International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**5**, pp.**26-34**, May **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

IJCSE
ISSN: 2347-2693 (E)

Survey Article

Survey on Machine Learning Techniques for Stock Market Prediction: Models, Challenges, and Future Directions

Payal Hake[1*], Sheetal Sonawane[2], Kalyani Waghmare[3], Sanket Dabade[4]

[1,2,3]Dept. of Computer Science, PICT College, Pune, 411043, Maharashtra, India
[4]B.Tech., IIT Bombay; Member, The Institute of Chartered Accountants of India

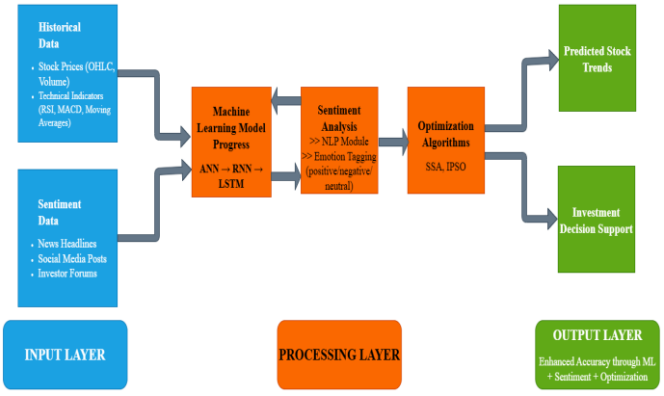
**Corresponding Author:* ✉

Received: 24/Mar/2025; **Accepted**: 26/Apr/2025; **Published**: 31/May/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i5.2634

Abstract: The stock market significantly impacts the global economy by shaping investment choices and contributing to financial stability. However, its dynamic, volatile, and non-linear nature makes stock price prediction a challenging yet essential task for investors, analysts, and researchers. Traditional forecasting methods, such as fundamental and technical analysis, often fail to capture complex market patterns. Recently, Machine Learning (ML) techniques have demonstrated effectiveness in forecasting stock prices by analyzing historical data and identifying intricate trends. This survey provides a comprehensive review of various ML models, including Artificial Neural Networks (ANNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and hybrid approaches, highlighting their effectiveness in stock market forecasting. Additionally, it explores the role of sentiment analysis in price prediction, as financial news, social media, and investor sentiment significantly influence market movements. The paper discusses key challenges, optimization strategies, and recent advancements in combining ML with sentiment analysis for enhanced predictive accuracy. By analyzing existing literature and identifying research gaps, this survey offers valuable insights into stock market prediction using machine learning.

Keywords: Stock Market Prediction, Machine Learning, Deep Learning, Sentiment Analysis, Optimization Algorithms

Graphical Abstract: The stock market plays a crucial role in the global economy, but its volatile and non-linear behaviour makes accurate price prediction challenging. This survey reviews various machine learning models such as ANN, RNN, LSTM, and hybrid approaches demonstrating their ability to capture complex patterns in stock data. It also highlights the impact of sentiment analysis and recent advancements in optimization techniques, offering insights into improved forecasting accuracy and future research directions.



1. Introduction

The complexity and unpredictability of financial markets have made stock price forecasting a challenging pursuit, attracting considerable interest from researchers and investors alike. With the advent of machine learning, a wide array of techniques has been applied to financial forecasting, offering significant promise in improving prediction accuracy. This survey focuses on reviewing existing literature that applies machine learning techniques to stock price prediction, particularly Long Short-Term Memory (LSTM) networks and sentiment analysis models. LSTM, a variant of recurrent neural networks, excels at capturing sequential dependencies in stock market data, making it valuable for precise forecasting. Meanwhile, sentiment analysis extracting public sentiment from news articles, social media, and other text-based data sources adds an additional layer of predictive information by incorporating human emotion and market perception. By optimising models with advanced algorithms like the Sparrow Search Algorithm, researchers have explored ways to enhance predictive power. This paper offers an in-depth review of the existing research landscape in this field, highlighting major challenges, methodologies, and potential

future developments. By refining machine learning models, especially LSTM, and integrating other techniques like sentiment analysis, we hope to offer insights into improving stock price prediction models, leading to better investment strategies and more profitable outcomes.

1.1 An Overview and the Need for Machine Learning

The stock market is a dynamic financial system where investors trade shares of publicly listed companies. It reflects a nation's economic performance and enables businesses to raise capital by offering ownership stakes. Investors, from individuals to institutional entities, participate in the stock market to grow their wealth by making informed decisions regarding buying or selling stocks. However, predicting stock prices is an inherently difficult task due to the dynamic, volatile, and multifaceted nature of the stock market. Several factors influence stock prices, including company earnings, industry performance, macroeconomic indicators (like inflation, interest rates, and GDP), and geopolitical events. In addition, investor sentiment, reflected in the psychological and emotional responses to market news, adds another layer of complexity to stock price movements. The traditional methods of stock price forecasting, such as fundamental and technical analysis, often struggle to account for all these variables effectively.

