A Method to Increase the Analysis Accuracy of Stock Market Valuation: A Case Study of the Nasdaq Index

Haixia Niu*
School of Finance, Henan Finance University; Zhengzhou Henan, 451464, China

***Abstract*—For a significant period, conventional methodologies have been employed to assess fundamental and technical aspects in forecasting and analyzing stock market performance. The precision and availability of stock market predictions have been enhanced by machine learning. Various machine learning methods have been utilized for stock market predictions. A novel, optimized machine-learning approach for financial market analysis is aimed to be introduced by this study. A unique method for improving the accuracy of stock price forecasting by incorporating support vector regression with the slime mould algorithm is presented in the present work. Other optimization algorithms were employed to enhance the prediction accuracy and the convergence speed of the network, which were Biogeography-based optimization and Gray Wolf Optimizer. An assessment of the proposed model's effectiveness in predicting stock prices was conducted through research employing Nasdaq index data extending from January 1, 2015, to June 29, 2023. Substantial improvements in accuracy for the proposed model were indicated by the results compared to other models, with an R-squared value of 0.991, a root mean absolute error of 149.248, a mean absolute percentage error of 0.930, and a mean absolute error of 116.260. Furthermore, not only is the prediction accuracy enhanced by the integration of the proposed model, but the model's adaptability to dynamic market conditions is also increased.**

***Keywords*—Machine learning; Nasdaq index; support vector regression; gray wolf optimizer; slime mould algorithm**

I. Introduction

The stability and safety of the stock market are considered of utmost importance, as it is regarded as a crucial element within national economies [1] [2]. The behaviors and consequences of the stock markets have become a vital area of scholarly investigation owing to the potential risks associated with them [3]. The forecasting of stock price trajectories is deemed a crucial responsibility as it serves the dual purpose of maintaining stability within financial markets by regulators and enabling investors to make informed choices while mitigating potential risks. The utilization of uncertain prediction processes and the subsequent generation of erroneous predictions can potentially lead to significant hazards [4]. Hence, the development of a robust and compelling predictive model is deemed crucial for mitigating potential hazards. The issue of stock market uncertainty is addressed by other theories, while conventional forecasting techniques rely on patterns that exhibit consistent behaviour across time. The aforementioned approach fails to account for the inherent volatility of the stock market, and when combined with the multitude of variables involved, the task of predicting stock values becomes a complex endeavor. Nevertheless, the emergence of machine learning (ML) [5] [6] is being considered. A comprehensive approach that utilizes a variety of algorithms to optimize performance in various situations is demonstrated by the solution offered. This emerging advancement exhibits considerable promise in its capacity to fundamentally transform our methodologies for forecasting stock market trends. The notion that trustworthy information can be discerned and patterns within a given dataset can be identified through machine learning is commonly accepted [7].

The anticipation of stock market trends has been an enduring area of interest for scholars, and machine learning methodologies are progressively assuming a more prominent role in this regard. By conducting a comparative analysis of state-of-the-art machine learning methods using a decade of daily historical data from the top 10 equities on the Casablanca Stock Exchange, our study contributes to this ongoing dialogue. It is worth mentioning that ensemble learning has been utilized in the past for this purpose, as demonstrated by Bilal et al. [8]. These efforts employed various classifiers, including ridge regression, LASSO regression, support-vector machine (SVM), k-nearest neighbors, random forest, and adaptive boosting. SVM, adaptive boosting, random forests, and SVM were discovered to perform exceptionally well in short-term forecasting, demonstrating the effectiveness of ensemble learning across a range of prediction horizons. Expanding upon the aforementioned groundwork, Sonkavde et al. [9] investigated a variety of methodologies including time series analysis, ensemble algorithms, deep learning, and supervised and unsupervised machine learning, in order to address the complexities associated with stock price classification and prediction. Furthermore, a comprehensive analysis of the Nasdaq stock market was presented by Ashfaq et al. [10], who employed a variety of machine learning regressors to forecast the opening prices of specific companies the following day. For the purpose of evaluation, they utilized metrics such as the mean square error and coefficient of determination, thereby enhancing our comprehension of predictive modelling within this particular domain. Agrawal et al. [11] presented a seminal study in which they described an algorithmic approach for predicting the stock market that utilized deep learning and non-linear regression. The researchers' investigation, which utilized a decade's worth of data from the New York Stock Exchange and Tesla Stock Price, demonstrated that their proposed solution outperformed currently available machine learning algorithms. However, in their inability to adapt to the volatile and unpredictable nature of financial markets, these methodologies frequently encountered obstacles. The trade-offs that existed between accuracy and computational efficiency exposed this dilemma.

The advantages and disadvantages of these literatures are presented in the Table I. In order to fill this void, our research presents an enhanced machine-learning methodology that integrates the slime mould algorithm and support vector regression in a synergistic fashion. The objective of this innovative approach is to surmount the limitations identified in prior investigations through the improvement of forecast accuracy and flexibility in the face of market volatility. Our methodology signifies a substantial deviation from traditional approaches, enabling a more intricate examination of intricate market data and surmounting the constraints intrinsic in conventional models. Our objective is to provide a comprehensive resolution to the persistent difficulty of precisely forecasting stock market fluctuations within a dynamic financial environment.

TABLE I. THE ADVANTAGES AND DISADVANTAGES OF THE RELEVANT LITERATURE

Authors	Advantages	Disadvantages
Bilal et al. [8]	Effective for short-term forecasting across a range of prediction horizons.	Lack of adaptability to unpredictable and volatile stock market.
Sonkavde et al. [9]	Addresses complexities of stock price classification and prediction.	Unable to adapt to stock markets that are volatile and unpredictable.
Ashfaq et al. [10]	Enhanced understanding of predictive modeling.	Unableness to adjust to unexpected and chaotic stock markets.
Agrawal et al. [11]	Outperformed existing machine learning algorithms.	Inability to adapt to volatile and unpredictable financial markets.

Linear regression is one of the machine learning models most frequently utilized for forecasting. A valuable statistical technique that can be employed to make predictions about a numerical outcome is provided by linear regression [12]. Artificial neural networks (ANN) [13], Support vector machines (SVM) [13], Decision trees [14], Random Forest [15], Logistic regression [16], Gradient boosting [17], time series forecasting [18], and more. All of these models possess the limitations and advantages that exert a substantial influence on the accuracy of predictions.

