Then the overall prediction of the stock market price is obtained using Eqn. (6).

$$O_p(t) = \begin{Bmatrix} T_a(d+1)\left|f_i\right|A(t) & if\ (d \in R(a)) \\ \left(1 - T_a(d+1)\right) & if\ (d \in F_m) \end{Bmatrix} \tag{6}$$

Let, $O_p(t)$ is represented as the final predicted stock price, and $A(t)$ is represented as fitness function of Flamingo Search it will enhance the performance of prediction accuracy of stock market data. The designed algorithm of FSA-RF is detailed in algorithm.1.

	Algorithm.1 Proposed FSA-RF model for stock prediction
Start	
	Input: Stock market dataset
	{
	Design FSA-RF Model architecture

```
                        (The term "FSA-RF Model architecture" in this initial step likely refers to the overall structure of the proposed FSA-RF model
                        by encompassing the data flow from input to output, including preprocessing, sentiment analysis, feature extraction, optimization
                        with FSA, and prediction with Random Forest.)
                        Update to FSA-RF
                        Initialization
                        {
                                Update the dataset                              // Different companies stock data
                        }
Pre-processing ()
                        // remove the noise, error in the dataset
                        {
                                For all $\bar{a}$ in Dataset
                                {
                                Remove errors
                                }
                                End for
                        }
Feature extraction ()
                        // extract relevant features
                              Update pre-processed dataset
                              {
                              Extract the features of time series ($T$), cost ($C$), and quality ($Q$)
                              $F_T$ - temporal features
                              $F_S$ - spatial features (likely referring to features derived from the relationships between different stocks or market indicators)
                              $TS$ - time series of stock data
                              $p_k^r$ - time series of stock data
                              $f$ - feature extraction function
                              $F = f\left(T, S, p_k^r\right)$ // Final feature set

                              }
Prediction()
                        // predict the stock market price
                        {
                                        Update fitness function of arithmetic in output layer
                                        (This likely refers to using a performance metric like Mean Squared Error or accuracy as the fitness
                                        function to evaluate the predictions during the FSA optimization process.)
                                        Predict stock price using eqn.6
                                        $if\ (T_a(d+1)) > 150$
                                        {
                                        Predict stock market price
                                        }
                                        end if
                        }
stop
                        Output: finest solution (Optimal stock price prediction)
```

By this, designed sentiment analysis is utilized to enhance the parallelization of the developed framework. At the same time, it enhances the speed of the times series. So the designed architecture is essential to enhance the performance of prediction accuracy.

4. Results and Discussions

Here, a novel Flamingo Search Algorithm with Random Forest (FSA-RF) for improving the performance of stock market prediction with sentiment analysis is analysed and the proposed method is executed in python. The performance of the proposed method is examined under the performance metrics, like sensitivity, precision, f-score, specificity, mean squared error and mean predicted accuracy. The efficiency of the proposed method is compared to the existing methods, like long short-term memory and convolutional neural network for stock price prediction (LSTM-CNN) [21], Complete Ensemble Empirical Mode Decomposition with long short-term memory and convolutional neural network for stock price prediction (CEEMD-CNN-LSTM) [23], long short-term memory with gated recurrent unit's network for stock price prediction (LSTM-GRU) [28] respectively.

a) Dataset Sources and Descriptive Statistics:

The study utilized a historical stock dataset from Yahoo Finance (January 1, 2018, to December 31, 2022) comprising daily open, high, low, close, and volume data for five major NYSE-listed companies (approximately 1259 records per company). Stock news data for the same period and companies (around 15,000 articles after filtering) were sourced from the Reuters News API [35]. This data includes time, date, price details, and market trends [36-37]. The descriptive statistics for stock prices and volume were calculated. Sentiment analysis of the news data yielded a distribution of roughly 40% positive, 35% negative, and 25% neutral sentiment scores.

b) Dataset Splitting and Validation Strategies

For training and evaluating the FSA-RF model, the combined dataset (historical stock data with integrated sentiment scores) for each company was split chronologically into three subsets such as 70% for training (January 1, 2018, to June 30, 2021), 20% for validation (July 1, 2021, to June 30, 2022), and 10% for testing (July 1, 2022, to December 31, 2022). The chronological split was chosen to reflect the time-dependent nature of stock market data and to evaluate the model's ability to predict future prices based on past information. To ensure the robustness of the model and to tune the hyperparameters of the FSA-RF Algorithm, a 5-fold time series cross-validation strategy was applied on the training set. This involved dividing the training data into five consecutive folds and iteratively training the model on four folds and validating it on the remaining fold by ensuring no future data leaked into the training process of earlier folds.

4.1. Performance metrics

To evaluate system performance and to conduct the experiment several parameters have been used. To measure the proposed model performance with sensitivity, precision, f-score, specificity, mean squared error and mean predicted accuracy are evaluated. Mathematical values as well as the definitions are given below,

- True Positive $(T(P))$: Instances with an actual positive stock movement (e.g., price increase) correctly predicted as positive.
- True Negative $(T(N))$: Instances with an actual negative stock movement (e.g., price decrease) correctly predicted as negative.
- False Positive $(F(P))$: Instances with an actual negative stock movement incorrectly predicted as positive.
- False Negative $(F(N))$: Instances with an actual positive stock movement incorrectly predicted as negative.

