Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

# Predictive Analysts Using Different Algorithm Machine And Deep Learning For Financial Market: Review Paper

Samah s Laftah[1,*], Sinan A Diwan[2]
[1]Department of software, Faculty of Computers & Information, Wasit University, Wasit, Iraq E-mail: Samah.bb67@gmail.com
[2] Department of software, Faculty of Computers & Information, Wasit University, Wasit, Iraq E-mail: sdiwan@uowasit.edu.iq
Samah.bb67@gmail.com

https://doi.org/10.46649/fjiece.v3.2.18a.27.5.2024

***Abstract.*** *This paper presents the development and effectiveness of deep and machine learning techniques in forecasting stock market trends. This paper review focus on key developments from 2020 to 2024 involves the integration of ensemble learning, sentiment analysis from social media, and various predictive algorithms such as LSTM, CNN, and ANN. These techniques improves predication accuracy by efficient processing large data groups and detect nonlinear patterns in stock price trends. in addition, the paper presented background about various machine learning algorithms such as KNN, naive Bayes, linear regression decision trees, SVM, random forests and deep learning models such as neural network, LSTM and CNN .the findings of this study to providing valuable insights to researchers and practitioners who aim to improve investment strategies and improve predictive accuracy By looking at the appropriate algorithm.*

***Keywords:*** *Stock price predication, Machine learning, CNN, LSTM, KNN.*

## 1. INTRODUCTION

The stock price represents the assessment of an individual stock within a company's shares that are available for public trading. Ownership of shares in a publicly listed firm confers individuals with partial ownership in that corporation. The original owners of the company initially offer shares for sale to obtain more investment, so enabling the company's expansion and development. An Initial Public Offering (IPO) refers to the process of selling stocks to the general public for the first time. More precisely, when there is a substantial demand for a specific stock resulting from a large number of purchasers in the market, the price of the stock will rise. In contrast, if the number of sellers exceeds the number of buyers, the stock price will decline as a result of an excess supply **[1]**.

Conventional method of stock price forecasting such as technical analysis used for analyse historical stock data by hand to identify trading opportunities. analyzing big amounts of data by hand can be challenging for investors and time consuming . To get over these limitations, researchers have changed to machine learning technique to help forecast stock prices. Machine learning technique can handle large amount of complex data, determined hidden patterns, and discovered complex relationships that hard to detect by hand **[2]**.

The aim of study is to evaluate and review models and algorithms used to predict stock price in the financial market, such as algorithms of machine and deep learning. The principal research question of this study is: "Which deep learning and machine learning algorithms provide the better accuracy to forecast stock price?"

The exploration of deep learning and machine learning techniques for predicting stock market trends has seen significant advancements over recent years. Starting in 2020, Isaac Kofi Nti et al. clarified the effectiveness of several ensemble learning techniques, such as stacking, bagging, boosting and mixing, when used with Neural Networks (NN), Decision Trees and Support Vector Machine. They showed that advanced clustering and mixing techniques could achieve forecast accuracies as high as 90% by analysing data from Bombay Stock Exchange (BSE-SENSEX), Ghana Stock Exchange (GSE), New York Stock Exchange (NYSE) and Johannesburg Stock Exchange (JSE). From January 2012 to December 2018. This work provided critical insights into the function of ensemble learning in improving stock market forecast accuracy, highlighting the significance of model selection within ensemble frameworks and launching a conversation on the computational efficiency trade-off **[3]**.

Building upon these foundations, M. Al Ridhawi in 2021 introduced a novel approach merging sentiment analysis from social media and financial stock data for stock market prediction. By employing an ensemble model that integrates Convolutional- Neural Network (CNN) models and Multi Layer Perceptron (MLP), LSTM, and focusing on stocks like AAPL, CSCO, IBM, and MSFT, Al Ridhawi's methodology showcased a prediction accuracy of 74.3%. This study underscored the value of merging financial data with sentiment analysis derived from social media, marking an important step forward in the predictive analysis of stock markets **[4]**.

Further expanding the scope of research, Bezan Lilauwala The potential of several machine learning algorithms involved Random Forest, LSTM, K-Nearest Neighbor Classification, and Linear Multivariate Regression, for stock price prediction was examined in 2022. Lilauwala's thesis demonstrated the complex strengths of these algorithms in predicting by using previous price data and financial technical indicators. It specifically highlighted the promising advantages of LSTM models when combined with historical and technical data. This research shed light on the difficulties involved in stock market forecasting in addition to underscoring the growing significance of machine learning in financial analytics **[5]**.