The method of research employed in this study is support vector regression (SVR), which is a resilient form of supervised learning. In SVR, an effort is made to minimize both structural risk and empirical risk while also reducing the range of trust in training examples. A significant level of efficacy is demonstrated by the previously mentioned methodology in addressing intricate nonlinear issues, particularly those characterized by a limited sample size. Notable efficacy in addressing intricate nonlinear issues, particularly those characterized by a limited sample size, is demonstrated by the mentioned methodology. A crucial role is played by SVR in mitigating risk and enhancing the overall predictive accuracy of future samples. This, in turn, facilitates the extraction of useful insights and enhances the decision-making process [19]. An in-depth understanding of the principles and advantages of SVR is imperative for achieving ideal outcomes and accomplishments in several fields, such as machine learning, data science, and related domains.

The behavior and performance of a model during training are significantly influenced by a range of hyperparameters. Several factors, such as model complexity, regularization power, and learning rate, among other considerations, are encompassed by the hyperparameters. To optimize the model's efficacy, the careful selection and meticulous refinement of appropriate hyperparameters are imperative. Engaging in this practice has the potential to greatly affect the precision, resilience, and the applicability of the model, enabling it to more effectively align with the dataset on which it is being trained. Hence, the meticulous fine-tuning of the hyperparameters is an essential stage in achieving the utmost performance of the model. Various techniques and algorithms are employed for the optimization of hyperparameters for models, such as grey wolf optimization (GWO) [20] [21], Biogeography-based optimization (BBO) [22], Grasshopper optimization algorithm (GOA) [23], Slime mould algorithm (SMA) [24], Moth flame optimization (MFO) [25], and more. Some of the above optimization techniques are nature-inspired. Three strategies were applied to improve the proposed model's hyperparameters: Biogeography-based optimization, grey wolf optimization, and the Slime mould algorithm.

The GWO algorithm, as proposed by Mirjalili et al. [21], is inspired by the hierarchical structure of leadership and hunting patterns observed in grey wolves. Fundamentally, the grey wolf species are categorized into four distinct groupings, namely alpha, beta, delta, and omega. A methodology known as BBO, proposed by Nazari et al. [26], is utilized. A pioneering methodology called SMA was presented in a study conducted by Chen et al. This methodology is inspired by the behavioural patterns exhibited by slime mould organisms in natural environments. One intriguing attribute of slime mould is its capacity to perceive the olfactory cues associated with the presence of food particles in the surrounding environment. The olfactory perception of food odors in the surrounding environment is a notable characteristic that contributes to the intriguing nature of slime mold. The SMA approach aims to replicate this inherent mechanism to optimize its efficacy in attaining its goals.

In the present study, a comprehensive dataset covering the period from January 2015 to June 2023 was analyzed using many models. To ensure the accuracy and reliability of the produced outputs, thorough training was conducted for the SVR method, taking into account a wide array of input factors. The criteria included several factors such as daily transaction volume, high and low prices, as well as the opening and closing prices. To assess the accuracy of the model's results, a comprehensive testing procedure was conducted, employing the same parameters as those used during the initial phase. The outcome of this intensive training and evaluation procedure yields a model capable of furnishing traders and investors with invaluable market insights, facilitating well-informed decision-making, and ultimately fostering profitable investments. The variable data was acquired from the stock market of the Nasdaq. In order to achieve its objectives, the research employed a methodology comprising multiple analytical stages. The research paper presents a comprehensive examination of the data source and its relevant elements in the subsequent section. The data are analyzed utilizing a variety of

methods, including the SMA optimizer, evaluation metrics, and the SVR model. Following the presentation of the analyses' results in the third section, those obtained from alternative methodologies are contrasted. The study concludes with a concise presentation of the results in the concluding section.

II. METHODOLOGY

A. Support Vector Regression

A very effective technique utilized in the domain of machine learning, particularly for nonlinear classification. The SVR has become a widely adopted technique among data scientists due to its ability to handle inputs with high dimensions effectively [27] [28]. The SVR, as seen in Fig. 1, has demonstrated remarkable efficacy in resolving classification problems, prompting its extension to tackle challenges associated with regression. Similar to its predecessor, the present improvement, referred to as SVR, demonstrates exceptional proficiency in managing intricate data sets. SVR and its many extensions have emerged as vital tools in the field of machine learning due to their ability to provide a flexible and accurate methodology for addressing regression and classification problems [29].

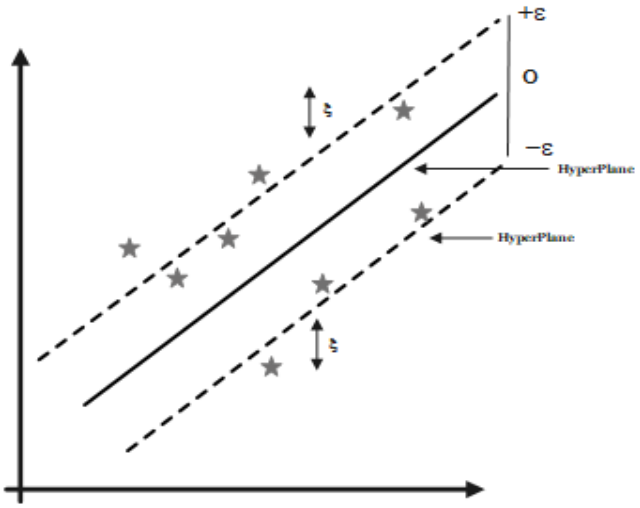

Fig. 1. The diagram of the SVR.

The initial description pertains to a linear function, denoted as f(x), which is characterized by the following mathematical form:

$$f(x) = w \times x + b \tag{1}$$

where, x is the input vector, b is a constant that has to be calculated, and w is the vector holding the parameters. In instances involving nonlinear problems, the data is transformed into a higher-dimensional space by the utilization of a nonlinear kernel.

$$f(x) = w\phi \times (x) + b \tag{2}$$

where, φ(x) is the kernel function.

The utilization of a feature space with larger dimensions can be employed to facilitate the mapping of data, hence enabling the implementation of a linear regression approach. The coefficients w and b can be obtained by minimizing a given function as follows:

$$min\frac{1}{2} \parallel w \parallel^2 + c\sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{3}$$

The exposure to:

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \text{with } \xi_i, \xi_i^* \geq 0, i = 1, \ldots . N \end{cases} \tag{4}$$

where, the positive and negative errors are denoted by ξ_i and ξ_i^*, respectively. The constant C>0 is a hyperparameter that enables the adjustment of the trade-off between the permissible error and the flatness of the function $f(x)$.

