AI powered stock price Prediction: Leveraging Machine Learning techniques for market forecasting

By

Karim Elerian

Seminar Paper

submitted to the Department of Business Informatics

at the Faculty of Management Technology

German University in Cairo

Student registration number/ ID: 55-25593

Date: 11th of December 2024

Supervisor: Associate Prof. Dr. -Ing. Maggie Mashaly

Table of Contents

Abbreviation List

ANN - Artificial Neural Network

AR - Autoregressive process

ARIMA - Autoregressive Integrated Moving Average model

ARL - Association Rule Learning

ARI-MA - Auto-regressive Integrated Moving Average

CNN - Convolutional Neural Network

CNN-BiLSTM-AM - Convolutional Neural Network - Bidirectional LSTM with Attention Mechanism

DPS - Dividend per Share

EPS - Earnings per Share

EWMA - Exponential Weighted Moving Average

GDP - Gross Domestic Product

HAM - Historical Average Model

IR - Interest Rates

LSTM - Long Short-Term Memory

LSTMRNN - Long Short-Term Memory Recurrent Neural Network

LS-SVM - Least Squares Support Vector Machine

MA - Moving Average

MAE - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

MBE - Mean Bias Error

ML - Machine Learning

MSE - Mean Squared Error

MVs - Macroeconomic Variables

P/E - Price-to-Earnings Ratio

RF - Random Forest

RL - Reinforcement Learning

RMSE - Root Mean Square Error

RNNs - Recurrent Neural Networks

RWH - Random Walk Hypothesis

SMA - Simple Moving Average

SVM - Support Vector Machine

UL - Unsupervised Learning

WMA - Weighted Moving Average

1.Introduction

In recent years, the volume of financial activities has risen alongside rapid economic growth, leading to increasingly complex trends in their behavior. Understanding these patterns and predicting future developments has become a key area of focus for both academic researchers and financial professionals. By employing various predictive methods, it is possible to gain insights into the broader trends of the financial market, while also helping investors and businesses make informed, strategic, data driven decisions at the micro level. Such predictions can ultimately guide investment strategies and optimize profit generation.

Financial data is often complex, incomplete, and noisy, making it challenging to predict market trends accurately. Traditional statistical methods, which rely on linear models, struggle to handle the nonlinearity of financial data. However, with the rise of machine learning techniques, there is a shift toward more effective prediction models. These methods are better suited to process large volumes of data and capture complex, nonlinear relationships, offering significant opportunities for more accurate financial forecasting. As a result, machine learning is becoming increasingly important for predicting financial market behavior and making data-driven decisions (Yu & Yan, 2020).

A key challenge, however, is determining which machine learning model is most suitable for the unique nature of financial data to achieve accurate stock price predictions. Therefore, the main goal of this thesis is to explore and evaluate the use of machine learning techniques for forecasting stock prices. To determine important success criteria, constraints, and opportunities, the study will examine important models like Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and LSTM networks. This study also aims to identify scientific gaps in current approaches and suggest topics for additional research, like how sentiment and macroeconomic data might improve prediction accuracy.

This study employs a literature review approach to explore the current state of research on machine learning-based stock price prediction. Through a systematic examination of existing literature, the thesis will assess various ML models' efficacy, identify key success factors, and highlight persistent challenges in the field. The scientific gap identified will guide the formulation of future research directions. The Literature Review is divided into multiple sections to provide a comprehensive

understanding of the relevant topics. Section 2.1 covers traditional stock prediction methods, financial market dynamics, and challenges like volatility. Section 2.2 introduces machine learning (ML), detailing its types (supervised, unsupervised, and reinforcement learning), deep learning, and evaluation metrics in the context of stock forecasting. Section 2.3 focuses on applying ML to stock price prediction, reviewing successful models and their challenges, and discussing obstacles like data quality and model interpretability.

2. Literature Review

2.1 Financial Market Stock Prediction

Forecasting stock prices is a crucial and complex task in the financial domain, as it plays a vital role in guiding investment decisions and mitigating financial risks. The dynamic and complex nature of stock markets, influenced by a multitude of interconnected factors, makes forecasting an inherently complex process. Over the years, researchers and practitioners have employed diverse methods to decode patterns and predict price movements. This section explores the core aspects of financial market stock prediction, focusing on the underlying dynamics of stock markets, traditional forecasting techniques, market volatility, and the relationship between risk and reward in stock prediction.

2.1.1. Stock Market Dynamics

Stock market dynamics refer to the factors that drive price movements in financial markets. According to (Sindhu, Bukhari, Sub-Campus, & Hussain, 2014), these factors can be divided into internal, external, economic and political. they conducted a structured survey collected from 55 investors involved in share business, and from data sourced from the Karachi Stock Exchange and sample enterprise annual reports, the researchers identified which factors most significantly affect stock prices. Internal factors, which are specific to a company and its operations, were analyzed, and the top two most influential internal factors out of 15 factors analyzed were found to be the Price-to-Earnings (P/E) ratio, which was considered the most influential and Earnings per Share (EPS) as the second most influential.

They also found that the most influential External factor was the rumors spread about a company. Other factors that drive stock price movements are economical factors as the supply of money has

a great impact on interest rates and price level demand was found to be the most influential economical factor. As to political factors, change in government policies affected stock price the most. In a different study conducted on the stock market of Nepal, the results revealed that Earnings per Share (EPS), Dividend per Share (DPS), market whims and rumors, and company profiles have a significant positive association with share prices. In contrast, interest rates (IR) and the (P/E) ratio showed a significant inverse relationship with stock prices (Thapa, 2019).