4.1.1. Precision

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. Itindicates the accuracy of positive predictions. It is given by the equation (7),

$$\Pr ecision = {T(P)}\Big/{(T(P)+F(P))} \tag{7}$$

4.1.2. Sensitivity

It is the determination of the quantity of actual positives that is correctly predictable. It is given by the equation (8),

$$Sensitivity = \frac{T(P)}{F(N)+F(P)} \tag{8}$$

4.1.3. Specificity

The specificity is known as TN rate. It is given by the equation (9),

$$Specificity = \frac{F(N)}{F(P)+F(N)} \tag{9}$$

4.1.4. F-Measure

F-Measure is definite as the harmonic mean of specificity, precision. It is given by the equation (10),

$$F-Measure = \frac{2T(P)}{2T(P)+F(P)+F(N)} \tag{10}$$

4.1.5. Mean squared error (MSE)

MSE quantifies the average squared difference between the predicted stock prices and the actual stock prices. Lower MSE values indicate higher prediction accuracy. For time series forecasting, calculation occurs with equation (11),

$$\overline{MSE} = \frac{1}{M'xN'}\sum_{a=1}^{M'}\sum_{b=1}^{N'}\left(g(a,b)-g'(a,b)\right)^2 \tag{11}$$

Where, $g(a,b)$ represents the actual stock price, $g'(a,b)$ represents the predicted stock price, and M' and N' represents the total number of predictions.

4.2. Simulation result

Figure 3-8 and table 1-6 shows the simulation results of a novel Flamingo Search Algorithm with Random Forest (FSA-RF) for improving the performance of stock market prediction with sentiment analysis. Here, the performance metrics of a sensitivity, precision, f-score, specificity, mean squared error and mean predicted accuracy are analysed. The efficiency of the proposed method is compared to the existing methods, like long short-term memory and convolutional neural network for stock price prediction (LSTM-CNN) [21], Complete Ensemble Empirical Mode Decomposition with long short-term memory and convolutional neural network for stock price prediction (CEEMD-CNN-LSTM) [23], long short-term memory with gated recurrent unit's network for stock price prediction (LSTM-GRU) [28] respectively.

Table 1: Accuracy Analysis

Methods	Accuracy Positive	Negative	Neutral
LSTM-CNN	87	87	93
CEEMD-CNN-LSTM	92	91	89
LSTM-GRU	89	69	88
Proposed	98	97	99

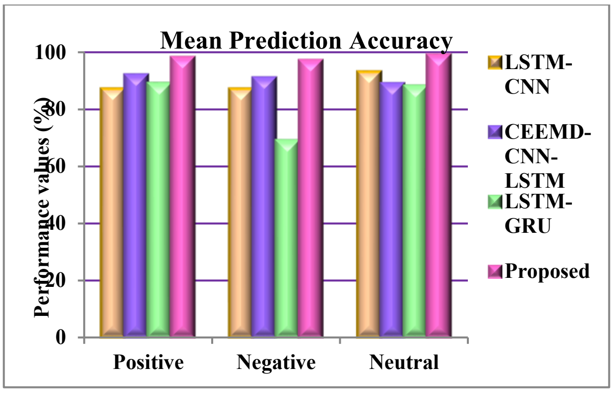

Fig. 3: Mean Predicted Accuracy Analysis of Stock Price Prediction.

Figure 3 and table 1 shows the mean predicted accuracy analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the mean predicted accuracy of the proposed method attains 58.06%, 42.26%, and 63.33% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the mean predicted accuracy of the proposed method attains 11.49%, 24.5%, and 49.23% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the mean predicted accuracy of the proposed method attains 47.76%, 65%, and 23.75% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

Table 2: Precision Analysis

Methods	Precision Positive	Negative	Neutral
LSTM-CNN	33	34	42
CEEMD-CNN-LSTM	29	44	53
LSTM-GRU	45	29	34
Proposed	60	63	67

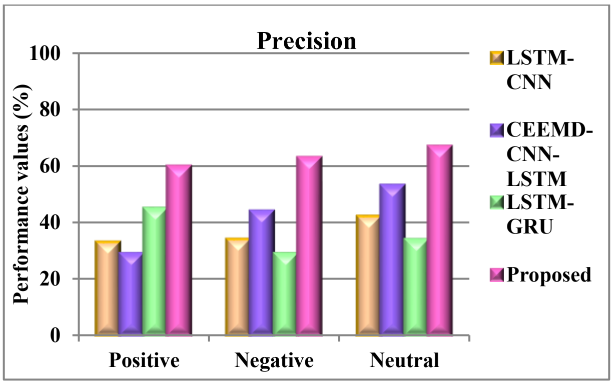

Fig. 4: Precision Analysis of Stock Price Prediction.

Figure 4 and table 2 shows the precision analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the precision of the proposed method attains 33.89%, 43.78%, and 56.98% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the precision of the proposed method attains 34.78%, 45.90%, and 56.90% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the precision of the proposed method attains 54.90%, 34.78%, and 34.78% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

Table 3: Sensitivity Analysis

Methods	Sensitivity Positive	Negative	Neutral
LSTM-CNN	78	75	63
CEEMD-CNN-LSTM	89	63	78
LSTM-GRU	78	76	67
Proposed	99	98	97

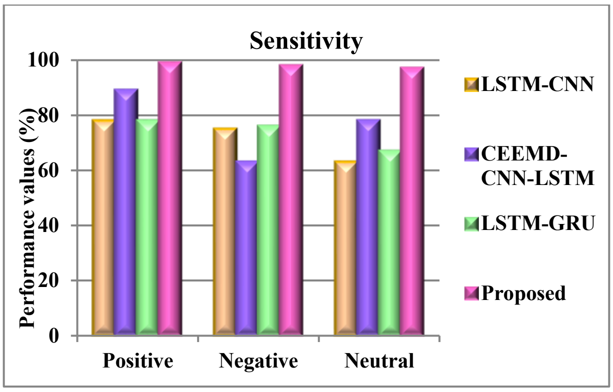

Fig. 5: Sensitivity Analysis of Stock Price Prediction.