B. Biogeography-based Optimization

The BBO algorithm operates by simulating the movement of various species, which is determined by the suitability of their respective habitats. This process allows for a comprehensive analysis of the complex relationships between different species and their environments, ultimately leading to more refined and accurate predictions about ecological patterns. In the context of an optimization issue, it might be argued that a solution bears resemblance to a habitat. An optimal approach for addressing population concerns involves the establishment of densely populated habitats that offer superior living circumstances for various species compared to alternative habitats. The environment in which living animals encounter significant challenges is where the least optimal solution within the population is found. Therefore, the superior solutions can attract the inferior solutions due to their shared characteristics. The method of sharing features is accomplished by the utilization of the outlined operators.

Operation of Migration: Migration is a procedure through which a poorer habitat is swapped with a better one based on emigration rates. The quantification of the influx of species in a given area is referred to as the immigration rate. The migration incidence is anticipated to be greater under a more favourable solution compared to a less favourable option.

On the contrary, the quantitative assessment of the number of individuals within a species that depart from their environment is known as the rate of migration. Consequently, the emigration rate is expected to be greater under a suboptimal solution compared to an optimal option. The basic form of BBO has utilized straight paths as shown in Eq. (4).

$$\mu_k = \frac{E \times k}{n} \lambda_k = I\left(1 - \frac{k}{n}\right) \tag{5}$$

$\boldsymbol{\mu_k}$: Migration amount of $\mathbf{k}^{\text{th}}$ habitat.

$\boldsymbol{\lambda_k}$: Migration amount of $\mathbf{k}^{\text{th}}$ habitat.

I: Maximum immigration amount.

E: Maximum emigration rate.

$\boldsymbol{n} =$: Maximum number of species that a habitat can support.

K: Number of species count.

Mutation: Mutation in BBO may be likened to an abrupt alteration in the environmental circumstances experienced by living organisms, such as those caused by natural disasters like earthquakes, volcanic eruptions, or tornadoes. Similarly, the random modifications in the genetic makeup of a species prompt its migration to a new habitat, as the previous habitat becomes unsuitable for its survival.

C. Gray Wolf Optimizer

The GWO is a unique optimization approach that has been developed through the utilization of a meta-heuristic method. The strategy, initially introduced by Mirjalili et al. [18], emulates the social hierarchy and hunting strategies employed by grey wolves. Alpha is often regarded as the most optimum choice, whereas Omega is positioned as the final challenger inside the hierarchical framework of leadership.

Three principal hunting techniques are utilized by the method in order to emulate the behaviours of wolves: The action of following, confining, and assaulting prey. To simulate the locomotion patterns of grey wolves during hunting activities in their natural habitat, the following relate was employed:

$$\vec{D} = \left|\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)\right|$$
$$\vec{X} \mid (t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D} \quad (6)$$

In which, $\boldsymbol{t}$ is the current iteration, $\vec{\boldsymbol{D}}$ denotes movement, $\vec{\boldsymbol{X}}_{\boldsymbol{p}}$ denotes prey location, $\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{C}}$ denotes coefficient vectors, and $\vec{\boldsymbol{X}}$ denotes the grey wolf's position. The coefficient vectors ($\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{C}}$) are constructed using the relationships shown below.

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a}$$
$$\vec{C} = 2 \cdot \vec{r}_2 \quad (7)$$

The spatial allocation of novel search representatives pertaining to omegas is modified by using data derived from alpha, beta, and delta in the following manner:

$$\vec{D}_a = \left|\vec{C}_1 \cdot \vec{X}_a - \vec{X}\right|, \vec{D}_\beta = \left|\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}\right|, \vec{D}_\delta = \left|\vec{C}_3 \cdot \vec{X}_s - \vec{X}\right| \quad (8)$$

$$\vec{X}_1 = \vec{X}_a - \vec{A}_1 \cdot \vec{D}_u, \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta, \vec{X}_3 = \vec{X}_b - \vec{A}_3 \cdot \vec{D}_\delta \quad (9)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (10)$$

The wolves, denoted by the subscripts $\alpha, \beta,$ and δ , converge to initiate a conclusive assault in order to accomplish the objective successfully. The variable $\vec{\boldsymbol{a}}$ is used in order to replicate the previous assault by altering a value from 2 to 0. On the other hand, the variable a represents a random variable that falls between the range of $-2\vec{a}$ and $2\vec{a}$. Consequently, if $\vec{a}$ is lowered, it will also lead to a decrease in the value of $\vec{A}$. The wolves were compelled to grasp their prey due to the factor tightly $|\vec{A}| < 1$. Grey wolves engage in cooperative hunting strategies by forming packs and exhibiting a hierarchical social structure. These packs are led by an alpha wolf, who guides the group's activities. The pack members disperse to forage for food individually, and then afterwards regroup to launch coordinated attacks. Wolves may separate in search of prey when $\left|\vec{\boldsymbol{A}}\right|$ has a random value greater than unity. The wolf count and generation number are considered to be the two most critical configuration parameters for the GWO algorithm. This means that the total number of objective function evaluations will be equal to the wolf population times the size of the generation or,

$$OFEs = N_W \times N_G \quad (11)$$

D. Slime Mould Algorithm

Li et al. introduced a unique approach called SMA, which draws inspiration from the behaviour of slime mould seen in natural environments [24]. The slime mould, which is represented in Fig. 3, uses its olfactory perception and detects the volatilized scent of nourishment in the air in order to approach its prey. Fig. 2 provides a comprehensive illustration of the general characteristics of SMA.

The behaviour of the slime mould may be mathematically described by the equation that goes as follows:

$$\overrightarrow{X(t+1)} = \begin{cases} \overrightarrow{X_b(t)} + \overrightarrow{v_b} \cdot \left(\overrightarrow{W} \cdot \overrightarrow{X_A(t)} - \overrightarrow{X_B(t)}\right) & r < p \\ \overrightarrow{v_c} \cdot \overrightarrow{X(t)} & r \geq p \end{cases} \quad (12)$$

In which, $X_b(t)$ reflects the specific area of the slime mould that now exhibits the greatest level of odour, the variables $X(t)$ and $X(t+1)$ represent the locations of the slime mould in iteration t and $t+1$, respectively. $X_A(t)$ and X_B represent two randomly selected sites of slime mould. The variable v_b varies over time within the range [$-a$, a], where r is a random number between 0 and 1. The parameter p is specified as ($a = \operatorname{arctanh}(-(\frac{t}{\max\_t}) + 1)$), and v_c is a linearly decreasing parameter ranging from 0 to 1.