2.1.2. Traditional Stock Price Prediction Techniques

Various stock price prediction techniques exist, including the Autoregressive Integrated Moving Average (ARIMA) model. This model utilizes historical data and breaks it down into three components: an Autoregressive (AR) process, which incorporates memory of past events; an Integrated (I) process that stabilizes or makes the data stationary, thereby making it suitable for forecasting; and a Moving Average (MA) of the forecast errors. As a result, the accuracy of forecasts improves with the length of historical data, allowing the model to learn over time (Emenike, 2010).

Additionally, other statistical models, such as the Simple Moving Average (SMA), Weighted Moving Average (WMA), Exponential Smoothing, and the Naive Approach, were frequently employed in the past for predicting stock prices. However, these methods have been found to be less accurate due to the chaotic and nonlinear nature of the stock market (Bhattacharjee & Bhattacharja, 2019). Furthermore, the Random Walk Hypothesis (RWH) claims that the market fluctuates up and down randomly therefore stock prices cannot be predicted using past data as there are no predictable patterns to guide such predictions (Chitenderu, Maredza, & Sibanda, 2014).

2.1.3. Financial Market Volatility

Volatility is an important factor in financial markets since it is commonly used as a measure of uncertainty. It plays an important role in many investment decisions and portfolio strategies since investors and portfolio managers must estimate their risk tolerance. An accurate estimate of the volatility of asset prices over the investment period is helpful in assessing the risk involved. Additionally, as option prices are based on market volatility, volatility plays a crucial role in the pricing of derivative securities. Efficient risk management and option pricing strategies rely

heavily on precise volatility predictions. In the context of finance volatility is viewed as the variance or standard deviation of returns over a specific period (Poon & Granger, 2003).

Similarly, (Ladokhin, 2009; Granger & Poon, 2001) agree that volatility represents the dispersion or spread of outcomes, typically linked to the standard deviation of returns in financial contexts. While both sources acknowledge that volatility measures both positive and negative outcomes, they distinguish between volatility and risk. Volatility refers to the overall spread of returns, whereas risk generally focuses on the uncertainty of negative outcomes. There are many classes of volatility models which include Historical volatility models, which use historical data to predict future volatility.

These models are considered the simplest type it can be as simple as using the Historical average model (HAM) which is just the mean standard deviation over a certain period to predict future volatility to more complex models like the Exponential weighted moving average (EWMA) which assigns exponentially decreasing weights to past data points, giving more importance to recent observations. Another class of volatility models is the Black-Scholes formula and implied volatility class which are widely used for pricing European options, where the volatility parameter is a key input.

Implied volatility refers to the volatility implied by the market prices of options, and it is often used as a proxy for market expectations of future volatility. The concept of implied volatility is important because it reflects market sentiment and plays a crucial role in the pricing of derivative securities. The volatility smile phenomenon, where implied volatility is not constant across different strike prices, is a notable feature of options markets and is used to price more complex financial derivatives (Ladokhin, 2009).

2.1.4. Risk and Reward in Stock Prediction

Risk is defined as the volatility of unforeseen events, which might represent the worth of assets, equity, or profits. Businesses are subject to a variety of risks, which can be categorized into financial and commercial risks. Financial risk is the possibility of suffering losses as a result of a company's operations in the financial markets. Business risk is those that a business voluntarily takes on to gain a competitive edge and increase value for shareholders (Jorion, 2007). In this context, momentum strategies, which involve purchasing stocks that have previously performed

well and selling those that have underperformed, present an intriguing relationship between risk and reward. Historically, these strategies yield average returns of about 1% per month over time horizons of 6 to 12 months, demonstrating statistically significant and substantial payoffs.

However, the lack of a consistent risk-based explanation for the momentum effect poses a significant challenge to asset pricing theory. Several risk-adjusted criteria are used in momentum techniques to measure stock performance. Conventional techniques have depended on evaluating previous performance, yet these methods frequently ignore the related risks. The use of the Sharpe ratio, which evaluates predicted excess return in relation to standard deviation, to non-Gaussian distributions might result in incorrect asset selection choices. On the other hand, modifications based on varying degrees of risk aversion are possible with the STARR ratio, which compares predicted excess return to conditional value at risk (CVaR). By contrasting predicted tail losses at various confidence levels, the Rachev ratio (R-ratio) helps investors weigh potential gains against probable losses and provides additional insight into tail risk (Rachev, Jašić, Stoyanov, & Fabozzi, 2007).

2.2. Machine Learning

The primary objective of machine learning, which is a branch of computer science, is to enable computers to "learn" on their own without the need for explicit programming. When it comes to machine learning, "learning" is normally accomplished by enhancing task performance through "experience," which is typically associated with data fitting. It can be difficult to distinguish between machine learning and conventional statistical approaches because such distinctions are frequently based more on the historical background of a methodology than on blatant methodological distinctions. Despite these parallels, machine learning is different philosophically and practically because it focuses on big, high-dimensional data sets and typically prioritizes predictive accuracy over hypothesis-driven inference. Furthermore, a growing focus on "Big Data" draws attention to problems and datasets that machine learning algorithms can effectively address, but which traditional statistical techniques might not be able to (Bi et al., 2019).

2.2.1. Types of Machine Learning

Machine learning serves as a cornerstone of artificial intelligence, offering diverse approaches to analyze data, extract patterns, and make predictions. These approaches are broadly classified into three types: supervised learning, unsupervised learning, and reinforcement learning, each tailored to address specific challenges in data driven problem solving.