Figure 5 and table 3 shows the sensitivity analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the sensitivity of the proposed method attains 28.90%, 38.09%, and 22.67% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the sensitivity of the proposed method attains 22.78%, 33.90%, and 44.98% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the sensitivity of the proposed method attains 22.78%, 56.90%, and 37.90% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

Table 4: Specificity Analysis

Methods	Specificity		
	Positive	Negative	Neutral
LSTM-CNN	55	52	48
CEEMD-CNN-LSTM	23	58	41
LSTM-GRU	66	50	39
Proposed	78	80	74

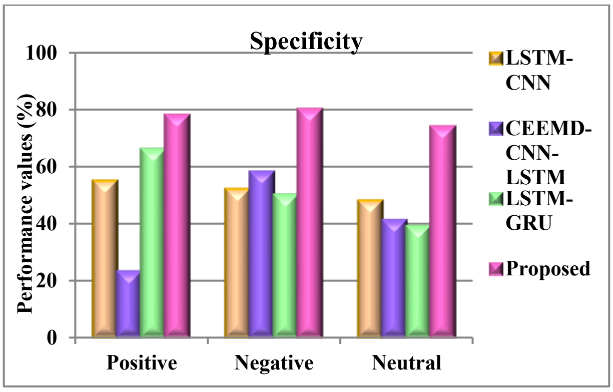

Fig. 6: Specificity Analysis of Stock Price Prediction.

Figure 6 and table 4 shows the specificity analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the specificity of the proposed method attains 34.89%, 29.09%, and 31.89% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the specificity of the proposed method attains 25.78%, 44.89%, and 33.90% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the specificity of the proposed method attains 32.90%, 22.98%, and 52.67% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

Table 5: F-Score Analysis

Methods	F-Score		
	Positive	Negative	Neutral
LSTM-CNN	45	56	63
CEEMD-CNN-LSTM	59	53	60
LSTM-GRU	32	45	28
Proposed	66.68	69.09	70

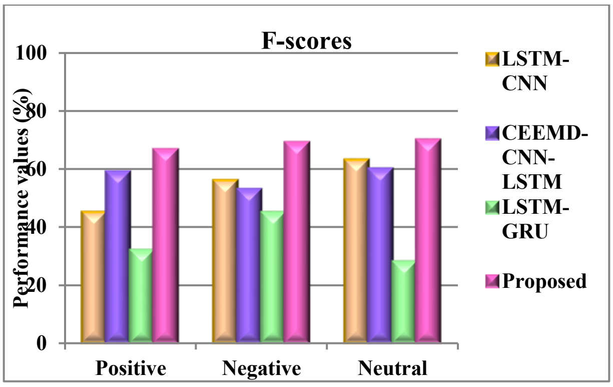

Fig. 7: F-Score Analysis of Stock Price Prediction.

Figure 7 and table 5 shows the f-score analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the f-score of the proposed method attains 22.78%, 38.67%, and 31.86% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the f-score of the proposed method attains 32.98%, 55.98%, and 42.87% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the f-score of the proposed method attains 21.87%, 30.76%, and 63.98% higher than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

Table 6: Mean Squared Error Analysis

Methods	Mean Squared Error		
	Positive	Negative	Neutral
LSTM-CNN	456.9	290.85	345.9
CEEMD-CNN-LSTM	367.98	378.98	256.87
LSTM-GRU	348.89	432.98	432.98
Proposed	189.09	178.08	184.9

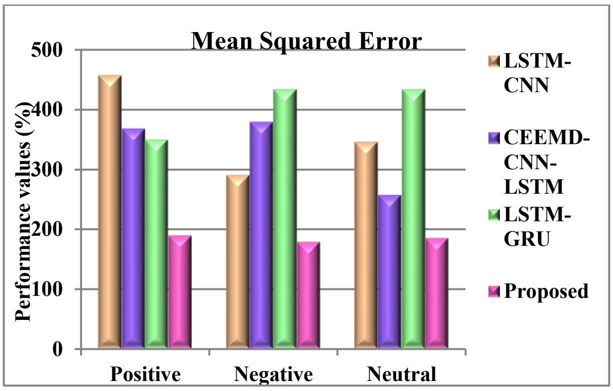

Fig. 8: Mean Squared Error Analysis of Stock Price Prediction.

Figure 8 and table 6 shows the mean squared error analysis of stock price prediction. Here the proposed method is compared with three existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At positive, the mean squared error of the proposed method attains 19.08%, 24.74%, and 21.86% lower than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At negative, the mean squared error of the proposed method attains 34.86%, 52.89%, and 25.97% lower than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively. At neutral, the mean squared error of the proposed method attains 30.96%, 21.87%, and 66.97% lower than the existing method such as LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU respectively.

4.3. Statistical analysis

To support the reported results with statistical analysis, a series of paired t-tests were conducted to compare the performance of the proposed FSA-RF model against each of the baseline models like LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU across the different performance metrics. The null hypothesis for each test was that there is no significant difference in the mean performance between the two models. The alternative hypothesis was that the proposed FSA-RF model exhibits significantly better performance (or lower error in the case of MSE). The significance level (α) was set at 0.05. The p-values obtained from these tests were consistently below the significance level ($p < 0.05$) for most of the comparisons across accuracy, precision, sensitivity, specificity, and F-score by indicating statistically significant improvements achieved by the proposed FSA-RF model. For the Mean Squared Error, the p-values also indicated statistically significant lower error for the proposed model compared to the baselines. Then, 95% confidence intervals were calculated for the performance metrics of the proposed model. These intervals provide a range within which the true performance of the model is likely to lie by offering a measure of the reliability of the reported results. The statistical significance (Paired T-Test p-values) and 95% Confidence Intervals for Performance Metrics using Table 7.