1.2 Why Machine Learning is Essential for Stock Market Analysis

Given the enormous volume of data generated daily and the non-linear relationships between various factors affecting stock prices, traditional statistical models often fall short in making accurate predictions. This is where Machine Learning (ML) comes into play. ML algorithms are capable of processing large datasets, identifying hidden patterns, and capturing the non-linear dependencies that exist in stock price movements. These algorithms can learn from past data and adjust their models dynamically as new data becomes available, making them highly suitable for stock price prediction. In particular, advanced techniques such as Long Short-Term Memory (LSTM) networks, which excel at understanding sequential data, are well-suited to the temporal nature of stock prices. They can capture trends, seasonality, and long-term dependencies that are often crucial for making accurate predictions. Additionally, incorporating sentiment analysis allows machine learning models to quantify investor emotions derived from news articles, social media, and market reports, further improving prediction performance. Machine learning also enables the optimization of model parameters to maximise prediction accuracy. Techniques like the Sparrow Search Algorithm (SSA) can be employed to fine-tune models, making them more robust and adaptable to changing market conditions. This combination of deep learning, sentiment analysis, and optimization techniques offers investors a powerful toolkit for making well-informed investment decisions. In this light, the application of machine learning algorithms to stock price prediction is not merely a technological advancement, it has become a necessity. As financial markets continue to evolve and the volume of available data increases, investors require sophisticated tools to keep up with market movements and seize profitable opportunities. Machine learning provides a cutting-edge approach that addresses the limitations of traditional methods, enabling investors to make more accurate predictions and, ultimately, better investment decisions.

1.3 Motivation for this Survey Paper

Stock market prediction is a critical area of research in financial analytics due to its impact on investment strategies, risk management, and economic decision-making. Traditional statistical methods, such as ARIMA and GARCH, have been widely used but fail to capture the complex, non-linear, and highly volatile nature of stock prices. The advent of machine learning (ML) and deep learning (DL) techniques has opened new avenues for predictive modeling, leveraging historical stock prices, technical indicators, and sentiment analysis. This survey aims to explore the effectiveness of ML and DL models, such as Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM), and Hybrid Optimization approaches, in improving forecasting accuracy. Additionally, sentiment analysis using social media and financial news is integrated to enhance predictions by incorporating investor behavior and market trends. The motivation behind this study is to provide a comprehensive review of existing models, highlight their limitations, and propose future research directions to enhance predictive accuracy and reliability in financial forecasting.

2. Stock Market Prediction Techniques

In the papers surveyed, techniques such as data mining, machine learning and deep learning are used to estimate and predict future stock prices. The advantages and disadvantages of these techniques are discussed. These techniques are Traditional Statistical Time Series Models, Hybrid and Advanced Machine Learning Techniques, Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), Deep Learning with Sentiment Analysis, and Optimization Algorithms. A generalized stock price prediction process pipeline using various algorithms, models and techniques is given in Figure 1.

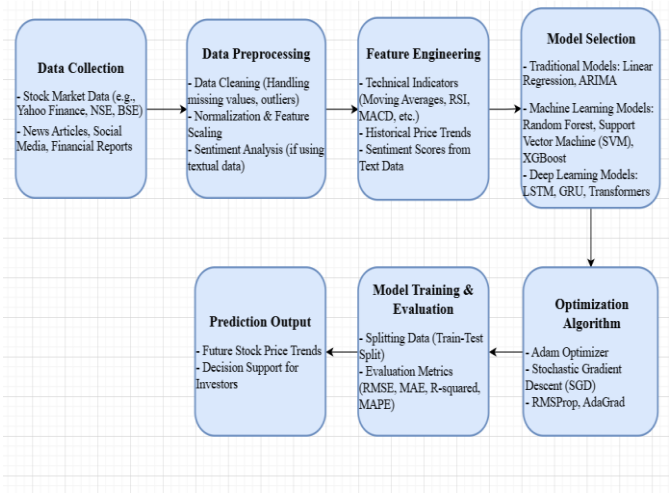


Fig. 1 Pipeline Architecture Diagram

3. Classification and Comparison

In this section should extend, not repeat the information discussed in Introduction [4]. In contrast, a Calculation Section represents a practical development from a theoretical basis [5].

3.1 Traditional Statistical Time Series Models Technique

Traditional statistical models like ARMA (Autoregressive Moving Average) and ARIMA (Autoregressive Integrated Moving Average) represent foundational approaches for time series forecasting. M. M. Rounaghi and F. N. Zadeh [1], applied ARMA to the financial data from the S & P 500 and London Stock Exchange to predict stock returns, while G. Bandyopadhyay [2], used ARIMA to forecast gold prices based on historical data. These models assume linearity in time series data, making them suitable for simpler, stationary datasets. ARMA and ARIMA models share common assumptions, such as requiring data stationarity. ARMA handles stationary data, while ARIMA extends ARMA to accommodate non-stationary data through differencing. Although these models have been widely used in finance, they are limited in their ability to capture non-linear relationships and are often outperformed by more advanced machine learning models in the presence of complex financial dynamics. However, they offer a baseline of comparison for more sophisticated approaches.