2.2.1.1. Supervised Learning

Supervised learning is a machine learning technique that focuses on learning a function that maps inputs to outputs through analyzing sample input-output pairs. This involves creating a function using labeled training data which consists of a set of training examples. In supervised learning, algorithms rely on external assistance, where the input dataset is divided into training and testing sets. The training set includes an output variable that has to be predicted or classified. To make predictions or classify data, the algorithms find patterns in the training dataset and apply them to the test dataset. There are many types of models that are considered supervised learning including Decision Trees and Naive Bayes classification.

A decision tree is a graphical depiction, organized like a tree, containing options and potential results. The decision rules or circumstances that direct the decision-making process are represented by the edges of the tree, whilst the nodes within it reflect options or occurrences. A decision tree's nodes represent attributes from the dataset being classified, and its branches represent the many values that an attribute may have. The tree's ability to separate data systematically according to circumstances is made possible by its hierarchical structure, which eventually aids in tasks involving regression or classification.

The Naive Bayes classification method relies on the assumption of predictor independence and is based on the Bayes Theorem. In essence, a Naive Bayes classifier assumes that a feature's existence in a class is independent of the existence of any other feature. Naive Bayes performs effectively in a variety of applications, especially text classification tasks, despite this strong assumption. It makes predictions using the conditional likelihood of events occurring and is commonly used for both classification and grouping (Mahesh, 2020).

2.2.1.1.1. Support Vector Machines (SVM)

SVMs are supervised learning models used for classification and regression problems that employ a collection of linear functions in a high-dimensional feature space. The method uses machine learning methods to improve prediction accuracy while reducing overfitting. SVMs use an optimization-based learning algorithm based on statistical learning theory, attempting to achieve the best balance of model accuracy and generalization by setting optimal decision limits for classification, SVMs contain kernel functions which are crucial because they transform input data into a higher-dimensional feature space, where non-linear relationships can be modeled using linear decision boundaries. The kernel allows SVMs to accurately capture these intricate patterns by mapping the data into a space where they can be linearly separated.

SVMs are flexible models that work well for both regression and classification issues. By identifying the best hyperplane in a feature space that divides the classes, SVM learns to map input data to predefined classes for classification. This procedure frequently involves feature extraction or selection, which helps identify the essential traits that set the classes apart. SVM models continuous data with linear or non-linear mappings to a higher-dimensional space using a modified loss function for regression tasks. The kernel approach, which is frequently applied in both situations, enables SVM to efficiently manage intricate, non-linear interactions (Jakkula, 2006).

The ARI-MA-LS-SVM model, which was created especially for stock market forecasting, is another form of SVM regression. To identify patterns and remove noise from stock data, it combines the Auto-Regressive Integrated (ARI) and Moving Average (MA) techniques for data preprocessing. To get more precise predictions, Least Squares Support Vector Machine (LS-SVM) regression is used after preprocessing. By lowering dimensionality and increasing prediction accuracy, this combination strengthens the model's capacity to manage the intricacies of stock price prediction (Xiao, C., Xia, W., & Jiang, J., 2020).

Choosing the right kernel function is one of the most significant challenges when using SVMs because common options like polynomial or Gaussian kernels are not always effective with certain data structures. To properly capture underlying patterns in complex or non-standard datasets, customized kernels will likely be required. Furthermore, SVMs depend on efficient feature mapping to prevent problems with high dimensionality, even though they perform well with high-

dimensional data. Scalability and ease of training are two advantages of SVMs; nonetheless, in order to maximize performance, a carefully selected kernel function is necessary (Jakkula, 2006).

2.2.1.1.2. Random Forest

Random forests are a supervised machine learning technique. They were developed using an approach that creates several decision trees from random samples of data, which was originally proposed to increase accuracy. Each tree is constructed using a distinct subset of the data, and the predictions are then averaged in a procedure known as bagging. This method, known as the Random Forest algorithm, increases the model's accuracy and prevents overfitting, making it applicable to both classification and regression applications. Individual decision trees are easily interpretable, but this interpretability is lost in random forests because many decision trees are aggregated. However, in exchange, random forests often perform much better on prediction tasks (Schonlau & Zou, 2020).

Random forests encounter multiple other obstacles, particularly when it comes to tuning parameters. Because they rely on the dataset and performance objectives, the ideal hyperparameter values must be carefully chosen. Improper tuning can cause overfitting, in which the model performs well on training data but badly on new data. It can also be difficult to tell the difference between algorithm variations and hyperparameters. The tuning and selection procedure is a little unclear because parameters like the splitting rule and the number of trees, for instance, can be regarded as either a hyperparameter or an alternative model variation. Moreover, more trees are required to improve prediction clarity and ensure convergence in random forests. This results in an increased need for more trees to achieve optimal performance. However, as the number of trees rises, so does computation time, which grows linearly making it computationally expensive (Probst, Wright, & Boulesteix, 2019).

However, the random forest's classification performance might not be improved by adding trees beyond a predefined threshold. Using random forests for regression is particularly effective for learning from limited samples in logging regression modeling, unlike other algorithms that struggle with insufficient data. Additionally, it is more resilient to large data errors than other algorithms, which could be significantly damaged by logging data inaccuracies. Furthermore, alternative algorithms may have difficulties with smoothness, continuity, and extrapolation in

logging regression applications, but random forests are well-suited for nonlinear or nearly linear conditions (Ao et al., 2019).

2.2.1.1.3. Long Short-Term Memory (LSTM) Networks

The Long Short-Term Memory (LSTM) network was created to address the vanishing gradient problem that impacts regular Recurrent Neural Networks (RNNs). In order to enable a steadier gradient flow, LSTMs add a special structure with nonlinear, data-driven controls inside the RNN cell. The gradient of the objective function with respect to the state of the cell which is crucial for parameter updates during gradient descent remains constant and does not decrease over time because of the LSTM cell's implementation of these controls. To construct a robust system that can retain information across longer sequences, the fundamental RNN architecture is modified. The output signal of a typical RNN is a modified representation of the internal state of the cell, and the behavior of the gradient is influenced by this feedback loop across successive time steps. In standard RNNs, this dependency causes gradient stability concerns during training, resulting in either vanishing or exploding gradients, which LSTMs effectively handle.