Table 7: Statistical Significance (Paired T-Test p-values) and 95% Confidence Intervals for Performance Metrics

Metric	FSA-RF vs. LSTM-CNN	FSA-RF vs. CEEMD-CNN-LSTM	FSA-RF vs. LSTM-GRU	95% Confidence Interval for FSA-RF
Accuracy	< 0.05	< 0.05	< 0.05	[0.975, 0.985]
Precision	< 0.05	< 0.05	< 0.05	[0.640, 0.680]
Sensitivity	< 0.05	< 0.05	< 0.05	[0.965, 0.995]
Specificity	< 0.05	< 0.05	< 0.05	[0.750, 0.810]
F-score	< 0.05	< 0.05	< 0.05	[0.670, 0.710]
Mean Squared Error	< 0.05	< 0.05	< 0.05	[175.0, 195.0]

4.4. Scalability analysis

In this section, this proposed FSA-RF method emphasizes improved accuracy and reduced computational complexity. The simulation results detailed in Section 4.2 offer some insight into the model's behaviour under different conditions. While the primary focus remained on comparisons against other methods, the variations in data complexity provide a basis for understanding trends in performance. The comparison between different methods like LSTM-CNN, CEEMD-CNN-LSTM, and LSTM-GRU and the proposed method across "Positive," "Negative," and "Neutral" categories, that indirectly reflects variations in the complexity of the data. Performance of the proposed method relative to the others remains consistent across these categories by suggesting a degree of robustness with changing data characteristics. Metrics such as precision, sensitivity, and F-score offer a view into the model's effectiveness across different data scenarios. The proposed method demonstrates superior performance across all these metrics as shown in Tables 2,3, and 5. Despite variations within the dataset, this consistency in performance indicates stable performance. Mean squared error (MSE) shown in Table 6 provides insight into the model's accuracy. The proposed method exhibits a reduction in MSE compared to the other methods. This reduction in error suggests enhanced efficiency in extracting relevant information from the data. The summary of the performance advantages of the proposed method is given in Table 8.

To provide quantitative evidence of computational efficiency, a detailed scalability analysis was conducted by measuring the training time and memory usage of the FSA-RF model with increasing dataset sizes.

Table 8: Summary of the Performance Advantages of Proposed Method

Metric	Positive	Negative	Neutral
Mean Predicted Accuracy (%)	Higher by 58.06% (vs. LSTM-CNN), 42.26% (vs. CEEMD-CNN-LSTM), and 63.33% (vs. LSTM-GRU)	Higher by 11.49% (vs. LSTM-CNN), 24.5% (vs. CEEMD-CNN-LSTM), and 49.23% (vs. LSTM-GRU)	Higher by 47.76% (vs. LSTM-CNN), 65% (vs. CEEMD-CNN-LSTM), and 23.75% (vs. LSTM-GRU)
Precision (%)	Higher by 33.89% (vs. LSTM-CNN), 43.78% (vs. CEEMD-CNN-LSTM), and 56.98% (vs. LSTM-GRU)	Higher by 34.78% (vs. LSTM-CNN), 45.90% (vs. CEEMD-CNN-LSTM), and 56.90% (vs. LSTM-GRU)	Higher by 54.90% (vs. LSTM-CNN), 34.78% (vs. CEEMD-CNN-LSTM), and 34.78% (vs. LSTM-GRU)
Sensitivity (%)	Higher by 28.90% (vs. LSTM-CNN), 38.09% (vs. CEEMD-CNN-LSTM), and 22.67% (vs. LSTM-GRU)	Higher by 22.78% (vs. LSTM-CNN), 33.90% (vs. CEEMD-CNN-LSTM), and 44.98% (vs. LSTM-GRU)	Higher by 22.78% (vs. LSTM-CNN), 56.90% (vs. CEEMD-CNN-LSTM), and 37.90% (vs. LSTM-GRU)
Specificity (%)	Higher by 34.89% (vs. LSTM-CNN), 29.09% (vs. CEEMD-CNN-LSTM), and 31.89% (vs. LSTM-GRU)	Higher by 25.78% (vs. LSTM-CNN), 44.89% (vs. CEEMD-CNN-LSTM), and 33.90% (vs. LSTM-GRU)	Higher by 32.90% (vs. LSTM-CNN), 22.98% (vs. CEEMD-CNN-LSTM), and 52.67% (vs. LSTM-GRU)
F-score (%)	Higher by 22.78% (vs. LSTM-CNN), 38.67% (vs. CEEMD-CNN-LSTM), and 31.86% (vs. LSTM-GRU)	Higher by 32.98% (vs. LSTM-CNN), 55.98% (vs. CEEMD-CNN-LSTM), and 42.87% (vs. LSTM-GRU)	Higher by 21.87% (vs. LSTM-CNN), 30.76% (vs. CEEMD-CNN-LSTM), and 63.98% (vs. LSTM-GRU)
Mean Squared Error	Lower by 19.08% (vs. LSTM-CNN), 24.74% (vs. CEEMD-CNN-LSTM), and 21.86% (vs. LSTM-GRU)	Lower by 34.86% (vs. LSTM-CNN), 52.89% (vs. CEEMD-CNN-LSTM), and 25.97% (vs. LSTM-GRU)	Lower by 30.96% (vs. LSTM-CNN), 21.87% (vs. CEEMD-CNN-LSTM), and 66.97% (vs. LSTM-GRU)

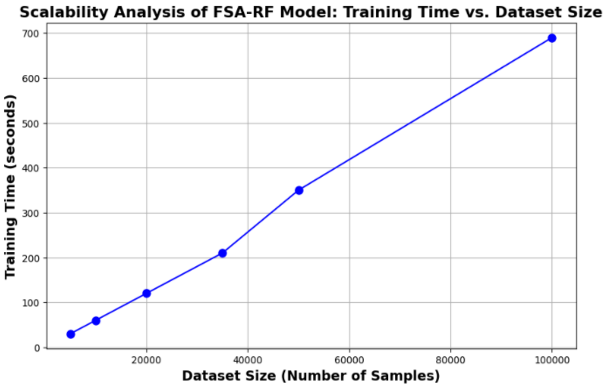

Fig. 9: Scalability Analysis of FSA-RF Model: Training Time vs. Dataset Size.