Continuing this trajectory of innovation, In 2023 Sama Hayder et al. delved into the predictive capabilities of LSTM and artificial neural network (ANN) algorithms for the Iraqi stock market index. Through a detailed examination over five years, their study found the LSTM algorithm to excel in forecasting accuracy, especially when the number of training iterations was optimized to 400. This research not only affirmed the LSTM's proficiency in stock market predictions but also opened avenues for future experiments across diverse datasets, further enriching the discourse on machine learning applications in economic forecasting **[6].**

In the 2024 study by Nazif Ayyildiz and Omer Iskenderoglu The usefulness algorithms of machine learning in predicting the movements directional of developed country stock market indexes is investigated in an article titled "How active is machine learning in predictions stock market?" The study tries to identify the most accurate prediction algorithm using decision trees, naive Bayes, logistic regression, k nearest neighbors, random forest, artificial neural networks and support vector machines. A wide range of developed economies are represented by the indexes analyzed, which include the FTSE 100, NIKKEI 225, CAC 40, DAX 30,NYSE 100, TSX and FTSE MIB. The results show that artificial neural networks are the better method for predicting movements in industrialized countries' stock markets, with the highest average prediction performance across many indices. This study emphasizes how machine learning can be used to improve **[7]**.

## 2. METHOD

In this part, we will explain data collection, pre-processing and the techniques applied in the studies mentioned previously and the most widely utilize to predict the stock price in the financial market**.**

## 2.1 Data Collection

For this study has been closed depend on a comprehensive review of relevant literature and studies in the same field. By focused on data from stock exchanges such as the New York Stock Exchange (NYSE), Bombay Stock Exchange (BSE-SENSEX), Iraqi stock exchange, and others.

## 2.2 pre-processing

Data preprocessing is a crucial step, particularly for projects involving large amounts of data. It entails transforming raw and random data to improve its quality by eliminating or cleaning out unwanted points, standardising it for everyday use, and enabling it to provide insightful information. The quality of the data has a greater influence than its quantity when it comes to producing excellent results. It includes organising, cleaning, scaling, normalising, and standardising data—that is, normalising and standardising data in addition to encoding data categorical. In project's pre-processing data phase unknown values, lost values and null values were cleared up, and any discrepancies were addressed and scale using Min-Max and standardise data. NumPy and Pandas were the two primary Python libraries used to preprocess the dataset; Matplotlib was utilised for data visualisation. When working with the datasets, NumPy fulfilled the roles of a scientific calculator, and Pandas library was appropriate for manipulation and data analysis. The data was visualised as charts using Matplotlib [8].

## 2.3 Assumptions Underlying Data Analysis Techniques

The data analysis technique utilize in this study are depend on certain assumptions, such as the independence of variables and the normal distribution of data**.** Data analysis techniques including

1. Efficient Markets Hypothesis

Several studies have examined the testing of the Efficient Market Hypothesis (EMH). E.F. Fama coined the phrase "efficient market" in a 1965 study, in which he posited that in efficient markets, competition would promptly incorporate the complete impact of new information on fundamental values into actual prices. It was hypothesized that as information emerges, it rapidly disseminates and promptly affects the pricing of shares. The implication of this hypothesis is that both technical analysis of stocks and fundamental study of firms would not result in significantly higher returns for investors compared to a simple "buy or hold" strategy .There are three types of EMH:

- Weak-Form Efficiency refers to the idea that the current price of an asset already reflects all relevant information from its past prices, and hence, past price information cannot be used to predict future prices. Consequently, it is not feasible to forecast the future gains of an asset by relying on technical analysis

- Semi-Strong Form Efficiency - This form of efficiency asserts that the present price of an item completely reflects all publicly available information. Public information encompasses historical stock prices, as well as the data disclosed in a company's financial statements, earnings and dividends announcements, the financial position of its competitors, and forecasts regarding macroeconomic conditions, among other things.

- Strong Form Efficiency refers to the concept that the current price of a financial asset reflects all available information, including both public and private information. This implies that no participant in the market can continuously generate profits by trading using non-public information [9].

**2**. Random Walk Theory

This theory presents price prediction in a distinct manner. The stock prices exhibit inconsistency in this technique, as they are derived from other attributes despite having the same distribution. Therefore, historical fluctuations or patterns in the value of assets cannot be employed to forecast future actions. In general, the random walk theory posits that the semi-strong efficient market hypothesis holds, meaning that public information is expected to be disseminated to all participants. This idea posits that the natural behavior of customers in the market follows a random walk pattern, indicating that making predictions is not feasible [10].

3. Analysis Philosophies

Recent research has shown that market prices do not adhere to the idea of random behaviour, indicating that it is indeed possible to predict financial markets. Within the realm of finance, two prominent trading theories exist. The fundamental analytical philosophy is based on the quantitative data of a corporation. The information may encompass sales data, the financial standing of the company, import/export volume, audit reports, corporate strength and investment, plant capacity, competition, terminal balance slips, and production indexes [11].