An LSTM cell works by dividing its internal operations into two independent but related parts: data and data control. The control component creates "gate" signals, which range from 0 to 1, that control the amount of the candidate data that is passed forward, while the data component creates the candidate data signals, which have values between -1 and 1. The cell calculates the percentage of data that can go on to the next cell state by multiplying the candidate data by the control signal(Sherstinsky, 2020).

LSTM networks have a substantial benefit in dealing with long sequences of data because they can distinguish between recent and earlier inputs, assigning different weights to each while avoiding information deemed unnecessary for predicting the next output. This distinguishing feature allows LSTMs to handle longer input sequences more successfully than classic recurrent neural networks, which are often limited to remembering only short sequences (Nelson, Pereira, & De Oliveira, 2017). This is particularly important in stock prediction, where the ability to detect and incorporate both historical trends and recent patterns can improve predicting accuracy. Stock prices follow a sequential pattern, with each value influenced by the previous price, making LSTM's memory capacity extremely useful in capturing these relationships.

2.2.1.2. Unsupervised Learning

Unsupervised Learning (UL) is a machine learning approach that finds patterns in datasets with unstructured or unlabeled data points. With this method, the system is given just the input data and not the associated output labels. Unsupervised learning does not require human supervision of the model. Rather, the system learns on its own by looking for patterns in the data. This way makes it possible for a machine learning model to function on data without the need for outside direction, which makes it essential for creating artificial intelligence systems that are capable of autonomous decision-making from vast volumes of unlabeled data. UL models are often employed to discover hidden patterns in large datasets and to categorize or cluster data points based on their similarities and differences. There are many types of unsupervised Learning models they include clustering, Association Rule Learning and Anomaly detection.

Clustering is a technique that organizes items into clusters based on similarities, sometimes referred to as cluster analysis. There are various types of clustering, such as probabilistic, overlapping, hierarchical, and partitioning clustering. Partitioning, such as K-means, divides data into clusters where each item can only belong to one cluster. Each data point begins in its own cluster in hierarchical clustering, and clusters are joined based on proximity to one another by iterative merging, creating a hierarchy. A data point with overlapping clustering might belong to more than one cluster; its membership value is determined by how many clusters it is associated with. Probability distribution is used to form clusters in probabilistic clustering. Different methods are used by each type of clustering based on the application and the data it is used for.

Association Rule Learning (ARL) is an unsupervised learning technique that identifies associations between variables in large datasets. ARL is focused on finding relationships between variables and can handle non-numeric data points, unlike many other machine learning techniques. In ARL, statements like “if/then” rules help reveal patterns within the data, supported by metrics such as support (how frequently the rule appears) and confidence (how often the if/then relationship holds). ARL is widely used in applications such as shopping cart analysis and web usage mining.

Anomaly detection is the technique of discovering outliers or anomalies in a dataset. These anomalies frequently indicate odd behavior, defective sensors, or unclean data that requires

cleaning up before more examination. When data points diverge from predicted patterns, anomalies can be identified. This could be an indication of illicit data transmission, for example, or irregular network activity. In domains such as military surveillance, fraud prevention, and intrusion detection, anomaly detection is frequently employed (Naeem, Ali, Anam, & Ahmed, 2023).

2.2.1.3. Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning problem in which an autonomous agent interacts with its environment, gathers information about its present state, and makes judgments. The environment sends out a reward signal, which may be favorable or negative, in response to these activities. Throughout the engagement, the agent's main goal is to maximize the expected cumulative reward signal. This type of learning is frequently applied to applications like gaming, autonomous systems, and robots that call for decision-making under uncertainty (Ernst & Louette, 2024).

One of the most popular reinforcement learning methods is Q-learning, which is distinguished by its off-policy approach. By dividing the behavior policy, the way the agent acts from the learning policy, the way it learns this approach employs off-policy control. Because of its simple Q-functions, Q-learning stands out from other reinforcement learning algorithms and serves as the basis for many other RL algorithms. However, reward storage constraints posed serious problems for early iterations of Q-learning. It gets harder to complete complex learning tasks when the number of alternative actions increases, and the amount of storage needed becomes inadequate. More specifically, learning gets more inefficient in systems with big state-action spaces. Different Q-learning algorithms have been developed to increase performance in different circumstances to meet these problems (Jang, Kim, Harerimana, & Kim, 2019).

2.2.2. Deep Learning and Neural Networks

Artificial neural networks (ANNs) are machine learning algorithms modeled after biological neural networks. Every ANN is made up of nodes, which are like cell bodies, and communicates with other nodes through connections, which are like axons and dendrites. Similar to how, in a functioning brain network, connections between neurons are reinforced when their outputs are linked. An ANN's node-to-node connections are weighted according to how well they can provide the intended result. A perceptron is a type of machine learning algorithm that uses a

line, plane, or hyperplane, respectively, to attempt to divide classes in two, three, or hyperdimensional spaces based on the data and their related targets that are fed into the algorithm. This method is similar to logistic regression because the sigmoid function transforms the features. But in contrast to logistic regression, it does not provide the likelihood that an instance belongs to a specific class rather, it just establishes class relationships.