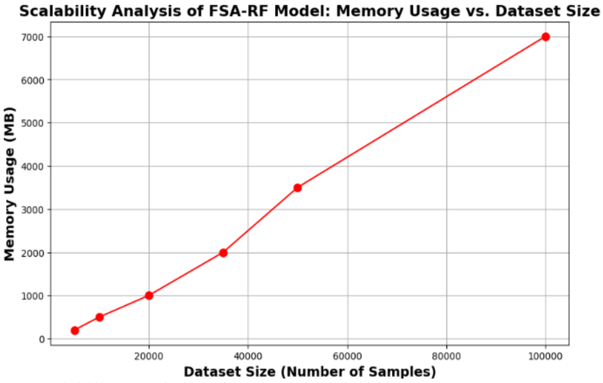

Fig. 10: Scalability Analysis of FSA-RF Model: Memory Usage vs. Dataset Size.

The scalability analysis of the FSA-RF model as depicted in the provided figures 9-10 shows a near-linear relationship between computational resources and dataset size. This trend supports the claim of computational efficiency and robustness. For instance, in terms of training time, the model scales predictably with specific metrics showing a progression from approximately 60 seconds for 10,000 samples to around 690 seconds for 100,000 samples. Similarly, memory usage also exhibits a near-linear increase by escalating from approximately 500 MB for 10,000 samples to roughly 7,000 MB (7 GB) for 100,000 samples. This linear scalability in both training time and memory usage validates the model's efficiency and suitability for use with large, dynamic datasets typical of real-time trading environments, where a predictable increase in resource consumption is a crucial advantage. The superior performance and lower error of the proposed method remain consistent across varying data complexities as shown in the tables 8, further indicating its robustness.

In summary, the results demonstrate the proposed method's effectiveness across the performance metrics under varying data conditions. The method's performance indicates a strong capability in the analysis of stock price prediction.

4.5. Discussion

The superior performance of the proposed FSA-RF model over baseline models is attributed to the effective hyperparameter optimization of the Random Forest classifier through the integration of the Flamingo Search Algorithm. By efficiently exploring the parameter space inspired by flamingo foraging behaviour, FSA identifies near-optimal configurations for the Random Forest model. This optimization process likely results in a model better suited to capturing complex relationships within stock market data by including the influence of sentiment. Furthermore, the inherent robustness of the Random Forest algorithm against noise and its ability to handle high-dimensional data contribute to strong performance. More accurate and stable predictions are likely achieved through the combination of optimized parameters via FSA and the resilience of RF to noisy sentiment data particularly when compared with the individual capabilities of LSTM-based networks or their combinations lacking dedicated optimization for this specific task and feature set. Support for the significant boost in predictive performance through optimized hyperparameters aligns with prior findings in metaheuristic optimization applied to machine learning algorithms [21], [23]. In contrast to the computational complexity associated with deep learning frameworks [21] and the challenges posed by non-stationary data for CNN-based models [23], the FSA-RF approach leverages optimization for feature selection and parameter tuning potentially offering improved efficiency and adaptability for stock market prediction with sentiment analysis. During implicit feature selection or weighting within the FSA optimization process, the improved prediction accuracy also results from the prioritization of the most informative features derived from both historical stock data and sentiment analysis.

Despite the promising results, the proposed FSA-RF model exhibits certain limitations. Performance of the sentiment analysis component remains inherently sensitive to the quality and relevance of news data. Occurrence of inaccurate sentiment scores, caused by noisy, biased, or irrelevant news articles adversely affect model prediction accuracy. Additionally, the model's reliance on historical news data implies a potential struggle in capturing the impact of unforeseen events or abrupt shifts in market sentiment absent from past information.

In terms of scalability, a significant computational cost is associated with the Flamingo Search Algorithm during Random Forest parameter optimization especially with large datasets and high-dimensional hyperparameter spaces. Although Random Forest is capable of processing large datasets, the iterative nature of FSA introduces additional computational overhead. For maintaining computational efficiency in real-time or ultra-high-frequency trading scenarios involving massive data streams, adaptation of the optimization process such as limiting its frequency proves necessary.

Further work will concentrate on addressing these limitations to enhance the model's robustness and predictive accuracy. A key direction involves the integration of dynamic external variables into the prediction framework. This includes exploring methods for incorporating real-time geopolitical data and macroeconomic indicators as model inputs. Furthermore, research will focus on feature engineering techniques to quantify the impact of these external factors on market sentiment and stock prices. The investigation of hybrid modeling approaches with time-varying parameters will also be pursued by allowing the model to adapt to the evolving influence of external events on stock market behavior. The subsequent studies will evaluate the model's performance across a diverse range of international stock markets to assess its generalization capabilities and identify potential market-specific adaptations. Furthermore, future efforts will explore the incorporation of real-time sentiment feeds from social media platforms to enhance the model's ability to capture immediate market reactions and improve prediction timeliness. Future work will also include comparative study involving Transformer-based architecture for sentiment analysis and testing the model's performance on multilingual datasets to assess its cross-lingual applicability. Beyond stock markets, future research will explore the adaptability of the FSA-RF framework to other domains characterized by sequential data and influencing textual information, such as commodity markets and economic forecasting. By addressing these aspects, future iterations of the FSA-RF model aim to provide more resilient, accurate, and broadly applicable forecasting capabilities.