Conversely, the technical analysis concept utilizes previous time series data to forecast future prices. Investors typically appreciate the concept of following previous investor actions. Consequently, technical assessments are founded upon this concept. By analyzing the quantities and previous prices, one can identify profitable possibilities by comparing the average actions with the current volume and price. Technical analysis also produces several technical indicators (TIs) such as Moving Averages, Relative Strength, Rate of Change Index, Moving Average Convergence Divergence, Commodity Channel Index, and so on. These indications assist traders in determining if assets are oversold or overbought, as well as whether the trend is weak or strong. Machine Learning approaches have increasingly employed TIs as input for system prediction. This approach has the ability to recognize intricate patterns in data and make predictions about future prices and trends [12].

## 2.4 Machine Learning Techniques

Machine learning has a significant impact in the financial field, that is, in the field of predicting the trend of the stock market. Financial markets are complex networks that are influenced by several interrelated elements. The ability to predict future movements in these markets can lead to significant economic benefits. Stock prices and market behaviour are commonly modelled using machine learning techniques such as support vector machines, gradient boosting and neural networks. These models have the ability to analyze big amounts of financial data and determined valuable patterns that may not be obvious to human analysts **[13] .**

### 2.4.1 **K nearest Neighbour (KNN)**

K-nearest neighbour (kNN) algorithm is nonparametric technique, easy and effective utilize for regression and classification problems. The output of k-NN is decided by the k closest training samples in the feature space. This algorithm works on the assumption that items with same characteristics are in close proximity, this meaning that it forecast that comparable data points are in close proximity to each other. Proximity metric are sometimes obtained used a distance metric, such as the Hamming distance , Manhattan and Euclidean, based on the nature of the data **[14] .**

The effectiveness of the k-NN algorithm relies significantly on the choice of the parameter k (representing the number of nearest neighbours) and the distance metric used to calculate the proximity between data points. Choosing the right number for k is very important. A smaller value of k makes the algorithm more responsive to noise, but a larger k leads to increased processing complexity and perhaps reduced accuracy, as it may include features from other classes. Usually, the value of k is chosen through cross-validation to confirm that the model can be effectively generalized to different data sets **[15] .**

The k-NN algorithm has utilized in financial markets, especially for prediction stock market trends. By checking past stock data, the k-NN algorithm can forecast future changes in stock prices by determining same patterns in stocks. The capability of the k-NN model to effective deal with nonlinear interactions between variables, which is a typical widespread in financial markets. Accurate prediction of stock market trends used this algorithm requires accurate determination of relevant features and fine-tuning of parameters to account for the constantly changing nature of financial data, which is prone to high volatility and influenced by multiple external factors **[16] .**

### 2.4.2 Decision Tree

Decision trees are kind of supervised learning algorithm utilize for regression and classification problems. They are nonparametric, this meaning they do not put assumptions about the fundamental distribution of the data. They decisions and their potential outcomes represent in a tree , in this tree nodes representing the decisions or outcomes , and branches representing the conditions that lead to those decisions or events. This tree enables professionals to visualize the decision-making process in a clear and comprehensible way. Decision trees are widely used in different fields such as manufacturing, medicine, marketing, and finance to help stakeholders make informed decisions by analyzing complex data. The main algorithm behind decision trees is called the Classification and Regression Tree (CART) algorithm, originally proposed by Breiman et al. In his influential research **[17].**

Building a decision tree involves dividing the data into subsets using a set of conditional control statements called partition rules. The purpose of choosing these rules is to maximize the homogeneity of the generated subgroups. When it comes to classification activities, the degree of similarity within a group is usually evaluated using measures such as Gini impurity or entropy. On the other hand, in regression tasks, variability or mean square error is used. A decision tree starts with a root node that contains the entire data set. Subsequent divisions are performed recursively until a stopping criterion is met. One possible criterion might be the maximum depth of the tree, the minimum amount of samples allowed per leaf, or the threshold of improvement in the splitting metric. The use of recursive partitioning in decision trees makes them very consistent and robust, but sometimes prone to overfitting, especially when very deep trees are used **[18] .**

### 2.4.3 Naïve Bayes (NB)

Naïve Bayes algorithm is a classification method that applies Bayes' theorem to build Bayesian Networks from a given dataset. It functions on the premise that every characteristic in a class in the dataset is distinct from the others. For example, an object's classification of 'A' is determined by specific qualities. Even though these characteristics may rely on one another or other characteristics, their independent existence together affects the likelihood that the object will be classified as "A," which is why it is called "Naïve." The main benefits of the Naïve Bayes algorithm are its efficiency and simplicity, which allow it to work with big datasets and outperform more intricate classification techniques. These are the main steps that make up this algorithm **[19] .**