3.2 Hybrid and Advanced Machine Learning Techniques

This class focuses on hybrid models, where traditional statistical models are enhanced with machine learning algorithms. In [3], A. Hossain and M. Nasser, introduced a combination of ARMA-GARCH (Generalised Autoregressive Conditional Heteroskedasticity) with Recurrent Support Vector Machines (RSVM) and Recurrent Relevance Vector Machines (RRVM). These machine learning models improve volatility forecasting by capturing non-linear patterns that traditional models fail to address. J. Chai et al. [4], builds on this by introducing least squares support vector machine (LSSVM) with parameter optimization, improving stock price volatility prediction. J. Borade [5], integrated Support Vector Regression with sentiment analysis, allowing for stock price prediction based on both numerical and textual data. These hybrid models blend traditional statistical methods with machine learning to enhance accuracy in financial forecasting. While ARMA-GARCH models capture linear dependencies in time series data, they struggle with non-linearities and outliers, which RSVM and RRVM handle more effectively. Similarly, the LSSVM model introduces a support vector machine architecture to optimize stock price predictions. These hybrid approaches mark a step forward from pure statistical models by leveraging the strengths of machine learning to address the limitations of linearity in traditional time series models.

3.3 Artificial Neural Networks (ANN)

ANNs, or artificial neural networks, are powerful computational models inspired by the human brain's neural architecture. In these papers, ANNs are used for stock price and market index prediction. In [6], A. Murkute and T. Sarode implemented a multi-layer perceptron ANN to predict stock prices, while A. H. Moghaddam et al. [7], applied ANN to predict NASDAQ's daily closing prices. M.Nabipour et al. [8], compared ANN with tree-based models like decision trees, gradient boosting, and XGBoost to identify the best models for stock market forecasting. ANNs are non-linear models that excel in capturing complex relationships within financial data, outperforming traditional time series models in many cases. However, ANNs can struggle with long-term dependencies and time lags in sequential data, issues that are better addressed by RNNs and LSTM models in subsequent classes. Nevertheless, ANNs serve as a stepping stone in the development of more advanced neural networks, laying the groundwork for future innovation.

3.4 Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models are designed to capture temporal dependencies in time series data, making them especially suitable for financial forecasting. LSTMs, in particular, solve the vanishing gradient problem present in traditional RNNs, allowing for the effective learning of long-term dependencies. In [9], A. Sherstinsky and in [10], Rohit Tetarwal and Rohit Tushir in [24] G. Ding and L. Qin provided foundational insights into the mechanisms of RNNs and LSTMs, while Y. Ji and et al. [11], S. Hochreiter and J. Schmidhuber [12], V. Gupta and M. Ahmad [13], and V. R. Madireddy [14] applied these models to stock price and trend prediction. LSTM models outperform traditional RNNs and ANNs by addressing the vanishing gradient problem and handling long-range dependencies in sequential data. These models are particularly effective for financial data, which often contain long-term trends that need to be learned. The introduction of hybrid models, such as LSTM combined with IPSO in Paper Y. Ji and et al. [11], shows the potential for further enhancing LSTM performance through optimization techniques. In this class, we see the progression from traditional neural networks (ANNs) to more sophisticated models like LSTMs that are better suited for the intricacies of financial data.

3.5 Deep Learning with Sentiment Analysis

This class incorporates sentiment analysis into financial forecasting models. Sentiment analysis, a technique for extracting subjective information from text, is used to gauge investor sentiment from news articles, social media, and forums. C.R. Ko and H.T. Chang [15] and Y. Li and Y. Pan[16] applied LSTM and deep learning models combined with sentiment analysis to predict stock prices. In [17], C.Kearney and S.Liu and in [18], Y. Rao and et al. in[23], Rohit Kumar and et. al. focused on developing emotional dictionaries and sentiment analysis algorithms to enhance the prediction of stock movements based on textual data. T. Matsubara and et al. [19] introduced a deep generative model to address the problem of overfitting in traditional models by incorporating news articles into stock price prediction. Sentiment analysis provides an additional dimension to financial forecasting by integrating qualitative data (e.g., news and social media) with traditional quantitative financial data. LSTM and ensemble deep learning models, which are

already effective in capturing long-term dependencies in stock prices, are further enhanced by sentiment analysis, improving their ability to predict market movements influenced by public sentiment. This integration shows how external factors like investor mood and news can be pivotal in financial forecasting, marking an evolution from purely data-driven approaches to more holistic models that consider both numbers and narratives.