$$p = tanh\,|S(i) - DF| \quad i = 1, 2, \ldots, n \quad (13)$$

where, DF denotes the fittest iteration overall, and $S(i)$ denotes the fitness of $\overrightarrow{X}$. Following is a definition of the weight W equation:

$$\overrightarrow{W\left(smell\ index(l)\right)} = \begin{cases} 1 + r.log\left(\frac{bF - S(i)}{bF - wF} + 1\right), condition \\ 1 - r.log\left(\frac{bF - S(i)}{bF - wF} + 1\right), others \end{cases} \quad (14)$$

$$smell\ index = sort(S) \quad (15)$$

The variable $S(i)$ represents the initial half of the population in the given equation. bF for the best fitness, wF for the poorest fitness, and the scent index for the sorted fitness values. By utilizing the formula provided, the spatial coordinates of the slime mould are revised:

$$\overrightarrow{X^*} = \begin{cases} rand(UB - LB) + LB & rand < z \\ \overrightarrow{X_b(t)} + \overrightarrow{v_b} \cdot \left(\overrightarrow{W} \cdot \overrightarrow{X_A(t)} - \overrightarrow{X_B(t)}\right) & r < p \\ \overrightarrow{v_c} \cdot \overrightarrow{X(t)} & r \geq p \end{cases} \quad (16)$$

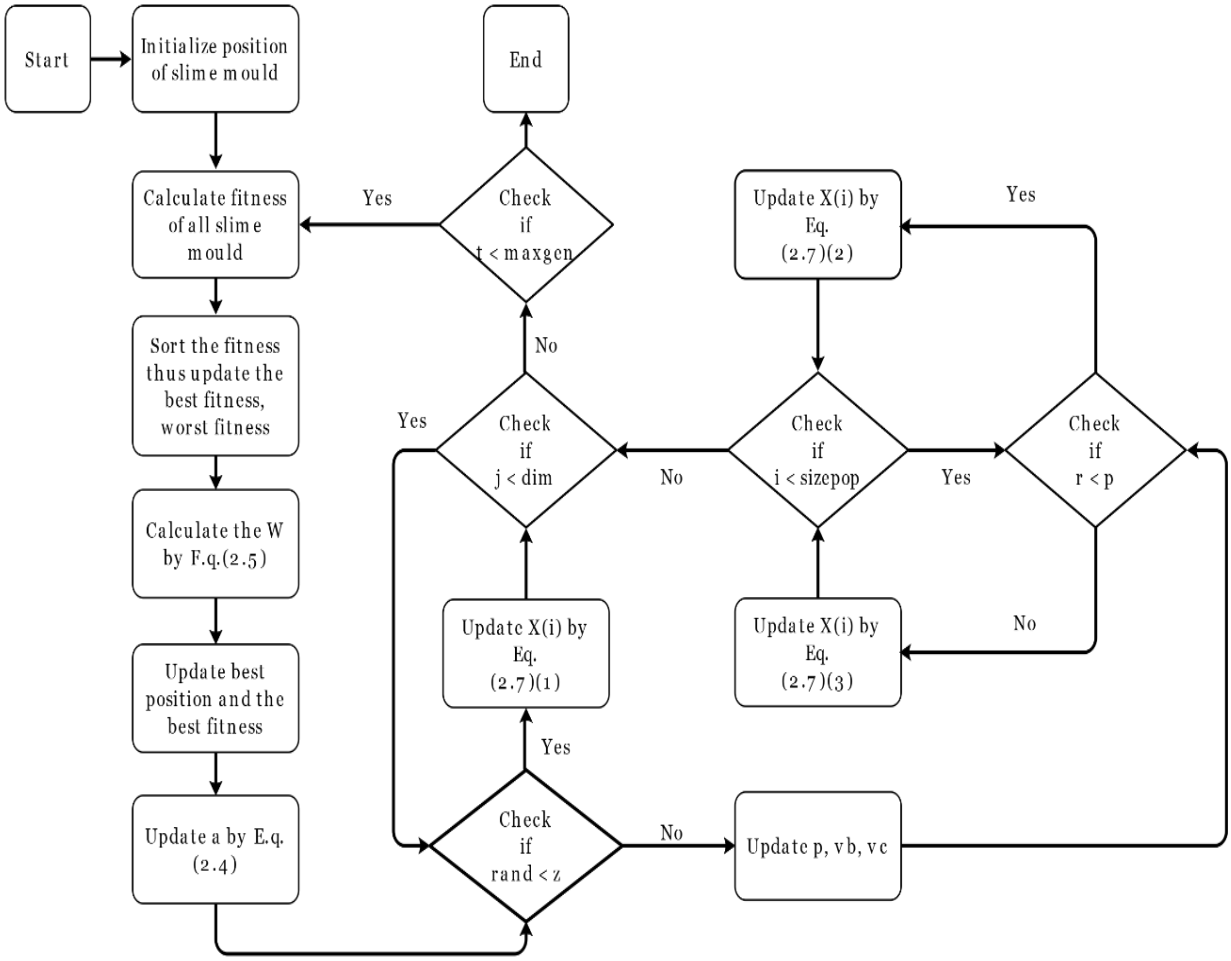

Fig. 2. The diagram of the Slime mould algorithm.

In this context, the parameter denoted as z is constrained to a range between 0 and 0.1. The terms LB and UB refer to the lower and upper borders of the search interval correspondingly.

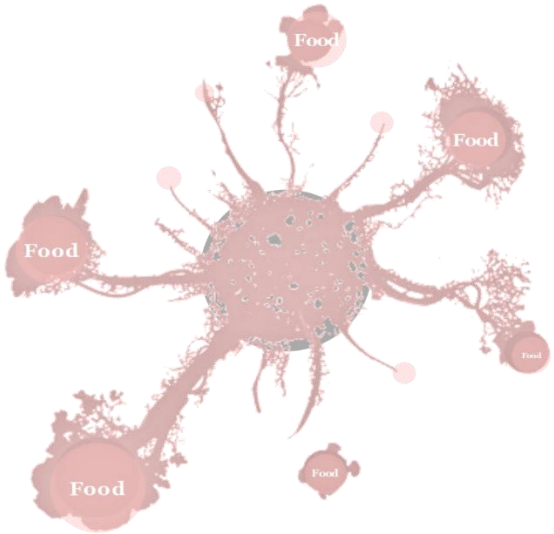

Fig. 3. The illustration of the SMA algorithm.

E. Dataset Collection

In order to conduct a thorough analysis, it is imperative to include trade volume and Open, High, Low, and Close (OHLC) prices during a designated period. Data for this investigation was collected via Yahoo Finance from January 2, 2015, to June 29, 2023. The Nasdaq, which was founded in 1971, is notable for being the inaugural electronic stock exchange in the globe. It distinguishes itself from conventional exchanges by functioning exclusively electronically and utilizing a digital infrastructure that optimizes the trading process. This allows for efficient and rapid transaction processing. In addition to technology firms, the Nasdaq Composite Index comprises companies from consumer services, healthcare, and a variety of other sectors. Due, in part, to its emphasis on emerging sectors and technology, the Nasdaq is a centre for innovative and expanding businesses, given its status as a major participant in international financial markets. Investors routinely track the Nasdaq Composite Index in order to assess the financial well-being of technology and growth firms, as well as to deduce wider economic patterns and investor sentiment.