### 2.4.4 Linear Regression

In machine learning, the linear regression algorithm is categorized as supervised learning. Rather of forecasting categories, this method predicts values that are substantially inside the range. Because of the presence of outliers, it only establishes a linear relationship between the dependent and the independent variable and performs poorly with non-linear data sets. After using this algorithm to anticipate stock market movements, researchers found that there were significant issues that needed to be resolved when attempting to predict daily stock prices. Investors cannot consistently invest money using the prediction of this algorithm **[8] .**

### 2.4.5 Support Vector Machine (SVM)

A Support Vector Machine is a discriminative classifier that is represented by the separating hyperplane. Stated differently, the technique creates the optimal hyperplane that uses the provided labelled training data (supervised learning) to classify new samples. In two dimensions, this hyperplane is a line that divides a plane in half, with each class located on one side. Support Vector Machine (SVM) is thought to be among the best algorithms available for time series prediction. The supervised technique can be used for both regression and classification. What the SVM does is plot data as a point in an n-dimensional space. These dimensions are the features that are plotted on particular coordinates. **[20] .**

### 2.4.6 Random Forest

An ensemble technique called random forest builds a forest of categorization trees. For large-scale data, ensemble approaches typically outperform single component models. The variance is decreased when a forest is created because many trees are constructed using a subset of rows and columns, or in our example, corporate quarters and financial factors **[21]**. Random forest is an extremely effective technique for classification since it uses the average prediction across all of the forest's trees for categorization. Here, the forest is created using incredibly randomized trees **[22]**.

When a tree node splits and the attributes chosen for the tree are both highly randomised, these extremely randomised trees are produced. Test results indicate that while the bias increases slightly over standard trees, the variance vanishes. However, compared to other machine learning techniques, random forests have a significant drawback in that they are extremely sluggish. While random forests aren't the ideal model for all types of data, they perform better than **[23]** more classifiers such as support vector machine [24] as well as conventional neural networks In random forest trees, splits are performed by computing the Gini impurity (IG(p)), which is maximum in the case of a uniform class distribution **[25]**.

### 2.5 Deep Learning Techniques

Deep learning system strives to attain these characteristics without any human intervention. Machine learning operates like to a compliant automaton, adhering to instructions without question. Patterns within the data are examined to generate forecasts. Deep learning can be likened to a robot that possesses the ability to learn autonomously. It has the ability to acquire more complex patterns and produce autonomous predictions. Deep neural networks are a subset of machine learning. This is a network model composed of neurons with various parameters and layers positioned between the input and output. Deep learning utilizes neural network architectures as its foundation. As a result, they are referred to as deep neural networks **[26]**.

### 2.5.1 Artificial Neural Network (ANN)

Bioinspired algorithms have shown remarkable effectiveness in artificial intelligence. Numerous scholars have shown that applying bio-inspired algorithms has greatly improved the research domain result. ANNs, fuzzy systems, artificial immune systems, swarm intelligence, and evolutionary computation are a few of these techniques **[27]**. An algorithm that is bio-inspired and modelled after the central nervous system of the brain is called an artificial neural network (ANN). Input/signals from neighbouring neurons or the environment. This signal will fire under certain conditions, informing all other connected artificial neurons about the situation. **[28]**.

The artificial neuron gathers the incoming inputs by computing its net input signal as a function with the relevant weights. These net input signals are the input used by the activation function to calculate the output signal of the artificial neuron(s). One or more artificial neurons stacked in layers make up an artificial neural network (ANN). An ANN normally consists of an input layer, one or more hidden layers, and an output layer. ANN capabilities include learning ability, generalization, resilience, mapping abilities, and parallel information processing. The ANN's Architecture Many ANN architectures have been developed, such as feedforward, recurrent, and spiking NN. Moreover, additional varieties of neural networks (NNs) exist, such as self-organizing NNs, radial basis function NNs, temporal NNs, multi-layer perceptron and single-layer NNs **[29]**.

### 2.5.2 Long Short Term Memory (LSTM)

Long short Term memory (LSTM) networks are type of recurrent neural network (RNN) architecture that overcomes the limitations of traditional RNNs by efficiently capturing relationships of long term in sequential data. Lstm introduced by Hochreiter and Schmidhuber in 1997 and since they have becomes a important items in variant fields such as natural language processing ,speech recognition and time series forecasting. LSTMs well appropriate for applications including sequential data with distant connections **[30]**.