An Artificial Neural Network (ANN) or multilayer perceptron is the name given to the model created when many perceptrons are connected. An input layer, one or more hidden layers, and an output layer make up an ANN in most cases. Up to three hidden layers can be found in simple ANNs, while deep neural networks might have tens or even hundreds of hidden layers, simple artificial neural networks (ANNs) can only have three. A feedforward neural network is an artificial neural network (ANN) that processes information in a unidirectional fashion, moving data from one layer to the next without providing feedback.

Using a feedforward Artificial Neural Network (ANN) for image identification tasks has limitations because each input corresponds to a single pixel and there are no spatial connections between neighboring nodes. Because of this, spatial context is lost, and nearby pixels—which are probably correlated are handled differently from farther away pixels. In order to overcome this problem, convolutional neural networks, or CNNs, were created, which maintain the spatial correlations between individual pixels. In contrast to conventional ANNs, CNNs convey information to individual nodes in the subsequent layer by using patches of pictures, as opposed to sending information to all nodes at once. An important aspect of image processing is that this method preserves the spatial context from which features were collected. Convolutional filters are used to process these patches; these filters can identify features like edges or textures (Choi, Coyner, Kalpathy-Cramer, Chiang, & Campbell, 2020).

2.2.3. Evaluation Metrics in Machine Learning

The effectiveness of machine learning (ML) algorithms is frequently assessed using a core set of metrics. The confusion matrix and ROC curve are important evaluation metrics for classification models, which predicts whether a condition will exist or not. The capacity of a model to distinguish between classes is quickly compared using the AUROC, whereas the ROC curve shows the true positive rate against the false positive rate. Mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) are metrics used to evaluate models

for regression tasks, which predict continuous variables (e.g., patient lifetime based on medical data).

Despite their technicality, the notion behind these measurements is simple. Lower values indicate greater quality. The MSE calculates the average squared difference between actual and forecasted values. In a similar vein, R2 evaluates how well the model explains the variation in the data, whereas MAE computes the average absolute difference. Furthermore, a popular method for assessing how well machine learning algorithms perform is cross-validation. Where Researchers divide the data into training and testing subsets rather than evaluating the model on an entirely fresh dataset.

Cross-validation ensures that a model avoids problems like overfitting or underfitting and helps ensuring that the model generalizes properly by training it on one subset and testing on another. The iterative procedure improves the model's capacity for prediction on a variety of datasets (Handelman et al., 2019). Other metrics include the F1 score which is a function of precision and is utilized when assessing effectiveness of an unbalanced dataset, the kappa statistic which is a measurement that indicates how closely the classifications made by the machine learning algorithm align with the data labeled as true positives, affecting the accuracy of the classifier as assessed by the expected accuracy (Naidu, Zuva, & Sibanda, 2023).

According to Bhandari et al. (2022) model performance was assessed using three key metrics: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and correlation coefficient (R), each of which provided distinct insights into model quality. RMSE, or the average squared difference between actual and anticipated values, captures the normal error magnitude in forecasts. Furthermore, MAPE provides a relative measure of error in the form of an average percentage, which is particularly valuable for evaluating forecast proportionality. Finally, the R value evaluates the linear correlation between actual and expected values; a high R implies that the predicted trend closely matches the actual price movements.

In order to account for stochastic changes, they ran each model several times. For real-world accuracy, they inversely converted the predictions from normalized data. The best model was chosen because it had the highest R value, which indicates little error and good trend alignment, and the lowest average RMSE and MAPE scores. The most trustworthy model for stock price prediction was found thanks to this rigorous evaluation process.

2.3. Using Machine Learning in Stock Price Prediction

Forecasting stock prices in financial markets is a challenging and important issue. Over time, an effective stock prediction model can reveal patterns and trends that might otherwise go overlooked, offering insightful information about market behavior. Machine learning (ML) has become an effective way to address this issue with the growing computer power available today, allowing for more precise forecasts therefore, improved financial forecasting decision-making (Polamuri, Srinivas, & Mohan, 2019).

2.3.1. Machine Learning Models for Stock Price Prediction

In one study, Vijh, Chandola, Tikkiwal, & Kumar (2020) utilized 10 years of historical stock data, from 2009 to 2019 obtained from Yahoo Finance. The dataset included stock information for five companies across various sectors: Nike, Pfizer, Goldman Sachs, Johnson & Johnson, and JP Morgan Chase & Co. For prediction purposes, only the daily closing prices were extracted from the dataset. To enhance predictive accuracy, six new features were engineered using the existing data which were the difference between the stock's high and low prices, The difference between the stock's close and open prices, Moving averages over 7, 14, and 21 days and the standard deviation of stock prices over the past 7 days

The study applied both Artificial Neural Networks (ANN) and Random Forest (RF) models to predict closing prices. The comparative analysis across five companies revealed that ANN consistently outperformed RF. For instance, ANN achieved an RMSE of 0.42 for Pfizer compared to 0.43 using RF. Similarly, ANN demonstrated better MAPE and MBE values across all companies. The results highlighted ANN's ability to model nonlinear relationships more effectively than RF. However, the study identified limitations in its dataset, which relied solely on stock price and volume attributes. The authors suggested incorporating external data, such as financial news or macroeconomic indicators, to improve future prediction accuracy.

In another study, Ji, Wang, & Yan (2021) proposed a stock price prediction method leveraging deep learning techniques, which integrates social media sentiment analysis with financial indicators. The research analyzed data from the "Oriental Fortune" website, including over 530,000 social media documents, and transaction data from the Tushare financial database. Text features were extracted using the Doc2Vec model and refined through dimensionality reduction using a stacked autoencoder (SAE), while financial data was denoised using Haar wavelet transforms. The

proposed approach used a Long Short-Term Memory (LSTM) network for prediction, achieving superior performance compared to baseline models like ARIMA and RNN.