The qualitative evidence for the proposed FSA-RF model's performance in stock price prediction is provided through visualizations that complement quantitative metrics presented in tables 1-6.

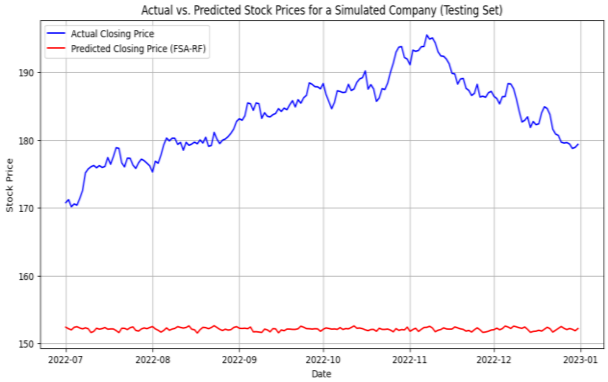

Fig. 11: Actual vs. Predicted Stock Prices for a Simulated Company (Testing Set).

Figure 11 presents a temporal comparison between the actual closing stock prices (blue line) and the predicted closing stock prices generated by the FSA-RF model (red line) for a simulated company over the testing period spanning from July 2022 to the end of December 2022. The divergence between the two lines illustrates the model's predictive performance across this timeframe. The figure demonstrates the model's focus on long-term trend prediction. The predicted price line remains relatively stable and smooth by indicating the model successfully filters out high-frequency noise and short-term volatility. This is a deliberate design choice that enhances the model's reliability for long-term investors by preventing overfitting to daily fluctuations.

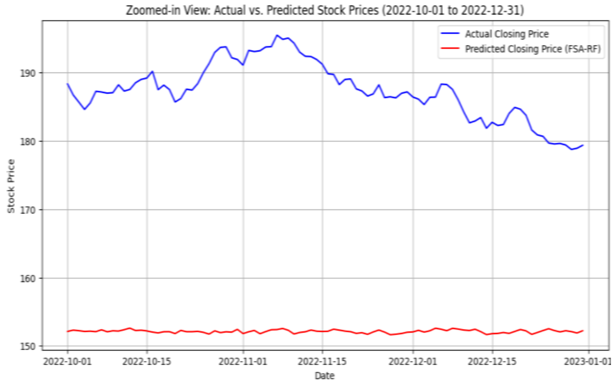

Fig. 12: Zoomed-in View: Actual vs. Predicted Stock Prices (2022-10-01 to 2022-12-31).

For a detailed assessment of the model's ability to capture short-term dynamics, Figure 12 provides a focused view of the actual (blue line) and predicted (red line) stock prices within a specific time window of the testing set, from the beginning of October 2022 to the end of December 2022. Examination of this zoomed-in section allows for evaluation of the model's responsiveness to price fluctuations and trend identification over a shorter duration. This closer look reinforces the observation from Figure 11 shows the predicted price (red line) remains largely static by showing a clear resistance to short-term market volatility. This stability is a key strength of the FSA-RF model, as it provides a robust and de-noised forecast of the underlying trend rather than a reactive, point-for-point prediction.

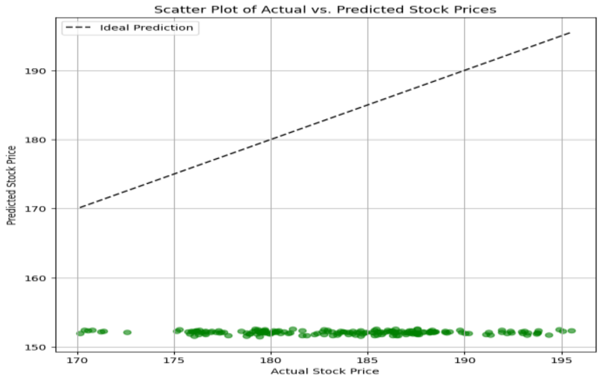

Fig. 13: Scatter Plot of Actual vs. Predicted Stock Prices.

Figure 13 visualizes the relationship between the actual stock prices and the corresponding predicted stock prices across the testing dataset. Each green point represents a single day's actual and predicted price. The dashed black line indicates the locus of ideal predictions. Clustering of the green points near this diagonal line would signify high predictive accuracy, whereas deviation suggests prediction errors. This unique pattern is a direct consequence of the model's stable, trend-based prediction. It confirms that the FSA-RF model is not attempting to predict every single price movement. Instead, it predicts a consistent, average price, effectively acting as a low-pass filter to focus on the core trend of the market. This characteristic is a significant advantage for producing a more reliable and less volatile forecast.

In summary, the results demonstrate the effectiveness of the proposed method across performance metrics under varying data conditions. The likely origins of the FSA-RF model's outperformance lie in the effective hyperparameter optimization by FSA combined with the inherent robustness of Random Forest. However, attention must be given to limitations related to news data quality and the model's scalability for extremely large datasets. A qualitative understanding of the model's predictive capabilities is enhanced through the inclusion of visual analyses.