3.6 Optimization Algorithms and Hybrid Techniques

This class focuses on optimization algorithms used to improve the accuracy and efficiency of financial forecasting models. In [20], Z. D. Aksehir and E. Kilic¸, tackled the issues of data imbalance and feature selection in CNN models for stock price prediction. J. Xue and B. Shen [21] and A. Fathy and et al. [22] introduced the Sparrow Search Algorithm (SSA) for optimising model performance, while Y. Ji and et al. [11], combined LSTM with Improved Particle Swarm Optimization (IPSO) for enhanced stock indices forecasting. Optimization techniques like SSA and IPSO are crucial for tuning machine learning models, improving their accuracy, and solving complex financial forecasting problems. By optimising hyperparameters and enhancing feature selection, these methods allow models such as CNNs and LSTMs to perform better in handling non-linearities and outliers in financial data. The integration of optimization algorithms marks a trend towards building more efficient, adaptable, and accurate models for financial forecasting. The classification of these papers reveals a clear evolution in financial forecasting methods. Initially, traditional time series models like ARMA and ARIMA were used for linear financial data. However, as the complexity of financial markets increased, these models were complemented and eventually replaced by more advanced machine learning techniques like SVM, ANN, and LSTM, which are capable of capturing non-linear relationships and long-term dependencies.

Table 1 provides a comparative analysis of various machine learning and deep learning models used for stock market prediction, highlighting their strengths and limitations. Traditional models like ARIMA are noted for their effectiveness in modeling linear trends but struggle with capturing non-linear patterns in financial data. Machine learning models such as SVM and Random Forest offer better handling of non-linearity but require extensive feature engineering and are sensitive to noise. Deep learning models, particularly LSTM, excel in capturing temporal dependencies in sequential data and outperform traditional approaches in terms of accuracy. However, they demand large training datasets and significant computational resources. The table underscores the trade-offs between interpretability, data requirements, and predictive performance, setting the stage for further exploration into optimized LSTM architectures and hybrid approaches involving sentiment analysis.

Table.1. Comparison of various models and their limitations

Models	Limitations
ARMA/ARIMA	Struggles with non-linearity, cannot handle sudden market changes
GARCH and ARMA-GARCH	Sensitive to outliers, poor at handling non-stationary data
Support Vector Regression (SVR)	Computationally expensive, does not scale well to large datasets
Decision Trees, Random Forest, XGBoost	Overfitting risk, requires hyperparameter tuning
Artificial Neural Networks (ANNs)	Prone to overfitting, requires large datasets
Recurrent Neural Networks (RNNs)	Suffer from vanishing gradient problem, limited memory
Long Short-Term Memory (LSTM)	Computationally expensive, sensitive to hyperparameters
Convolutional Neural Networks (CNNs)	Struggles with sequential dependencies in stock data
Hybrid Models (LSTM + Optimizers like IPSO, SSA)	Computationally intensive, difficult to implement
Sentiment Analysis (BERT, LSTM-based models)	Hard to quantify impact, requires high-quality text data

4. Relative Advantage and Disadvantage of each Technique

Table 2 provides a comparative analysis of various prediction techniques used in stock market forecasting, highlighting their respective advantages and disadvantages. Traditional statistical models like ARIMA are appreciated for their interpretability and suitability for linear data, but they often fail to capture complex nonlinear patterns. In contrast, machine learning techniques such as Random Forest and Support Vector Machine (SVM) offer better performance on nonlinear datasets, though they may struggle with overfitting and require significant data preprocessing. Deep learning models like LSTM demonstrate superior accuracy by effectively capturing temporal dependencies in stock price data, but they demand large datasets and high computational power. Sentiment analysis-based models complement these methods by integrating market sentiment derived from news and social media, enhancing prediction quality, although they face challenges in accurately interpreting natural language. This comparative overview underscores the importance of selecting appropriate models based on the nature of the data and the specific prediction goals.

Table.2. Comparison of prediction techniques, their advantages, and disadvantages

Technique	Advantages	Disadvantages
ARMA [1]	Effective for capturing short-term linear trends in stationary data	Fails to capture non-linear dynamics and complex market behaviors
ARIMA [2]	Captures non-stationary time series data; good for long-term trends	Assumes linear relationships, does not handle sudden market shifts
ARMA-GARCH [3]	Accounts for volatility clustering in time series data	Sensitive to outliers, does not handle non-linearity well
Hybrid Least Square SVM (LSSVM) [4]	Combines linear regression and SVM, improves forecasting accuracy	Computational complexity, requires parameter tuning
Support Vector Regression (SVR) [5]	High prediction accuracy, integrates sentiment analysis	Limited to short-term predictions and a single stock
Multi-layer Perceptron ANN [6]	Captures non-linear patterns, better than statistical models	Sensitive to overfitting, lacks generalizability
Feed forward ANN [7]	Good for short-term predictions with high accuracy ($R^2 > 0.94$)	Limited timeframe, lacks long-term prediction ability
ANN and Tree-based models [8]	Effective for stock prediction	Generalizability to different markets is unclear
LSTM [9]	Captures long-term dependencies, solves vanishing gradient problem	Computationally expensive, difficult to train
LSTM [10]	Predicts multiple outputs like opening/high/low prices	Limitations in dataset details and loss function
LSTM + IPSO [11]	Handles non-linearity, improves accuracy with hyperparameter tuning	Generalizability to other markets not explored
LSTM Trend Prediction [13]	High trend accuracy, financial return potential	Doesn't explore generalization to unseen data
LSTM for BSE [14]	Accurate for long-term dependencies	Real-world issues (volatility) not addressed
LSTM + Sentiment [15]	Improves accuracy (12.05% RMSE reduction)	Sentiment analysis might be too simplistic
Ensemble Deep Learning [16]	Combines textual and numerical data, high accuracy	Needs better model interpretability and generalization
Sentiment Analysis [17]	Captures emotions in financial text	Unclear link between sentiment and market movement
Emotional Dictionary [18]	Fine-grained dictionaries predict emotions well	Sarcasm/real-world accuracy untested
Deep Generative Model [19]	Reduces overfitting, beats conventional methods	Poor generalizability to other news/assets
2D-CNN + Feature Selection [20]	Better accuracy with feature selection	Limited to Dow 30, generalization not shown
Sparrow Search Algorithm (SSA) [21]	Fast convergence, avoids local optima	Needs refinement for complex tasks
SSA for Micro-grid [22]	Works for single/multi-objective optimization	Real-world deployment not explored