F. Dataset Description

The process of ensuring the high quality of the raw data is a crucial and fundamental stage in the pursuit of obtaining meaningful insights. The process of data preparation is of utmost importance in attaining this objective. The process encompasses a variety of tasks, such as the elimination of undesirable data, the establishment of standardized formats for widespread applicability, and the arrangement of information in a manner that promotes the retrieval of valuable insights. This has special significance in initiatives that entail substantial quantities of data because data quality takes precedence over mere numerical values. Data preparation activities encompass a range of tasks, such as encoding categorical data, cleaning and organizing data, scaling, standardization, and normalization in accordance with established industry standards. By engaging in these activities, the precision and dependability of the insights derived from the data can be enhanced. To achieve data scaling and normalization, Min-Max scalers were utilized in the project's data pre-processing step. This approach facilitated the removal of inconsistencies, as well as the measurement of null, missing, and unknown values. The methodology was illustrated using the data from the Nasdaq index. The dataset encompasses the timeframe spanning from January 2015 to June 2023 and has undergone several preparatory procedures, such as standardization.

The process of transforming numerical attributes inside a dataset to a specific range, commonly ranging from zero to one, is referred to as feature scaling. This technique is alternatively recognized as Min-Max normalization or data preparation. The objective is to maintain the relative relationships between the values while ensuring that all features are brought to a comparable scale. The consideration of input feature quantity holds particular significance in machine learning algorithms that exhibit sensitivity towards this aspect. The formula for the data normalization procedure is as follows:

$$\text{XScaled} = \frac{(X - Xmin)}{(Xmax - Xmin)} \tag{17}$$

G. Evaluation Metrics

Evaluation metrics are commonly used in the fields of statistics and machine learning to assess the effectiveness of forecasting models quantitatively. They aid in evaluating the predictive accuracy of a model on data that has not yet been observed. The selection of the optimal evaluation metric is contingent upon the particular analytical objectives and the nature of the predictive task at hand. Performance indicators such as R-squared ($\boldsymbol{R^2}$), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Relative Squared Error (RSE), and Mean Absolute Error (MAE) were utilized in this study to evaluate the prediction accuracy of the developed forecasting models. Below, a compilation of mathematical formulas pertaining to these metrics is provided:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{19}$$

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100 \tag{20}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{21}$$

$$RSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(\bar{y} - \hat{y}_i)} \tag{22}$$

$$MSE = \frac{1}{N}\sum_{k=0}^{n}\binom{n}{k}(Fi - Yi)b^2 \tag{23}$$

III. Result and Discussion

A. Statistical Results

Table II displays the statistical characteristics, such as mean, std., min, 25%, 50%, 75%, max, and variance. The central tendency of a data set is quantified by the mean, a statistical term. The calculation involves the sum of all values inside the dataset, followed by the division of the resulting sum by the total count of values. The dispersion or spread of data points relative to the mean is quantified by the standard deviation, a statistical metric. The minimum, minimum value, or minimal observation is the term used for the smallest data point or value in a dataset. It assists in finding the lowest number within a batch of data as well as understanding the range and distribution of the data. The degree to which individual data points in a dataset differ from the average (mean) of the dataset is described by the statistical concept of variance. A low variance denotes that data points are close to the mean, whereas a high variance suggests that data points are dispersed from the mean. Quantiles, specifically the 25th, 50th, and 75th percentiles, play a pivotal role in comprehending the data distribution. As indicated by the 25th percentile (also referred to as the first quartile), 25% of the observed values are situated below this threshold. As an illustration, a 25% percentile of 5776.33 for the opening price indicates that 25% of the dataset's initial prices fall below this threshold. As the 50th percentile, which is also equal to the median, divides the dataset in half, it is a significant indicator. With a median value of 7833.27, the proportion of closing prices that occur either above or below this value is half. The third quartile, or 75th percentile, indicates that 75% of the data are located below this value. A 75% percentile of the close price (1,590.78) indicates that 75% of the closing prices fall below this threshold. These quantiles offer valuable insights regarding the market's overall trend and stability. A significant disparity between the 25th and 75th percentiles, for instance, may suggest that stock prices are more volatile. A comprehension of the distribution of the majority of data points, namely stock prices, can assist analysts and investors in recognising customary price ranges, detecting anomalies, and identifying substantial changes in market trends.

The closing price data is shown in Fig. 4, whereby it has been partitioned into two distinct zones for the purposes of training and testing. This strategy ensures the precision of the data while aiding consumers in acquiring reliable insights.

TABLE II. RESULTS OF THE STATISTICS FOR THE OHCLV MODELS THAT WERE PRESENTED

	Open	High	Low	Volume	Close
mean	8744.356	8805.287	8677.574	3143.8	8745.821
Std.	3332.744	3362.163	3298.311	1551.37	3332.058
min	4218.81	4293.22	4209.76	706.88	4266.84
25%	5776.33	5821.95	5769.39	1908.94	5793.83
50%	7829.03	7867.15	7791.98	2318.76	7833.27
75%	11573.14	11699.63	11476.66	4416.84	11590.78
max	16120.92	16212.23	16017.23	11621.19	16057.44
variance	11107186	11304139	10878852	2406747	11102609