LSTM networks are consist of memory cells that control the flow of information inside the cell. Unlike traditional RNNs, which suffers from gradient regression when trying to train long range connections, LSTM make certain a continuous flow of error, this enabling LSTM to retain knowledge through extended periods of time. An LSTM cell comprises of different crucial elements: input gate, cell state, forget gate and output gate. Each of these components has a specific role in regulating the flow of information and maintaining relevant information **[31]**.

### 2.5.3 Convolutional Neural Network (CNN)

One sort of deep neural network that is most common used in analyze and perform artificial intelligence operation in visual images is the convolutional neural network, also known as Conv-Net or CNN. They are sometimes called space invariant artificial neural networks (SIANN) or shift invariant because it’s shared weights architecture and translation invariance. They offer natural language processing, image classification recommender systems, medical image analysis and image and video recognition applications. The term "convolutional neural network" refers to the usage of the convolution mathematical operation by the neural network **[32]** .

Two hidden layers input and output layers, make up a convolutional neural network. The convolutional and pooling layers that make up a CNN's hidden layers are usually arranged in a specific way before being flattened and succeeded by a fully linked network. Due to their outputs and inputs are masked by the activation function and final convolution, these layers are known as the hidden layers. In order to more precisely weight the final result, backpropagation is frequently used in the final convolution. Convolution layers are used to extract features from images by applying filters or kernels with particular dimensions less than the picture's dimensions over the whole image. The model may acquire these filters during training, or they may be pre-determined, such as the ones that identify edges and different shapesIn addition to the filter, a stride length is also applied to the picture. The number of pixels that can be moved across the input image is called the stride. With a filter size of 3 × 3 and a stride length of 2, Figure 4 illustrates the convolution operation performed on an image with dimensions of 7 × 7. In order to retrieve critical features or reduce the amount of extensive features extracted by the convolution layers, pooling layers are included. Typically, maximum pooling or average pooling are employed. The fully linked network then processes the fewer features, leading to speedy and accurate prediction generation. Figure 5 shows how the max pooling process works with stride length=2 and filter size 3× 3 **[33]**.

### 2.6 Ensemble Method

Ensemble methods are extensively employed in the fields of statistics and machine learning. EMs offer techniques for aggregating multiple individual classifiers or predictors in order to establish a committee and generate decisions that are more extensive and accurate than those made by any single classifier or base predictor [34].

EMs prioritize highlighting the favorable attributes of specific classifiers while minimizing their limitations. Opitz and Maclin categorize ensemble algorithms into two types: cooperative and competitive ensemble classifiers. Ensemble learning refers to the practice of training separate classifiers independently, using either the same or other datasets, but with different parameters. The final prediction, or anticipated output, is obtained by calculating the average of each unique single or base classifier output, or other similarities. On the other hand, the cooperative ensemble employs a strategy of dividing and conquering. The prediction task is divided into multiple subtasks, with each subtask being assigned to a specific classifier based on its unique traits and nature. The cumulative sum of the predictions from all subtasks yields the final prediction output [35].

### 2.6.1 Gradient boosting

Gradient boosting is a powerful method for machine learning. This method combines predictions from different models to improve the overall performance of classification and regression tasks. Gradient boosting is part of ensemble learning algorithms that combine their predictions to create models that are stronger than their individual components. Gradient enhancement involves gradually adding new models to deal with residual errors from previous models. It is widely used due to its ability to increase forecast accuracy and its efficiency in managing diverse data sources [36].

The basic concept of this algorithm arose from the realization that merging a more number of weak learners can generate a one resilient learner. Each learner in the sequence analyzes the errors of previous models and also makes adjustments to the future model to correct those errors. This iterative correction process continues until including more models no longer significantly improves performance. In gradient boosting, the term “gradient” refers to using the gradient descent process to minimize losses while combining new models [37].

#### 2.6.1.1 XGBOOST

A machine learning tool called Extreme Gradient Boosting (XGBoost) is built on massively parallel Gradient Boosting. It is an improved version of the GB technique, an approach that uses residual optimization to achieve great efficiency, flexibility, and portability. To make the predicted value of the tree group close to the true value, XGBoost creates K regression trees and offers parallel boosting trees. It can swiftly and precisely resolve a wide range of scientific data issues and has a great generalization capacity. The fundamental principle of XGBoost, an enhanced version of GBDT, is optimizing the objective function.[38].

Massively parallel Gradient Boosting is the foundation of a machine learning method called Extreme Gradient Boosting (XGBoost). It is an enhanced variant of the GB technique, which achieves remarkable efficiency, flexibility, and portability through residual optimization. XGBoost generates K regression trees and provides parallel boosting trees in order to get the predicted value of the tree group as near to the true value as possible. It has an excellent ability to generalize and can quickly and accurately answer a wide range of scientific data difficulties. Optimizing the goal function is the cornerstone of XGBoost, an improved form of GBDT.[39].