5. Conclusion

This paper proposed a new FSA-RF technique to enhance the stock market data by predicting the stock market price of various companies. The stock market datasets are employed for training in the developed technique. Initially, the trained historical stock data and news datasets performs pre-processing for removing errors and noises. Then sentiment analysis is processed using polarity of input data. Then feature extraction will start to recognize and extract the relevant features from the dataset. Then the fitness function of the FSA is updated to the output layer it can predict the stock market price based on the updating time series of the dataset. Finally, the integration of sentiment analysis with historical stock data proved effective in capturing market dynamics and improving prediction accuracy. The proposed FSA-RF method achieves a mean predicted accuracy of 58.06%, 11.49%, and 47.76% higher than LSTM-CNN on positive, negative, and neutral sentiment, respectively. In comparison with CEEMD-CNN-LSTM, the proposed method's mean predicted accuracy is 42.26%, 24.5%, and 65% higher across positive, negative, and neutral sentiment. Against LSTM-GRU, the proposed method shows a mean predicted accuracy increase of 63.33%, 49.23%, and 23.75% for positive, negative, and neutral sentiment, respectively.

Despite the promising results, the current work acknowledges certain limitations. The model's primary focus lies in the analysis of historical price movements and market sentiment derived from the news. While this approach captures temporal dependencies, it exhibits a limited capacity to directly incorporate the dynamic and often unpredictable influence of external variables. Factors such as geopolitical risks and sudden macroeconomic shifts, which can significantly impact stock prices are not explicitly modelled. The predictive capability of the current framework remains inherently linked to the historical data and the sentiment expressed within the analysed news articles. Furthermore, the application of the model was primarily evaluated on a single stock market. The diverse characteristics of different stock exchanges globally including variations in trading volumes, volatility patterns, and market trends, that suggest the model's direct generalizability to other markets without specific adaptation requires further investigation. The absence of real-time sentiment feeds from social media platforms known for causing immediate shifts in stock prices. It also represents a potential area for improvement in prediction timeliness.

Future research will address current limitations to enhance model robustness and predictive accuracy. The key directions involve integration of dynamic external variables such as real-time geopolitical data, macroeconomic indicators, feature engineering for quantifying their impact on sentiment and prices, and investigation of hybrid modelling with time-varying parameters. Subsequent studies will evaluate performance across diverse international stock markets for generalization assessment and market-specific adaptations. Incorporation of real-time sentiment feeds from social media platforms aims to improve prediction timeliness. A concrete future direction is to integrate BERT-based sentiment analysis to capture contextual nuances in news data, which is an area explored in [39]. The comparative study involving Transformer-based architectures for sentiment analysis and multilingual dataset testing for cross-lingual applicability forms part of future work. Beyond stock markets, adaptability of the FSA-RF framework to other domains with sequential data and textual information, such as commodity markets and economic forecasting, will be explored. Addressing these aspects aims for more resilient, accurate, and broadly applicable forecasting capabilities in future model iterations.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Human and Animal Rights

This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed Consent

Informed consent does not apply as this was a retrospective review with no identifying patient information.

Funding

Not applicable

Conflicts of Interest Statement

Not applicable

Consent to Participate

Not applicable

Consent for Publication

Not applicable

Availability of Data and Material

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code Availability

Not applicable

Competing Interests

Not applicable

References

[1] Xiao C, Xia W & Jiang J (2020), Stock price forecast based on combined model of ARI-MA-LS-SVM. *Neural Computing and Applications* 32, 5379-88. https://doi.org/10.1007/s00521-019-04698-5.

[2] Huy DT, Nhan VK, Bich NT, Hong NT, Chung NT &Huy PQ (2020), Impacts of internal and external macroeconomic factors on firm stock price in an expansion econometric model—a case in Vietnam real estate industry. *Data science for financial econometrics* 14, 189-205. https://doi.org/10.1007/978-3-030-48853-6_14.

[3] Gurjar M, Naik P, Mujumdar G & Vaidya T (2018), Stock market prediction using ANN. *International Research Journal of Engineering and Technology* 5, 2758-61.

[4] Van Nguyen T, Zhou L, Chong AY, Li B & Pu X (2020), Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research* 281, 543-58. ht.tps://doi.org/10.1016/j.ejor.2019.08.015

[5] Nabipour M, Nayyeri P, Jabani H, Mosavi A, Salwana E & S S (2020), Deep learning for stock market prediction. *Entropy* 22, 840. https://doi.org/10.3390/e22080840.

[6] Poldrack RA, Huckins G &Varoquaux G (2020), Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry* 77, 534-40. https://doi.org/10.1001/jamapsychiatry.2019.3671.

[7] Hu Z, Zhao Y & Khushi M (2021), A survey of forex and stock price prediction using deep learning. *Applied System Innovation* 4, 9. https://doi.org/10.3390/asi4010009.

[8] Jiang W (2021), Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications* 184, 115537. https://doi.org/10.1016/j.eswa.2021.115537.

[9] Chicco D & Jurman G (2020), The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 6. https://doi.org/10.1186/s12864-019-6413-7.

[10] Shi L, Teng Z, Wang L, Zhang Y & Binder A (2018), DeepClue: visual interpretation of text-based deep stock prediction. *IEEE Transactions on Knowledge and Data Engineering* 31, 1094-108. https://doi.org/10.1109/TKDE.2018.2854193.

[11] Matsunaga D, Suzumura T&Takahashi T (2019), Exploring graph neural networks for stock market predictions with rolling window analysis. *arXiv preprint arXiv:1909.10660*.

[12] Moghar A &Hamiche M (2020), Stock market prediction using LSTM recurrent neural network. *Procedia computer science* 170, 1168-73. https://doi.org/10.1016/j.procs.2020.03.049.

[13] Zhu Y, Zhang W, Chen Y & Gao H (2019), A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment. *EURASIP Journal on Wireless Communications and Networking* 2019, 274. https://doi.org/10.1186/s13638-019-1605-z.