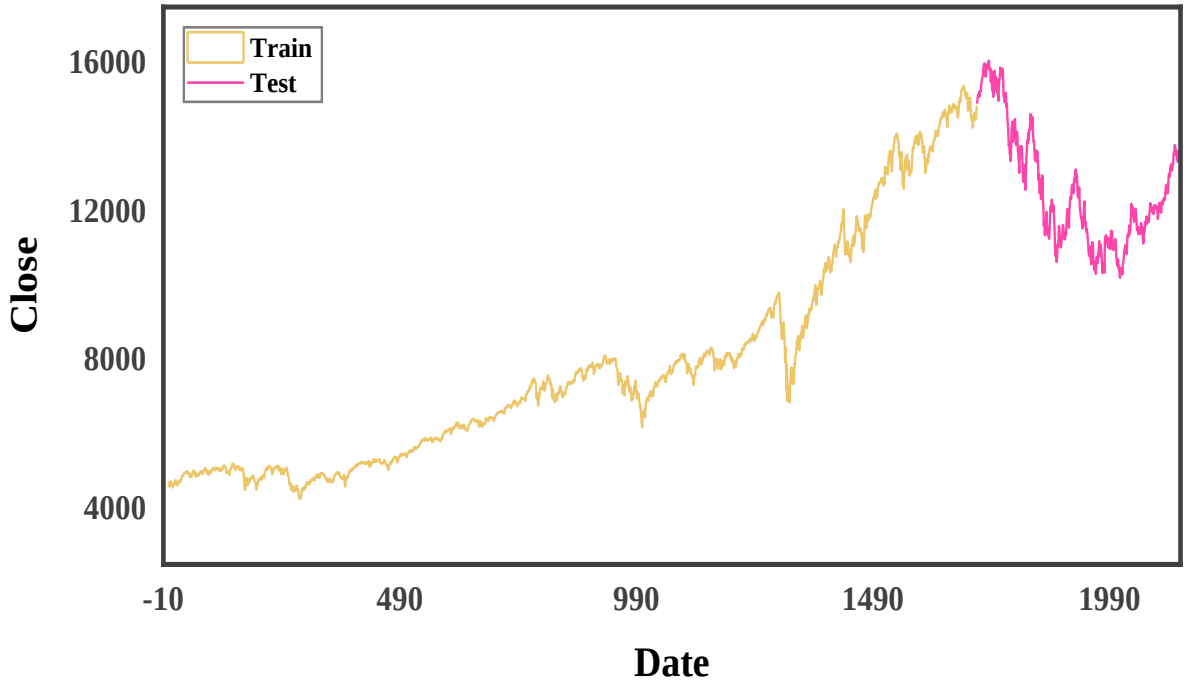

Fig. 4. Data set division into the train and test.

B. Comparative Analysis

The algorithms BBO, GWO, and SMA, which were chosen for comparison in this research, are widely acknowledged for their efficacy and efficiency in a range of optimization tasks, including financial market analysis. The selection of these algorithms is based on their established history of success in handling intricate optimization tasks, which is consistent with the aims of our research. Relevant benchmarks, these algorithms have been extensively implemented and cited in prior research for comparable applications. Their widespread usage and recognition within the scientific community signifies their acceptability and validation, thereby furnishing a strong foundation for comparative analysis. The exclusion of other algorithms is warranted due to the fact that our research is primarily concerned with algorithms that have demonstrated exceptional potential in the analysis of financial markets. Furthermore, an excessive number of algorithms may compromise the precision and comprehensibility of the comparative analysis.

In order to accomplish the crucial objective of forecasting the Nasdaq index, an identical dataset was used by each model. In this article, a meticulous analysis and evaluation of the outcomes of each model were conducted to provide a complete and informative comparison of their respective performances. Establishing a comprehensive and equitable comparison requires the clarification of performance indicators used for evaluating the models. By employing a diverse set of significant metrics, as explained in the methodology section, the models were assessed. A comprehensive evaluation of the performance of each model can be conducted by using a variety of indicators, thus facilitating the determination of the model that most effectively meets the established requirements. All the many nuances of how each model performed are displayed in a thorough Table III with the findings.

TABLE III. THE OUTCOMES OF THE MODEL'S EVALUATION CRITERIA AND TIME COMPUTING

Train/Test	Metrics	SVR	BBO-SVR	GWO-SVR	SMA-SVR
TRAIN SET	R^2	0.980	0.985	0.989	0.992
	RMSE	417.512	356.900	303.741	264.583
	MAPE	5.482	4.240	3.028	1.763
	MAE	382.054	329.413	225.660	158.062
	RSE	590.803	505.028	429.810	374.396
	MSE	174316.292	127377.539	92258.490	70004.188
TEST SET	R^2	0.972	0.981	0.988	0.991
	RMSE	265.540	217.667	173.450	149.248
	MAPE	1.666	1.376	1.073	0.930
	MAE	215.186	174.385	134.495	116.260
	RSE	376.411	308.550	245.870	211.564
	MSE	70511.446	47379.041	30084.774	22275.026
Time	Second	0.156	163.85	221.81	139.94

Initially, the obtained outcome was used in the selection process of the SVR model. The decision to develop the SVR model was made after a comprehensive review of the data due to its exceptional performance. Between the commencement of 2015 and the midpoint of 2023, the Nasdaq index data underwent a procedure that included the selection of relevant data and normalization. The collection of important insights will benefit the decision-making process using this detailed technique. The evaluation score for SVR alone is 0.972 in R^2, which, as shown in Table III, has increased due to improvements in the problematic optimizers. The R^2 criteria values for the BBO, GWO, and SMA algorithms are 0.981, 0.988, and 0.991, respectively, indicating the possibility of selecting the optimal course of action. When compared to other optimizers, superior results are produced by the SMA optimizer. The findings of the RMSE model shown in Table III also confirm the superiority of the SMA optimizer. The RMSE values for SVR, BBO-SVR, GWO-SVR, and SMA-SVR are 265.54, 217.667, 17.3450, and 149.248, respectively. Furthermore, the hybrid models outperform the SVR model, suggesting that tweaking the model's hyperparameters can be beneficial for optimization. However compared to SVR, hybrid models require longer running times. The rationale is that while hyperparameters of the SVR are manually set, hybrid-model hyperparameters are optimized via metaheuristic algorithms. A consistent upward trend in all metrics is observed in the training set, progressing from the SVR model to the SMA-SVR model. More precisely, the R^2 value, which indicates the extent to which the independent variables account for the variability of the dependent variable's variance, increases from 0.980 in the SVR model to 0.992 in the SMA-SVR model. This suggests that the SMA-SVR model exhibits superior predictive capability regarding the outcomes. The RMSE, an indicator of prediction error standard deviation, exhibits a substantial reduction from 417.512 in the SVR model to 264.583 in the SMA-SVR model. This pattern is similarly evident in MAPE and MAE, where SMA-SVR exhibits a significant decline in errors, signifying its exceptional precision in prognostication. In addition, the reduction in RSE and MSE provides additional evidence that, among the four models evaluated for the train set, SMA-SVR exhibits the most accurate predictions and the lowest error rates. Similar results are observed when the SMA-SVR model is applied to the test set. The model attains the maximum R^2 value of 0.991, indicating an exceptional capacity for variance prediction. The SMA-SVR model exhibits the lowest values of RMSE, MAPE, and MAE, which further validate its exceptional accuracy and dependability when forecasting on unseen data. The comparatively reduced RSE and MSE values of SMA-SVR, in contrast to the alternative models, provide additional validation for its status as the most precise and dependable model across training and testing scenarios. Fig. 5 shows evaluation of the suggested model's performance in comparison to other models during training.