## 2.7 Comparative Analysis of Techniques

In this part, we offers a comparative analysis of different machine learning and deep learning techniques common utilize for financial market prediction. The techniques that is compared involve Linear Regression, K-Nearest Neighbours (KNN ),Decision Tree, Random Forest, Naive Bayes, Long Short-Term Memory (LSTM), Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGBoost) and Convolutional Neural Networks (CNN), and Each algorithm is evaluated depend on three main criteria: accuracy, speed (both in training and prediction phases), and interpretability.

- **Decision tree** this model provide a middle level of accuracy and are better interpretable, as their decision-making process can be visualize. They are also fast in phases of prediction and training, making them suitable for high decision making process.

- **K nearest neighbours (KNN )** this model provide high accuracy but suffer from slow performance in particular with large datasets, and this model have low interpretability because the complexity of distance calculations between data points.

- **Linear regression** this model have high interpretability and speed, while it typically struggles with achieve better accuracy in complex's stock market forecasting.

- **Naïve Bayes** this model provide fast calculations and moderate accuracy, also medium interpretability because their probabilistic nature.

- **Random forest** this model have high accuracy and reasonable speed, while their interpretability is moderate because the ensemble nature of the model, which combined multi decision trees.

- **Artificial neural networks (ANN)** this model have high accuracy and capture complex patterns in data, while is slow to train and less interpretability.

- **Long short term memory (LSTM)**this model is suitable for time series data due to provide high accuracy for prediction, but are computationally intensive and hard to interpret.

- **Convolutional Neural Networks (CNN)** this model is also offers high accuracy, especially for data with spatial structures, but is slow and less interpretable.

- **Extreme Gradient boosting (XGBOOST)** this model have high accuracy and moderate speed, though its complexity minimize interpretability.

## 3. CONCLUSIONS

This paper has presented the various methods that are applied to forecast the stock market. Some deep and machine learning techniques used by researchers are presented in this paper. After studying the techniques presented in detail in this paper, we can conclude that these techniques are useful for predicting stock price accurately because they deal with non-linear relationships and that choosing the most efficient technique depends on whether the data is complex or simple, as well as on the speed or slowness of the model. This study summarize that in terms of high accuracy and if the patterns are complex, one can choose LSTM or ANN. In terms of speed and simplicity, one can choose decision tree or naïve Bayes, and for moderate accuracy and interpretability one can choose XGBOOST or random forest. As a result, researchers can benefit from this study in identifying the appropriate technique according to the type of research problem, and they can also learn about more machine and deep learning techniques.

## REFERENCES

[1] A. Tipirisetty, "Stock price prediction using deep learning," 2018.
[2] C. S. Ku, J. Xiong, Y.-L. Chen, S. D. Cheah, H. C. Soong, and L. Y. J. M. Por, "Improving Stock Market Predictions: An Equity Forecasting Scanner Using Long Short-Term Memory Method with Dynamic Indicators for Malaysia Stock Market," vol. 11, no. 11, p. 2470, 2023.
[3] I. K. Nti, A. F. Adekoya, and B. A. J. J. o. B. D. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," vol. 7, no. 1, p. 20, 2020.
[4] M. Al Ridhawi, "Stock Market Prediction Through Sentiment Analysis of Social-Media and Financial Stock Data Using Machine Learning," Université d'Ottawa/University of Ottawa, 2021.
[5] B. Lilauwala, "Stock Market Predicting Using Machine Learning and Data Analytics Techniques," California State University, Northridge, 2023.
[6] S. H. A. AlHakeem, N. J. Al-Anber, H. A. Atee, and M. M. J. J. o. T. Amrir, "Iraqi Stock Market Prediction Using Artificial Neural Network and Long Short-Term Memory," vol. 5, no. 1, pp. 156-163, 2023.
[7] N. Ayyildiz and O. J. H. Iskenderoglu, "How effective is machine learning in stock market predictions?," vol. 10, no. 2, 2024.
[8] M. Bansal, A. Goyal, and A. J. P. C. S. Choudhary, "Stock market prediction with high accuracy using machine learning techniques," vol. 215, pp. 247-265, 2022.
[9] L. R. Marwala, "Forecasting the stock market index using artificial intelligence techniques," 2010.
[10] E. F. J. F. a. j. Fama, "Random walks in stock market prices," vol. 51, no. 1, pp. 75-80, 1995.

[11] I. Klioutchnikov, M. Sigova, and N. J. P. c. s. Beizerov, "Chaos theory in finance," vol. 119, pp. 368-375, 2017.