5. Research Gaps

Some limitations of the existing techniques are discussed below. These research gaps, in turn, indicate directions for future research.

1. ARMA [1] and ARIMA [2] models have been widely used in finance. However, ARMA fails to capture non-linear dynamics and complex market behaviors. ARIMA assumes linear relationships, does not handle sudden market shifts. While ARMA-GARCH [3] models capture linear dependencies in time series data, but they struggle with non-linearities and outliers. RSVM, RRVM [3] are sensitive to outliers but are complex to implement.
2. ANN [5] is sensitive to overfitting but lacks generalizability to new data. ANN and Tree-based model's [9] generalizability to different markets is unclear. Multi-layer Perceptron and LSTM [7], [8] are computationally expensive and difficult to train. It is reported that 2D-CNN [10] also have limited generalizability beyond the Dow 30 index.
3. LSTM with IPSO [11] focused on stock indices and generalizability to other markets is not explored. SVR [12] is limited to short-term predictions and a single stock (Apple). SVM [13] is less effective for long-term predictions and doesn't consider broader variables.
4. Noise Trader Model [13] is limited to short-term anomalies, does not consider long-term effects. Sentiment Analysis with AZFinText [14] has limited accuracy for neutral and positive sentiment, and doesn't cover long-term trends.
5. LSTM with Sentiment Analysis [16] sentiment analysis might be overly simplistic. LSTM for Stock Price Trend Prediction [21] Doesn't explore model generalization to unseen data.

6. Insights into stock price prediction

Stock price prediction is a complex task influenced by both long-term historical trends and short-term sentiment variations. To capture these dynamics, we provide a

generalized framework that integrates historical performance analysis with sentiment driven market variations.

6.1 Core Components of the Framework

The framework is based on two primary components:

1. Historical Price Variations and Analysis (Long-Term): This represents the impact of historical stock performance, incorporating both fundamental and technical indicators. It reflects long-term trends, patterns, and seasonality in stock prices.
2. Sentiment Variation and Analysis (Short-Term): This accounts for the influence of market sentiment, driven by news, social media trends, and other real-time factors. Sentiment analysis helps capture short-term market reactions that may not be reflected in historical data.

6.2 Mathematical Representation

The generalized market index can be expressed as a combination of historical performance and sentiment analysis:

$$\text{Market Index} = F_1 + F_2 + F_1 \cdot F_2 + \epsilon \quad (1)$$

where:

- F_1 = Function representing the historical performance component
- F_2 = Function representing the sentiment analysis component
- ϵ = Error term accounting for random fluctuations and noise

Since F_1 and F_2 are not linearly additive (they interact in complex, non-linear ways), the variance of the market index can be derived using the following formula:

$$\text{Var}(\text{Market Index}) = \text{Var}(F_1) + \text{Var}(F_2) + \text{Var}(F_1 \times F_2) + \epsilon \quad (2)$$

6.3 Sectoral Index Modeling

The sectoral index represents the performance of a specific industry or sector, influenced by both the broader market trends and sector-specific sentiment:

$$\text{Sectoral Index} = F_1 + F_2 + F_1 \times F_2 + \epsilon \quad (3)$$

The variance for the sectoral index is given by:

$$\text{Var}(\text{Sectoral Index}) = \text{Var}(F_1) + \text{Var}(F_2) + \text{Var}(F_1 \times F_2) + \epsilon \quad (4)$$

6.4 Stock-Level Prediction

Traditional statistical models like ARMA (Autoregressive Moving Average) and ARIMA

Within each sector, individual stocks exhibit their own volatility, which is influenced by:

- Overall market volatility (derived from the market index)
- Sectoral volatility (derived from the sector index)
- Stock-specific sentiment and fundamentals

The generalized stock price model can be expressed as:

$$\text{Stock Price} = G_1(\text{Market Index Volatility}) + G_2(\text{Sectoral Volatility}) + G_3(\text{Stock-Specific Sentiment}) + \epsilon \quad (5)$$