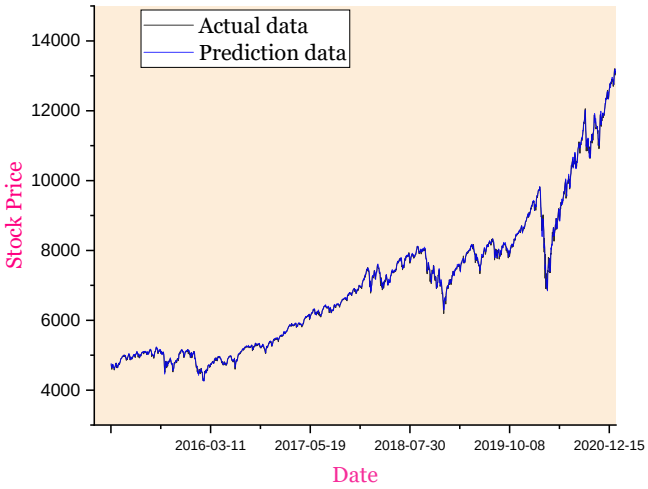

Fig. 5. Evaluation of the suggested model's performance in comparison to other models during training.

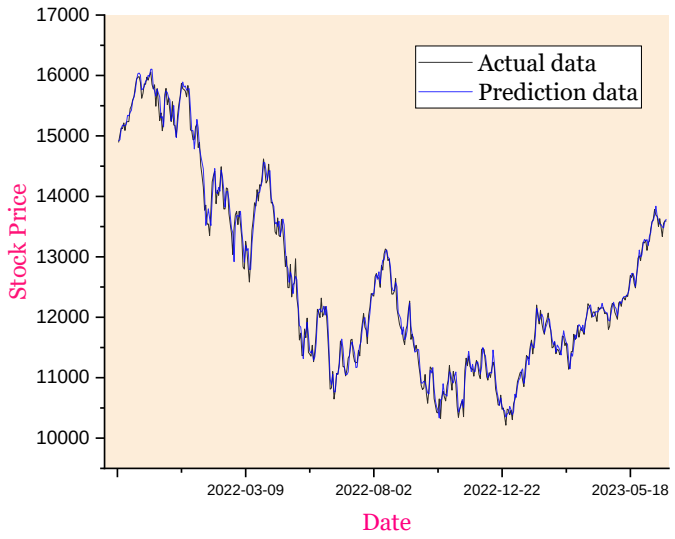

Fig. 6. Evaluation of the suggested model's performance in comparison to other models during testing.

The outcomes of the conducted tests are illustrated in Fig. 7 and Fig. 8, showcasing a robust connection between the model and the empirical data. Among the models tested, the SMA-SVR model exhibited superior performance in comparison to the individual SVR, BBO-SVR, and GWO-SVR models. It is worth mentioning that the utilization of the optimizer approach resulted in a substantial enhancement in the performance of the SVR model. Fig. 6 and Fig. 7 provide a comprehensive examination of the four models, therefore validating the superior performance of the chosen model. These results indicate that the SMA-SVR model is a potential method for precisely forecasting the intended results in the context. These results indicate that the SMA-SVR model is a potential method for precisely forecasting the intended results in the context. The proposed model has the following limitations:

The high R-squared value of 0.991 may potentially signify the occurrence of overfitting in the model. This is the result of an excessive learning load imposed by the training data, which may include noise and outliers, and may consequently hinder the model's performance when applied to novel, unseen data.

Fig. 7. Evaluation metrics outcomes for the attending models during train.

Fig. 8. Evaluation metrics outcomes for the attending models during the test.

The impact of external variables on market volatility: The model may not comprehensively capture the stock market's vulnerability to various elements, including economic indicators, political occurrences, and market sentiment. This constraint serves to underscore the intrinsic indeterminacy of the market.

The effectiveness of the model is widely recognized to be significantly contingent upon the provision of high-quality and exhaustive data. The predictive accuracy may be substantially affected by constraints in the data, including biases or incomplete historical information.

The augmentation of computational demands and complexity is an acknowledged consequence of integrating support vector regression with the slime mould algorithm. Potentially reducing the efficacy of real-time predictions, this could require additional processing time and power.

Market-Specific Variability in Generalizability: It has been noted that the efficacy of the model may differ when applied to various financial instruments or stock markets. This feature highlights the necessity for additional verification and assessment.

It is observed that the performance of the model is highly susceptible to changes in the hyperparameters it is configured with. Potentially inconsistent performance across a variety of datasets may result from the intricacy associated with fine-tuning these parameters.

Restriction on the Study's Scope: The study's concentration on particular algorithms and optimization techniques may have overlooked alternative approaches that have yet to be investigated but could have proven to be more effective.

Concerns Regarding Regulatory Compliance, Transparency, Fairness, and Ethics: The application of sophisticated predictive models in the realm of stock trading gives rise to ethical and regulatory issues. Careful attention should be given to these critical considerations.

Further research could be focused on a number of critical domains in light of the study's findings and methodologies:

Algorithm Improvement: To further enhance the accuracy and adaptability of predictions, further improvements could be made to the integration of support vector regression and the slime mould algorithm through the investigation of additional parameters or alternative configurations.

Conducting comparative analyses with novel or less prevalent optimization algorithms could provide fresh perspectives on fluctuating market conditions by evaluating the performance of the present model.

Analyses of Real-Time Data: The model's practical efficacy and resilience in authentic market situations could be evaluated through its implementation in a real-time trading environment.

Additional Market Segments: To assess the model's adaptability across various market sectors, it might be advantageous to expand its scope to encompass a more extensive array of financial instruments, including bonds, commodities, and cryptocurrencies.

The model's performance in diverse economic and regulatory environments could be better comprehended through cross-market analysis, which involves the application of the model to stock markets in different countries or regions.

Enhancement through Inclusion of Supplementary Data Sources: The predictive accuracy of the model could potentially be improved through the integration of alternative data sources, such as social media trends, news sentiment analysis, or economic indicators.

The potential for enhanced prediction capabilities and the ability to process more intricate data patterns could be investigated through the incorporation of deep learning techniques with the existing model.

The model could be rendered accessible to a wider spectrum of users, including individuals lacking technical expertise, through the development of a user interface that is intuitive and easy to use.

In order to aid investors in comprehending and alleviating potential losses, risk management functionalities might be integrated into the model.

An evaluation of the model's consistency and dependability over prolonged durations could be accomplished through the implementation of a long-term performance analysis.