[12] R. D. Edwards, J. Magee, and W. C. Bassetti, *Technical analysis of stock trends*. CRC press, 2018.

[13] S. Jansen, *Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python*. Packt Publishing Ltd, 2020.

[14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21-27, 1967.

[15] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 1992.

[16] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert systems with applications*, vol. 42, no. 4, pp. 2162-2172, 2015.

[17] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[18] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81-106, 1986.

[19] D. Mahajan Shubhrata, V. Deshmukh Kaveri, R. Thite Pranit, Y. Samel Bhavana, and P. Chate, "Stock market prediction and analysis using Naïve Bayes," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 4, no. 11, pp. 121-124, 2016.

[20] V. K. S. Reddy and K. Sai, "Stock market prediction using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, pp. 1033-1035, 2018.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[22] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3-42, 2006.

[23] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *arXiv preprint arXiv:1605.00003*, 2016.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273-297, 1995.

[25] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359-366, 1989.

[26] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2015.

[27] A. P. Engelbrecht, *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[28] R. E. Uhrig, "Introduction to artificial neural networks," in *Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics*, 1995, vol. 1: IEEE, pp. 33-37.

[29] J. Zupan, "Introduction to artificial neural network (ANN) methods: what they are and how to use them," *Acta Chimica Slovenica*, vol. 41, no. 3, p. 327, 1994.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[31] F. A. Gers, J. Schmidhuber, and F. J. N. c. Cummins, "Learning to forget: Continual prediction with LSTM," vol. 12, no. 10, pp. 2451-2471, 2000.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[33] S. Mukherjee, B. Sadhukhan, N. Sarkar, D. Roy, and S. De, "Stock market prediction using deep learning algorithms," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 82-94, 2023.

[34] L. Olson, N. Tjikhoeri, and E. J. a. p. a. Guzmán, "The Best Ends for the Best Means: Ethical Concerns in App Reviews," 2024.

[35] S. Anirudh R, N. Krishnan, K. J. N. Ravi Kumar, and K. Ravi Kumar, Solar Radiation Forecasting Using Gradient Boosting Based Ensemble Learning Model for Various Climatic Zones in India, "Solar Radiation Forecasting Using Gradient Boosting Based Ensemble Learning Model for Various Climatic Zones in India."

[36] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[37] J. Friedman, "Trevor Hastie i Robert Tibshirani, The elements of statistical learning, sv. 1," ed: Springer series in statistics Springer, Berlin, 2001.

[38] T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1-4, 2015.

[39] T. Liwei, F. Li, S. Yu, and G. Yuankai, "Forecast of lstm-xgboost in stock price based on bayesian optimization," *Intell. Autom. Soft Comput*, vol. 29, no. 3, pp. 855-868, 2021.