6.5 Factors

Factors which influence stock price are as follows:

1. Factors which affect historical performance – These are Fundamental Factors such as Earnings and Revenue Growth, Earnings Per Share (EPS), Price-to-Earnings Ratio (P/E Ratio), Dividend Payouts, Debt Levels and Financial Health; Macroeconomic Factors such as Interest Rates, Inflation, GDP Growth, Unemployment Rates; Industry and Sector Specific Factors such as Regulatory Changes, Technological Advancements, Supply Chain and Raw Material Costs.
2. Factors which affect investor sentiment – These are factors such as Positive News, Negative News, Media Bias, Fake News and Misinformation, FOMO (Fear of Missing Out), FOLO (Fear of Losing Out), Breakouts or Resistance Levels, Wars, trade disputes, political instability.

6.6 Model Validation and Optimization Steps

Once the model is built, the following steps will be performed to enhance its accuracy and reliability:

1. Sensitivity Analysis: Identify the key drivers of market volatility and assess their impact on the final prediction. Perform what if analyses to test the model's response to external shocks.
2. Optimization: Implement advanced optimization techniques (e.g., SSA, Adam) to minimize the model's error term and enhance predictive accuracy. Fine-tune hyperparameters of the LSTM and sentiment analysis models to reduce overfitting and improve generalization.
3. Validation: Use real-world stock market data (e.g., Nifty 50) for back testing and validation. Compare the predicted results with actual market movements to measure the model's performance using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared values.

By addressing these challenges and leveraging cutting-edge advancements in AI, sentiment analysis, and optimization techniques, future research can bridge the gap between theoretical models and practical, high-performance financial forecasting systems.

7. Results and Discussion

This survey paper provides a comprehensive review of machine learning techniques for stock market prediction, highlighting a clear evolution from traditional statistical models to advanced AI-driven methodologies. Initially, models like ARMA and ARIMA were foundational for time series forecasting, suitable for simpler, stationary datasets and linear trends. However, their limitations in capturing non-linear relationships and sudden market shifts became apparent as market complexity increased. The progression led to

hybrid models that combine statistical approaches with machine learning algorithms, such as ARMA-GARCH with Recurrent Support Vector Machines (RSVM) and Recurrent Relevance Vector Machines (RRVM), which address non-linear patterns more effectively. Artificial Neural Networks (ANNs) further enhanced this capability by excelling in capturing complex non-linear relationships within financial data, often outperforming traditional methods. Despite their strengths, ANNs can struggle with long-term dependencies and time lags, leading to the development of Recurrent Neural Networks (RNNs) and, more specifically, Long Short-Term Memory (LSTM) networks, which are designed to capture temporal dependencies and solve the vanishing gradient problem, making them particularly effective for financial data with long-term trends.

The integration of sentiment analysis marks a significant advancement by incorporating qualitative data like news and social media, providing an additional dimension to financial forecasting that accounts for investor emotions and market perception. This approach enhances the predictive power of models by integrating real-time investor opinions, leading to improved accuracy, as seen in studies combining LSTM with sentiment analysis. Furthermore, the review underscores the critical role of optimization algorithms like the Sparrow Search Algorithm (SSA) and Improved Particle Swarm Optimization (IPSO) in fine-tuning model parameters, thereby enhancing accuracy and efficiency in complex financial forecasting problems. While these advanced machine learning and deep learning models offer superior performance compared to traditional methods, they often come with limitations such as high computational costs, demanding large datasets, sensitivity to hyperparameters, and challenges in interpretability and generalizability across diverse markets. Sources

8. Conclusion

The application of machine learning and deep learning has significantly advanced financial forecasting techniques, moving beyond the limitations of traditional statistical models like ARIMA and GARCH, which often struggle with the non-linear and volatile nature of stock prices. The integration of machine learning and deep learning, particularly Long Short-Term Memory (LSTM) networks and various hybrid models, has demonstrated superior performance in predicting stock price movements by effectively capturing complex patterns and long-term dependencies in financial data. A crucial component in this evolution is sentiment analysis, which incorporates textual sentiment from news articles, social media, and financial reports, thereby enriching models with insights into investor psychology. Furthermore, the performance of these predictive models is significantly enhanced through optimization techniques such as the Sparrow Search Algorithm (SSA) and Improved Particle Swarm Optimization (IPSO), which dynamically fine-tune hyperparameters for improved adaptability and accuracy.

Despite these advancements, the field faces ongoing challenges, including overfitting, the inherent "black box" nature of complex deep learning models, high computational costs, and the unpredictable dynamics of financial markets. A generalized model has been suggested that integrates historical performance and sentiment analysis to capture both long-term trends and short-term market variations, offering a comprehensive and scalable approach by modeling market, sectoral, and stock-level volatility using non-linear variance functions. This framework also incorporates sensitivity analysis, optimization, and validation steps to enhance predictive accuracy, making it adaptable across diverse markets and geographies.