The examination of the model's reaction to market anomalies or exceptional circumstances, such as unanticipated global events or financial crises, may have valuable implications.

An investigation into the feasibility of attaining a more precise or resilient prediction outcome might involve the implementation of ensemble techniques, wherein the current model is combined with additional predictive models.

Further investigation could be warranted into the ethical and regulatory ramifications associated with the utilization of sophisticated AI for forecasting the stock market, with a specific focus on the principles of trading transparency and fairness.

IV. CONCLUSION

By employing stock prediction techniques to evaluate asset values and ascertain prevailing market trends, a substantial competitive advantage can be gained by both individual and institutional investors. Informed decisions regarding purchasing, selling, or retaining stocks can be made by investors through the utilization of historical data and advanced algorithms. The implementation of this particular method holds significant importance for investors committed to making prudent investment decisions, as it effectively mitigates risks and increases the likelihood of achieving profitable outcomes. Many predictive algorithms and data sources were employed in this research to evaluate the complex and ever-changing domain of stock prediction. The findings suggest that the accuracy of forecasting might potentially be enhanced through the utilization of a hybrid model or an ensemble technique. Finally, the construction and evaluation of the prediction model underscored the need to use data-driven insights to establish dependable decision-making processes. This highlights the benefits of a data-centric strategy in the modern, rapidly changing business environment, as well as the possible applications of predictive analytics across a wide variety of sectors. The objective of this research was to develop models with enhanced predictive capabilities for stock prices, enabling traders and investors to use these algorithms to make informed decisions on the optimal timing and price for purchasing stocks.

In this study, the following findings were reached:

- First, the data preparation and normalization process were completed, potentially impacting how the prediction model is presented. The chosen model was then prepared to begin its data analysis.
- The best model was selected, findings were analyzed, and hyperparameters were adjusted to increase the model's effectiveness.
- The identification of the most accurate optimization technique as the primary optimizer of the model was achieved through a comparative analysis of the outcomes produced by several optimization algorithms. Among the three optimization algorithms, namely BBO, GWO, and SMA, the SMA technique exhibited the highest performance in terms of the R^2 evaluation criterion, with a result of 0.991, surpassing the results of 0.981 and 0.988 achieved by BBO and GWO, respectively.

REFERENCES

[1] S. Claessens, J. Frost, G. Turner, and F. Zhu, “Fintech credit markets around the world: size, drivers and policy issues,” BIS Quarterly Review September, 2018.

[2] W. Li et al., “The nexus between COVID-19 fear and stock market volatility,” Economic research-Ekonomska istraživanja, vol. 35, no. 1, pp. 1765–1785, 2022.

[3] Z. Wang et al., “Measuring systemic risk contribution of global stock markets: A dynamic tail risk network approach,” International Review of Financial Analysis, vol. 84, p. 102361, 2022.

[4] Z. Li, W. Cheng, Y. Chen, H. Chen, and W. Wang, “Interpretable click-through rate prediction through hierarchical attention,” in Proceedings of

the 13th International Conference on Web Search and Data Mining, 2020, pp. 313–321.

[5] B. Mahesh, "Machine learning algorithms-a review," International Journal of Science and Research (IJSR).[Internet], vol. 9, no. 1, pp. 381–386, 2020.

[6] S. Ray, "A quick review of machine learning algorithms," in 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, 2019, pp. 35–39.

[7] E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, and L. Serrano, Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques. IGI global, 2009.

[8] A. E. L. Bilali, A. Taleb, M. A. Bahlaoui, and Y. Brouziyne, "An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset," J Hydrol (Amst), vol. 599, p. 126510, 2021.

[9] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications," International Journal of Financial Studies, Vol 11, Iss 94, p 94 (2023), Jan. 2023, doi: 10.3390/ijfs11030094.

[10] N. Ashfaq, Z. Nawaz, and M. Ilyas, "A comparative study of Different Machine Learning Regressors For Stock Market Prediction," 2021. doi: 10.48550/arxiv.2104.07469.

[11] S. C. Agrawal, "Deep learning based non-linear regression for Stock Prediction," IOP Conference Series: Materials Science and Engineering ; volume 1116, issue 1, page 012189 ; ISSN 1757-8981 1757-899X, 2021, doi: 10.1088/1757-899x/1116/1/012189.

[12] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in An Introduction to Statistical Learning: With Applications in Python, Springer, 2023, pp. 69–134.

[13] P. Chhajer, M. Shah, and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction," Decision Analytics Journal, vol. 2, no. November 2021, p. 100015, 2022, doi: 10.1016/j.dajour.2021.100015.

[14] S. B. Kotsiantis, "Decision trees: a recent overview," Artif Intell Rev, vol. 39, pp. 261–283, 2013.

[15] L. Breiman, "Random forests," Mach Learn, vol. 45, pp. 5–32, 2001.

[16] R. E. Wright, "Logistic regression.," 1995.

[17] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Front Neurorobot, vol. 7, p. 21, 2013.

[18] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," Int J Forecast, vol. 22, no. 3, pp. 443–473, 2006.

[19] W. S. Noble, "What is a support vector machine?," Nat Biotechnol, vol. 24, no. 12, pp. 1565–1567, 2006.

[20] H. Rezaei, O. Bozorg-Haddad, and X. Chu, "Grey wolf optimization (GWO) algorithm," Advanced optimization by nature-inspired algorithms, pp. 81–91, 2018.

[21] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46–61, 2014.

[22] D. Simon, "Biogeography-based optimization," IEEE transactions on evolutionary computation, vol. 12, no. 6, pp. 702–713, 2008.

[23] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: theory and application," Advances in engineering software, vol. 105, pp. 30–47, 2017.

[24] S. Li, H. Chen, M. Wang, A. A. Heidari, and S. Mirjalili, "Slime mould algorithm: A new method for stochastic optimization," Future Generation Computer Systems, vol. 111, pp. 300–323, 2020.

[25] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," Knowl Based Syst, vol. 89, pp. 228–249, 2015.

[26] A. Hadidi and A. Nazari, "Design and economic optimization of shell-and-tube heat exchangers using biogeography-based (BBO) algorithm," Appl Therm Eng, vol. 51, no. 1–2, pp. 1263–1272, 2013.

[27] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat Comput, vol. 14, pp. 199–222, 2004.

[28] E. H. Houssein, M. Dirar, L. Abualigah, and W. M. Mohamed, "An efficient equilibrium optimizer with support vector regression for stock market prediction," Neural Comput Appl, vol. 34, no. 4, pp. 3165–3200, 2022, doi: 10.1007/s00521-021-06580-9.

[29] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat Comput, vol. 14, pp. 199–222, 2004.