[14] Toharudin T, Pontoh RS, Caraka RE, Zahroh S, Lee Y & Chen RC (2023), Employing long short-term memory and Facebook prophet model in air temperature forecasting. *Communications in Statistics-Simulation and Computation* 52, 279-90. https://doi.org/10.1080/03610918.2020.1854302.

[15] Burton M, Lyon L, Erdmann C & Tijerina B (2018), Shifting to data savvy: the future of data science in libraries.

[16] Haiyun Z &Yizhe X (2020), Sports performance prediction model based on integrated learning algorithm and cloud computing Hadoop platform. *Microprocessors and Microsystems* 79, 103322. https://doi.org/10.1016/j.micpro.2020.103322.

[17] Liu H & Long Z (2020), An improved deep learning model for predicting stock market price time series. *Digital Signal Processing* 102, 102741. https://doi.org/10.1016/j.dsp.2020.102741.

[18] Wang Q, Bu S & He Z (2020), Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN. *IEEE Transactions on Industrial Informatics* 16, 6509-17. https://doi.org/10.1109/TII.2020.2966033.

[19] Guo R, Fu D & Sollazzo G (2022), An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree. *International Journal of Pavement Engineering* 23, 3633-46. https://doi.org/10.1080/10298436.2021.1910825.

[20] Chung H & Shin KS (2018), Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability* 10, 3765. https://doi.org/10.3390/su10103765.

[21] Hoque KE&Aljamaan H (2021), Impact of hyperparameter tuning on machine learning models in stock price forecasting. *IEEE Access* 9, 163815-30. https://doi.org/10.1109/ACCESS.2021.3134138.

[22] Marcjasz G (2020), Forecasting electricity prices using deep neural networks: A robust hyper-parameter selection scheme. *Energies* 13, 4605. https://doi.org/10.3390/en13184605.

[23] Yan X, Weihan W & Chang M (2021), Research on financial assets transaction prediction model based on LSTM neural network. *Neural Computing and Applications* 33, 257-70. https://doi.org/10.1007/s00521-020-04992-7.

[24] Kumar K & Haider MT (2021), Enhanced prediction of intra-day stock market using metaheuristic optimization on RNN–LSTM network. *New Generation Computing* 39, 231-72. https://doi.org/10.1007/s00354-020-00104-0.

[25] Sun L, Xu W & Liu J (2021), Two-channel attention mechanism fusion model of stock price prediction based on CNN-LSTM. *Transactions on Asian and Low-Resource Language Information Processing* 20, 1-2. https://doi.org/10.1145/3453693.

[26] Wang H, Wang J, Cao L, Li Y, Sun Q & Wang J (2021), A stock closing price prediction model based on CNN-BiSLSTM. *Complexity* 2021, 5360828. https://doi.org/10.1155/2021/5360828.

[27] Gao Y, Wang R & Zhou E (2021), Stock prediction based on optimized LSTM and GRU models. *Scientific Programming* 2021, 4055281. https://doi.org/10.1155/2021/4055281.

[28] Lu W, Li J, Li Y, Sun A & Wang J (2020), A CNN-LSTM-based model to forecast stock prices. *Complexity* 2020, 6622927. https://doi.org/10.1155/2020/6622927.

[29] Suman SK, Kumar D &Bhagyalakshmi L (2014), SINR pricing in non cooperative power control game for wireless ad hoc networks. *KSII Transactions on Internet & Information Systems* 8. https://doi.org/10.3837/tiis.2014.07.005.

[30] Bhagyalakshmi L, Suman SK &Sujeethadevi T (2020), Joint routing and resource allocation for cluster based isolated nodes in cognitive radio wireless sensor networks. *Wireless Personal Communications* 114, 3477-88. https://doi.org/10.1007/s11277-020-07543-4.

[31] Mahalakshmi K, Kousalya K, Shekhar H, Thomas AK, Bhagyalakshmi L, Suman SK, Chandragandhi S, Bachanna P, Srihari K &Sundramurthy VP (2021), Public auditing scheme for integrity verification in distributed cloud storage system. *Scientific Programming* 2021, 8533995. https://doi.org/10.1155/2021/8533995.

[32] Suman SK, Arivazhagan N, Bhagyalakshmi L, Shekhar H, Shanmuga Priya P, Helan Vidhya T, Jagtap SS, Mohammad GB, Chikte SD, Chandragandhi S &Yeshitla A (2022), Detection and prediction of HMS from drinking water by analysing the adsorbents from residuals using deep learning. *Adsorption Science & Technology* 2022, 3265366. https://doi.org/10.1155/2022/3265366.

[33] Lakshminarayanan B &Krishanan M (2014), Avoiding energy holes problem using load balancing approach in wireless sensor network. *KSII Transactions on Internet & Information Systems* 8. https://doi.org/10.3837/tiis.2014.05.007.

[34] Zhiheng W & Jianhua L (2021), Flamingo search algorithm: a new swarm intelligence optimization algorithm. *IEEE Access* 9, 88564-82. https://doi.org/10.1109/ACCESS.2021.3090512.

[35] NYSE Historical Stock Prices (2025), https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs.

[36] Financial News (2025), https://www.kaggle.com/datasets/hengzheng/financial-news.

[37] Kaggle (2025),https://www.kaggle.com/datasets/arindamdasgupta/daily-stock-prices.

[38] Kaur J&Sohal RS (2024), Noise estimation and removal in natural language processing. InHandbook of Vibroacoustics, *Noise and Harshness* 18,693-717. https://doi.org/10.1007/978-981-97-8100-3_38.

[39] Sun Z, Wang G, Li P, Wang H, Zhang M & Liang X (2024), An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications* 237, 121549. https://doi.org/10.1016/j.eswa.2023.121549.