9. Challenges and Future Frontiers

While machine learning and deep learning models have shown substantial progress in financial forecasting, significant challenges and research gaps persist, particularly regarding their generalizability across different markets and asset classes. Many existing studies have focused on specific stock indices or regions, limiting the broader applicability of their findings. Future research should aim to enhance generalizability by expanding model training to include diverse datasets from various global exchanges, commodities, and cryptocurrency markets, thereby improving robustness across the financial ecosystem. Another critical area for future efforts is fine-tuning model interpretability, as deep learning models often operate as "black boxes," making their decision-making processes opaque; thus, emphasizing explainable AI (XAI) techniques will be crucial for improving transparency in financial forecasting. Expanding data sources to include real-time economic reports, government policies, and investor sentiment from social media platforms could provide a more comprehensive predictive framework, with financial documents like annual reports and earnings calls offering potential for enhanced long-term trend prediction. Furthermore, addressing emotional and behavioral factors more precisely beyond simple positive, neutral, or negative classifications, by quantifying nuanced emotions like fear, greed, and panic, will allow future models to capture investor sentiment fluctuations more effectively. Handling market volatility, influenced by unexpected events such as economic crises and geopolitical conflicts, could benefit from hybrid models integrating reinforcement learning and real-time anomaly detection techniques for improved adaptability in high-volatility scenarios. Finally, reducing computational complexity is vital, as many deep learning models demand substantial computational power, hindering real-time predictions; therefore, developing lightweight, efficient models with optimized architectures will be critical for making predictive systems accessible for real-world applications.

Data Availability

Stock Market Data Sources for ML Model

Accurate stock price prediction requires high-quality historical data. Various platforms provide real-time and historical stock prices, including Yahoo Finance, NASDAQ, Shanghai Stock Exchange, NSE India, and APIs like Alpha Vantage and Tushare. These sources enable downloading

stock prices in CSV format or fetching data via APIs for ML model integration.

Stock Price Data Sources (US, China, India, Japan, Iran, Australia)

- United States (NASDAQ, NYSE, S & P 500): Yahoo Finance, NASDAQ, NYSE, Alpha Vantage API, Quandl
- China (Shanghai, Shenzhen Stock Exchange, CSI 300): SSE, SZSE, JoinQuant, Tushare API
- India (NIFTY 50, BSE): NSE, BSE, Moneycontrol, MCX
- Japan (Nikkei 225): Tokyo Stock Exchange (TSE) and Nikkei Indexes(Nikkei) provide official market data.
- Iran (Tehran Stock Exchange): Tehran Securities Exchange Technology Management Co (TSETMC) offers historical and real-time stock price data.
- Australia (S & P/ASX 200): Australian Securities Exchange (ASX) and S & P ASX Indices (S & P) serve as primary sources for Australian market data.

Sentiment Analysis Sources (US, China, India)

Sentiment analysis enhances stock prediction by incorporating real-time investor opinions. Social media and finance forums such as StockTwits, Reddit, Weibo, and Moneycontrol provide valuable stock sentiment insights.

Traditional Statistical Time Series Models Technique

- United States: Reddit, Stock-Twits, Yahoo Finance Boards, Twitter (StockMarket, NASDAQ, NYSE)
- China: Xueqiu, Eastmoney, Weibo
- India: Moneycontrol Forum, Trade Brains, Twitter (NSE, BSE, StockMarketIndia)

Acknowledgements

It is my pleasure to present a Survey on “Machine Learning Techniques for Stock Market Prediction: Models, Challenges, and Future Directions”. I genuinely express my gratitude to my guide Prof. Dr. S. S. Sonawane, Department of Computer Engineering for her guidance and help. She has constantly supported me and has played a crucial role in the completion of this report. Her motivation and encouragement from the beginning till end made this survey paper a success. I would also like to thank my parents-in-law Dr. B. M. Dabade and Dr. S. G. Kejgir for their guidance and support.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding source

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Authors’ Contributions

Payal Hake and Sanket Dabade researched and analyzed the literature, and wrote the first draft of the manuscript. Sheetal and Kalyani provided guidance and support for ideating the paper. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

References

[1] M. M. Rounaghi and F. N. Zadeh, “Investigation of market efficiency and financial stability between SP 500 and London stock exchange: Monthly and yearly forecasting of time series stock returns using ARMA model,” Phys. A, Stat. Mech. Appl., Vol.**456**, pp.**10–21**, **2016**

[2] G. Bandyopadhyay, “Gold price forecasting using ARIMA model,” J. Adv. Manage. Sci., Vol.**4**, No.**2**, pp.**117–121**, **2016**

[3] A. Hossain and M. Nasser, “Recurrent support and relevance vector machines-based model with application to forecasting volatility of financial returns,” J. Intell. Learn. Syst. Appl., Vol.**3**, No.**4**, pp.**230–241**, **2011**

[4] J. Chai, J. Du, K. K. Lai, and Y. P. Lee, “A hybrid least square support vector machine model with parameters optimization for stock forecasting,” Math. Problems Eng., **2015.**