The LSTM model integrated with Doc2Vec and SAE yielded a mean absolute error (MAE) of 0.019, a root mean square error (RMSE) of 0.110, and an R-squared (R^2) value of 0.957. These metrics were significantly better than the LSTM model using only financial features, which had an MAE of 0.046, an RMSE of 0.579, and an R^2 value of 0.774. This demonstrated that incorporating investor sentiment and social media data substantially enhanced the model's predictive accuracy. However, limitations were noted, including the reliance on a single platform for sentiment data, which may not represent the broader investor population. Future research could include data from diverse platforms and explore multi-stock predictions to improve robustness.

A study by Nti, Adekoya, and Weyori (2019) on stock price prediction used historical data from 2002 to 2018 from a variety of sources, including the official website of the Ghana Stock Exchange. The dataset contained important information about price movements, opening and closing prices, past closing prices, annual high and low prices, and bid-offer prices for a variety of stocks in a number of industries. The Bank of Ghana and outside databases were also used to obtain macroeconomic variables (MVs), such as GDP figures and inflation rates, which gave the stock price prediction model a more comprehensive framework.

The predictive model was constructed using a three-phase approach. The required datasets were collected and combined in the first stage. To determine the most important predictors of stock price changes, the second phase concentrated on feature selection utilizing the Random Forest (RF) method and an improved leave-one-out cross-validation technique. In phase three, stock prices were predicted using a Long Short-Term Memory Recurrent Neural Network (LSTMRNN) model that had been trained on the most important attributes found in phase two. These assessment metrics were used to evaluate the LSTMRNN model to the traditional time-series model, ARIMA. According to the results, LSTMRNN performed better than ARIMA in terms of all error measures when it came to stock market price prediction.

In particular, the accuracy of the LSTMRNN model was higher (89.57%) than that of ARIMA (62.34%), indicating that LSTMRNN was more adept at capturing the intricate, nonlinear linkages present in stock price fluctuations. Additionally, by applying feature selection using the Random Forest algorithm, which found the most significant predictors, the model's predictive performance

improved. This led to a more effective model with less computing time and a 7.1% reduction in prediction errors. The study also found that macroeconomic variables, such as currency rates and Ghana's net foreign assets, had a substantial correlation with stock prices in various industries, indicating that macroeconomic factors are important in predicting stock market volatility, especially in Ghana. Recommendations for future work included incorporating socioeconomic and behavioral factors, as well as expanding the feature pool to include social media sentiment and web financial news.

Madhu, B., Rahman, M. A., Mukherjee, A., Islam, M. Z., Roy, R., & Ali, L. E. (2021) conducted a comparative study evaluating the predictive performance of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for forecasting stock option prices. This study used a dataset from Yahoo Finance to perform stock price prediction experiments, with a particular focus on 2015 SPY option prices. Performance optimization involved trial and error adjustments to SVM kernel functions and ANN model settings, including regularization parameters and the number of training cycles.

The results of testing both models with the improved parameters revealed that the ANN model performed better than the SVM model. In particular, the SVM model recorded a testing RMSE of 0.409254, whereas the ANN model obtained a training RMSE of 1.743 and a testing RMSE of 0.274418. According to these results, the ANN model performed better at forecasting the price of stock options. Despite this, the study acknowledged the need for further research to understand ANN's limitations and explore potential improvements in option price prediction.

In a study by Xiao, Xia, and Jiang (2020), stock price prediction was carried out using a combination of the Least Squares Support Vector Machine (LS-SVM) and an auto-regressive integrated moving average (ARI-MA) model. The dataset comprised stock data from Auto-desk (002227) over 102 trading days in 2015. The prediction model was implemented using the LS-SVM Lab toolbox, a simple yet effective tool for pattern recognition and regression within MATLAB. The results indicated that the ARI-MA-LS-SVM model, which incorporated attribute reduction, outperformed the basic LS-SVM model in terms of prediction accuracy and error reduction.

For example, the Mean Squared Error (MSE) for ARI-MA-LS-SVM was 0.015, compared to 0.0174 for the LS-SVM model and 0.0342 for the RS-SVM model. Additionally, the Root Mean

Squared Error (RMSE) for ARI-MA-LS-SVM was 0.108, which was substantially lower than the other models. The experiment demonstrated that ARI-MA-LS-SVM not only achieved lower error rates but also showed more stability in its predictions, particularly in small- and medium-sized stock markets. This model's superior performance suggests that attribute reduction, using rough set theory, can significantly enhance stock price forecasting. However, it also suggested areas for further improvement, including the optimization of SVM kernel parameters and the potential for real-time prediction via integration with quantitative trading platforms.

In a study by Lu, Li, Wang, and Qin (2021), a novel approach called CNN-BiLSTM-AM was proposed for stock price prediction which is a hybrid deep learning model that combines three components: Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Attention Mechanism (AM). its performance was compared with several other methods, including MLP, CNN, RNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, and BiLSTM-AM. The experiment used daily trading data from the Shanghai Composite Index (000001), between 1991 and 2020, obtained from the Wind database.

The results revealed that the CNN-BiLSTM-AM model consistently outperformed all other methods in terms of prediction accuracy. Among the eight models, CNN-BiLSTM-AM showed the highest degree of fitting between the predicted and real values, with the least error and the highest R^2 value of 0.9804, indicating its superior predictive power. The comparison of error metrics showed that CNN-BiLSTM-AM achieved the smallest Mean Absolute Error (MAE) of 21.952 and Root Mean Square Error (RMSE) of 31.694, while the R^2 value was closest to 1. The study also noted that LSTM outperformed RNN, as its MAE and RMSE were lower, and its R^2 value was higher. Furthermore, BiLSTM improved upon LSTM by reducing MAE and RMSE, and increasing R^2.