[5] J. Borade, “Stock prediction and simulation of trade using support vector regression,” Int. J. Res. Eng. Technol., Vol.**7**, No.**4**, pp.**52–57**, **2018**

[6] A. Murkute and T. Sarode, “Forecasting market price of stock using artificial neural network,” Int. J. Comput. Appl., Vol.**124**, No.**12**, pp.**11–15**, **2015**

[7] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, “Stock market index prediction using artificial neural network,” J. Econ., Finance Administ. Sci., Vol.**21**, No.**41**, **2016**

[8] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. Shahab, “Deep learning for stock market prediction,” Entropy, Vol.**22**, No.**8**, pp.**840**, **2020**

[9] A.Sherstinsky,“Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” Phys. D, Nonlinear Phenomena, Art. no. 132306, Vol.**404**, **2020.**

[10] G. Ding and L. Qin, “Study on the prediction of stock price based on the associated network model of LSTM,”Int.J.Mach.Learn.Cybern.,Vol.**11**, No.**6**, pp.**1307–1317, 2019.**

[11] Ji, A. W. Liew, and L. Yang,“A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast,” IEEE Access, Vol.**9**, pp.**23660–23671, 2021.**

[12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., Vol.**9**, No.**8**, pp.**1735–1780**, **1997.**

[13] V. Gupta and M. Ahmad, “Stock price trend prediction with long short term memory neural networks,” Int. J. Comput. Intell. Stud., Vol.**8**, No.**4**, pp.289, **2019.**

[14] V. R. Madireddy, “Stock market prediction in BSE using long short-term memory (LSTM) algorithm,” Int. J. Innov. Res. Comput. Commun. Eng., Vol.**6**, No.**1**, pp.**561–565**, **2018**

[15] C.R. Ko and H.-T. Chang, “LSTM-based sentiment analysis for stock price forecast,” PeerJ Comput. Sci., Vol.**7**, pp.**408, 2021.**

[16] Y. Li and Y. Pan, “A novel ensemble deep learning model for stock prediction based on stock prices and news,” Int. J. Data Sci. Anal., Vol.**13**, No.**2**, pp.**139–149**, **2021.**

[17] C. Kearney and S. Liu, “Textual sentiment in finance: A Survey Of Methods and models,” Int. Rev. Financial Anal., Vol.**33**, pp.**171–185**, **2014.**

[18] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, “Building emotional dictionary for sentiment analysis of online news,” WorldWideWeb,Vol.**17**, No.**4**, pp.**723–742**, **2013.**

[19] T. Matsubara, R. Akita, and K. Uehara, “Stock price prediction by deep neural generative model of news articles,” IEICE Trans. Inf. Syst., Vol.E101.D, No.**4**, pp.**901–908**, **2018.**

[20] Z. D. Aksehir and E. Kilic, “How to handle data imbalance and

feature selection problems in CNN-based stock price forecasting," IEEE Access, Vol.**10**, pp.**31297–31305**, **2022**
[21] J. Xue and B. Shen, "A novel swarm intelligence optimization approach: Sparrow search algorithm," Syst. Sci. Control Eng.,Vol.**8**, No.**1**, pp.**22–34**, **2020.**
[22] A. Fathy, T. M. Alanazi, H. Rezk, and D. Yousri,"Optimal energy management of micro-grid using sparrow search algorithm," Energy Rep., Vol.**8**, pp.**758–773**, **2022.**
[23] Rohit Kumar, Rohit Gajbhiye, Isha Nikhar, Dyotak Thengdi, Sofia Pillai, "Stock Market Prediction using Deep Neural Networks," International Journal of Computer Sciences and Engineering, Vol.**7**, Issue.**4**, pp.**24-28**, **2019.**
[24] Rohit Tetarwal, Rohit Tushir, "Simulation Based Exploration of Stock Market Using LSTM Model ," International Journal of Computer Sciences and Engineering, Vol.**11**, pp.**26-29, 2023**

AUTHORS PROFILE

Payal Hake received her B.Tech. (Computer Science and Engineering) degree in 2020 from SRTM University. She is currently appearing for Master of Computer Engineering at Pune Institute of Computer Technology, Pune. Her research area is Artificial Intelligence and Deep Leaning.

Sheetal Sonawane has completed her Ph.D. from COEP Technological University, Pune in Computer Engineering. She completed her Master's degree in Computer Engineering from Pune Institute of Computer Technology. She is Associate Professor at Computer Engineering department of PICT, University of Pune. Her areas of interest are data mining and biological networks.

Kalyani Waghmare is Assistant Professor in Computer department at PICT College, Pune. She has completed her Master's Degree in IT. She has completed her B.E in Computer Science and Engineering. Her research interest include: Data Mining, Distributed, Signal processing.

Sanket Dabade completed his B. Tech. from IIT Bombay in 2014. He qualified as a Chartered Accountant in 2019. He has cleared CFA Level 1 and 2 exams. He works as an investment professional at a Private Equity/Venture Capital firm investing in Fintech companies. His research interests include the intersection of Artificial Intelligence, Deep Leaning and Finance.