Incorporating the CNN layer further enhanced performance, with CNN-BiLSTM achieving even lower error metrics and higher R^2. The introduction of Attention Mechanism (AM) to the CNN-BiLSTM model further refined predictions, with improvements in MAE, RMSE, and R^2. However, they mentioned recommendations for future research which are that it should focus on fine-tuning the model's parameters to improve the accuracy of the results. Additionally, it should explore the model's potential for application in various other time series prediction fields, such as forecasting gold prices, oil prices, weather patterns, and earthquake occurrences, among others. To provide a

concise comparison of recent studies on machine learning models for stock price prediction, Table 1 summarizes the methodologies, results, advantages, and disadvantages of six key research works.

Table 1: Summary of Key Studies on Machine Learning Models for Stock Price Prediction

Citation	**Methodology**	**Result**	**Advantages**	**Disadvantages**
Vijh et al. (2020)	Artificial Neural Network (ANN) and Random Forest (RF) with engineered features like moving averages and standard deviations.	ANN outperformed RF (e.g., Pfizer RMSE: ANN 0.42, RF 0.43). Better MAPE and MBE values for ANN.	ANN effectively models nonlinear relationships, RF aids in feature importance analysis.	Limited to price and volume attributes; no external data like macroeconomic indicators or news incorporated.
Ji et al. (2021)	LSTM combined with Doc2Vec for social media sentiment analysis, SAE for dimensionality reduction, and Haar wavelet transforms for denoising.	LSTM with sentiment data achieved MAE 0.019, RMSE 0.110, R^2 0.957, outperforming financial-only models (MAE 0.046, RMSE 0.579, R^2 0.774).	Integrated sentiment data significantly improves predictive accuracy.	Relies on data from a single social media platform, which may not be representative.

Nti et al. (2019)	LSTMRNN for prediction; RF for feature selection; used macroeconomic variables like GDP, inflation rates, and currency exchange rates.	LSTMRNN accuracy: 89.57% vs. ARIMA: 62.34%. RF reduced computing time and prediction errors by 7.1%.	Incorporating macroeconomic data adds robustness; RF improves efficiency and reduces errors.	Specific to Ghana; results may not generalize to other regions or markets.
Madhu et al. (2021)	ANN vs. SVM with hyperparameter tuning and feature optimization using SPY option prices dataset from Yahoo Finance.	ANN achieved testing RMSE of 0.274 vs. SVM's 0.409.	ANN better captures patterns in complex data, outperforming SVM in prediction accuracy.	Limited focus on stock option prices; requires further research to address ANN limitations.
Xiao et al. (2020)	ARI-MA-LS-SVM with rough set-based attribute reduction using MATLAB's LS-SVM toolbox.	ARI-MA-LS-SVM achieved lowest MSE (0.015) and RMSE (0.108), outperforming basic LS-SVM (MSE 0.0174)	Attribute reduction enhances model accuracy and stability, especially for small datasets.	Results are limited to a small stock market dataset; lacks exploration of real-time application.

		and RS-SVM (MSE 0.0342).		
Lu et al. (2021)	CNN-BiLSTM-AM (Convolutional Neural Network, Bidirectional LSTM, Attention Mechanism) compared to traditional and hybrid deep learning models (e.g., LSTM, BiLSTM).	CNN-BiLSTM-AM achieved highest R^2 (0.9804) with lowest MAE (21.952) and RMSE (31.694), outperforming other methods.	Superior accuracy and ability to model complex temporal relationships; Attention Mechanism enhances predictions further.	Computational complexity; requires fine-tuning and potential exploration for broader applications.

2.3.2. Challenges and Limitations in ML-based Stock Prediction

AI-based stock market prediction faces a number of challenges, including the stock market's inherent uncertainty, reliance on historical data, complex model structures, and overfitting. These difficulties make it more difficult for machine learning models to provide accurate stock price predictions. Significant barriers are also presented by problems including ethical consequences, human bias, regulatory concerns, and the high expense of putting AI ideas into practice. Despite these drawbacks, there are still chances to improve data quality, make AI models easier to understand, and combine machine learning with other technologies. In order to overcome these obstacles and raise the precision and dependability of AI-based stock prediction models, research is still ongoing (Jain & Vanzara, 2023).

Other challenges mentioned by (Soni, Tewari, & Krishnan, 2022) include Prediction accuracy is one of the difficulties encountered. Certain research have found that accuracy rates often range between 50 and 70 percent, even when models employ many technical indications. While adding more indicators can be beneficial, maintaining consistently high accuracy is still challenging. Moreover, feature modification may occasionally impair model performance. For example, when features are converted to binary values, important information is lost, which lowers the model's prediction ability.

The size of the dataset is another important consideration. The robustness and generalizability of many models are constrained by the fact that they are trained on comparatively small datasets. The prediction accuracy would probably increase with the use of a larger dataset. Furthermore, many models' emphasis on short-term forecasting (such projecting the closing price of the next day) makes it difficult to make long-term investment decisions, which require a different kind of analysis. The issue of overfitting and underfitting is also prevalent in ML-based stock prediction models. The complexity of these models and the challenge of selecting the appropriate features can lead to models that either fail to capture relevant patterns (underfitting) or perform well only on the training data (overfitting).

3. Research Gap

While machine learning models, such as Support Vector Machines, Random Forest, and Long Short-Term Memory networks, have been successful in predicting stock prices, there are considerable gaps in the utilization of diverse data sources, like including social media sentiment from multiple platforms and to include macro economical data from multiple countries. Most of the current models, such as ANN and RF models (Vijh et al., 2020), rely on price and volume attributes and do not take into consideration external data sources such as social media sentiment or macroeconomic indicators. Ji et al. (2021) showed the viability of integrating an LSTM model with sentiment analysis gathered from social media, achieving an R-squared of 0.957, MAE of 0.019, and RMSE of 0.110. That is significantly better than financial-only models, showing the strong influence that sentiment data can have on predictive modeling. However, focusing on just one social media source may provide a limited view and fail to reflect the full spectrum of sentiment across different platforms.

The integration of sentiment from multiple platforms, along with macroeconomic variables like GDP and inflation rates, could be added to improve the prediction accuracy and generalizability, as explored in studies such as Nti et al. 2019, who obtained an accuracy of 89.57% for the LSTMRNN model, which significantly outperformed other traditional models like ARIMA, whose accuracy was 62.34%. This further indicates the strength introduced by integrating macroeconomic data with machine learning techniques. However, their work was limited to one country, suggesting an opportunity to explore the integration of global economic factors for broader applicability.

This leaves a critical gap in the literature: existing research has yet to fully explore models that integrate global macroeconomic variables with sentiment data aggregated across multiple social media platforms. Such integration could provide a more comprehensive understanding of market dynamics, enhancing predictive accuracy and generalizability. The failure to account for the influence of global economic interdependence and diverse sentiment sources not only limits the robustness of existing models but also reduces their applicability in a highly interconnected and volatile market environment. Addressing this gap could unlock new possibilities for developing sophisticated predictive systems that better reflect the complexity of modern financial markets.

References

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering, 174*, 776-789.

Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications, 9*, 100320.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology, 188(12)*, 2222-2239.

- Bhattacharjee, I., & Bhattacharja, P. (2019, December). Stock price prediction: a comparative study between traditional statistical approach and machine learning approach. In *2019 4th international conference on electrical information and communication technology (EICT)* (pp. 1-6). IEEE.

Chitenderu, T. T., Maredza, A., & Sibanda, K. (2014). The random walk theory and stock prices: evidence from Johannesburg stock exchange. *The International Business & Economics Research Journal (Online), 13*(6), 1241.

Chan, L. K., Jegadeesh, N., & Lakonishok, J. (1996). Momentum strategies. *The journal of Finance, 51*(5), 1681-1713.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology, 9*(2), 14-14.

Ernst, D., & Louette, A. (2024). Introduction to reinforcement learning. Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P, 111-126.

Emenike, K. O. (2010). Forecasting Nigerian stock exchange returns: Evidence from autoregressive integrated moving average (ARIMA) model.

Granger, C. W., & Poon, S. H. (2001). Forecasting financial market volatility: A review. Available at SSRN 268866. https://doi.org/10.2139/ssrn.268866

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... & Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology, 212*(1), 38-43.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University, 37*(2.5), 3.

Jain, R., & Vanzara, R. (2023). Emerging Trends in AI-Based Stock Market Prediction: A Comprehensive and Systematic Review. *Engineering Proceedings, 56*(1), 254.

Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. *IEEE access, 7*, 133653-133667.

Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. International Journal of Crowd Science, 5(1), 55-72.

Jorion, P. (2007). *Value at risk: the new benchmark for managing financial risk*. McGraw-Hill.

Ladokhin, S. (2009). Forecasting volatility in the stock market. Unpublished Thesis, VU University Amsterdam, Faculty of Science.

Lu, Wenjie, Jiazheng Li, Jingyang Wang, and Lele Qin. "A CNN-BiLSTM-AM method for stock price prediction." Neural Computing and Applications 33, no. 10 (2021): 4741-4753.

Madhu, B., Rahman, M. A., Mukherjee, A., Islam, M. Z., Roy, R., & Ali, L. E. (2021). A comparative study of support vector machine and artificial neural network for option price prediction. Journal of Computer and Communications, 9(05), 78-91.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet], 9*(1), 381-386.

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.

Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference* (pp. 15-25). Cham: Springer International Publishing.

Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In *2017 International joint conference on neural networks (IJCNN)* (pp. 1419-1426). Ieee.

Nti, K. O., Adekoya, A., & Weyori, B. (2019). Random forest based feature selection of macroeconomic variables for stock market prediction. American Journal of Applied Sciences, 16(7), 200-212.

Poon, S. H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of economic literature, 41*(2), 478-539.

Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. Int. J. Recent Technol. Eng, 8(1), 1224-1228.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9*(3), e1301.

Rachev, S., Jašić, T., Stoyanov, S., & Fabozzi, F. J. (2007). Momentum strategies based on reward–risk stock selection criteria. *Journal of Banking & Finance, 31*(8), 2325-2346.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3-29.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena, 404*, 132306.

Sindhu, M. I., Bukhari, S. M. H., Sub-Campus, B. B., & Hussain, A. (2014). Macroeconomic factors do influencing stock price: a case study on Karachi stock exchange. *Journal of Economics and Sustainable Development, 5*(7), 114-125.

Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: a systematic review. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012065). IOP Publishing.

Thapa, K. B. (2019). Influencing factors of stock price in Nepal. *NCC Journal, 4*(1), 113-120.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6), 1609-1628.

Xiao, C., Xia, W., & Jiang, J. (2020). Stock price forecast based on combined model of ARI-MA-LS-SVM. *Neural Computing and Applications, 32*(10), 5379-5388.

Declaration

I herewith declare that this report is in full accordance with the Plagiarism Guidelines of the Faculty of Management & Technology at the GUC.

Karim Elerian