## Figure and tables

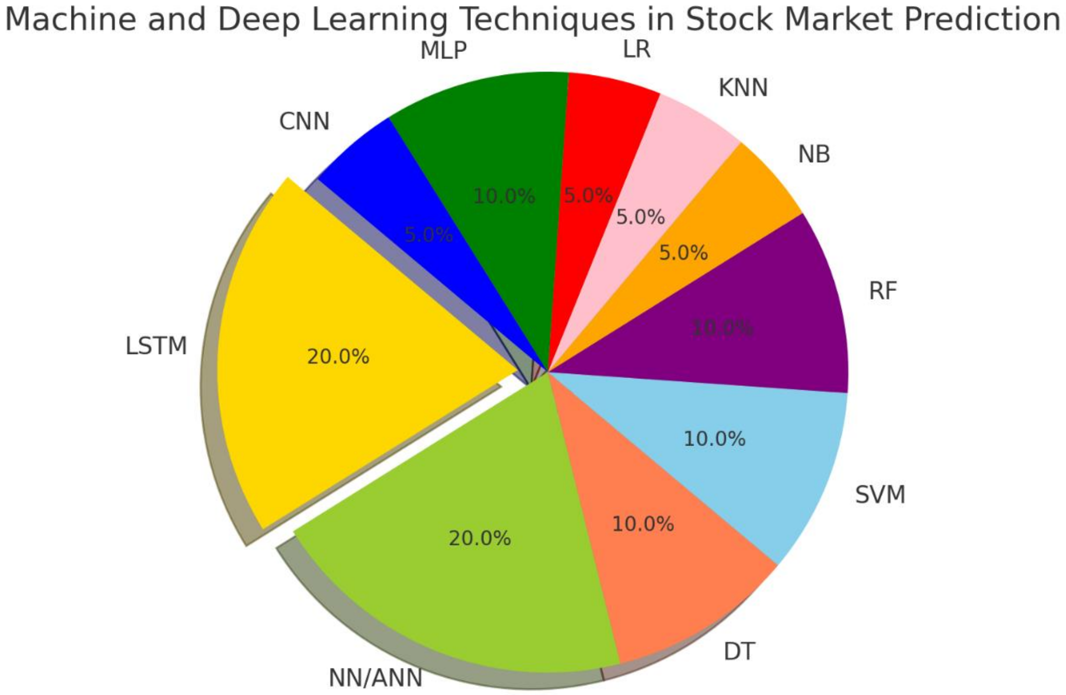

**Fig. 1. Summary of deep and machine learning techniques in studies**

**Table 1. Statistical Test and Rationale**

| Statistical Tests | Metric | Purpose |
|---|---|---|
| Descriptive Statistics | Mean, median, mode, standard deviation, variance, skewness, and kurtosis | To summarize and describe the main features of a dataset |
| Normality Tests | Shapiro-Wilk Test, Kolmogorov-Smirnov Test, Anderson-Darling Test | To determine whether the data follows a normal distribution |
| Correlation Analysis | Pearson Correlation, Spearman Rank Correlation | To measure the strength and direction of the relationship between two variables |
| Hypothesis Testing | t-test, ANOVA (Analysis of Variance) | To compare means between groups and determine if there are statistically significant differences |
| Regression Analysis | Mean square error, root mean square error | To model the relationship between a dependent variable and one or more independent variables |
| Goodness-of-Fit Tests | Chi-Square Test, R-squared (Coefficient of Determination) | To assess how well the model fits the observed data |
| Time Series Analysis | Augmented Dickey-Fuller Test, KPSS Test | To analyze time series data for stationarity and to identify trends, cycles, and seasonal effects |

**Table 2. Hyperparameters of machine and deep learning techniques**

| Model | Hyperparameters | Training details |
|---|---|---|
| **Decision tree** | Max_depth, min_samples_ split, min_samples_ leaf, max_features | 80 – 20 % split for training |
| **KNN** | N_nighbors, distance metric,weights . | 80 – 20 % split for training |
| **Random forest** | N-estimators , min_samples_ split, min_samples_ leaf, max_features | 80 – 20 % split for training |
| **Linear regression** | No hyperparameters for linear regression | reduces residual sum of squares between observed and predicted values using ordinary least squares (OLS) method |
| **Naïve Bayes** | No hyperparameters for naive bayes but for Multinomial Naive Bayes is Alpha | Calculate conditional probabilities for features that given each class and previous probabilities of each class. |
| **SVM** | Kernel, C (Regularization Parameter), Gamma, Degree. | Finding the better hyperplane that maximize margin between classes, including solve a quadratic programming problem. |
| **ANN** | Number of hidden layers ,Number of neurons , Activation function such as ( ReLU, sigmoid, tanh ) , Learning rate, Batch size ,Number of epochs ,Dropout rate and Optimizer . | Trained of this algorithm utilize backpropagation to reduce a loss function such as mean squared error for regression, cross-entropy for classification. |

| | | |
|---|---|---|
| **LSTM** | Number of layers , Number of units per layer ,Learning rate ,Batch size , Number of epochs ,Dropout rate and Activation function | Trained also utilize backpropagation through time (BPTT) to reduce loss. |
| **CNN** | Number of convolutional layers ,Number of filters per layer, Filter size (kernel size) , Pooling size ,Learning rate , Batch size , Number of epochs ,Dropout rate ,Activation function and Optimizer | Trained also utilize backpropagation through time (BPTT) to reduce loss. |
| **Gradient boosting** | Number of boosting stages (n_estimators) ,Learning rate , Maximum depth of individual estimators ,Minimum samples split, Minimum samples leaf ,Subsample and Loss function | This technique that build models sequentially, each trying to correct the errors of the prior models and Used decision trees as base learners. Trained utilize gradient descent to reduce specified loss function |
| **XGBOOST** | Number of n_estimators ,Learning rate , Max_depth ,Gamma and Lambda. | This technique that build models sequentially, each trying to correct the errors of the prior models and Used decision trees as base learners. Trained utilize gradient descent to reduce specified loss function |

**Table 3. Comparison between studies**

| References | Prediction Techniques | Stocks/Index | Performance metric | Result |
|---|---|---|---|---|
| (Isaac Kofi Nti et al., 2020) | ensemble learning methods | Ghana Stock Exchange (GSE), Johannesburg Stock Exchange (JSE), Bombay Stock Exchange (BSE-SENSEX) and New York Stock Exchange (NYSE) | Accuracy | Stacking :85%<br>Blending : 95% |
| (M. Al Ridhawi,2021) | Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), And Convolutional Neural Network (CNN) Models. | AAPI,CSCO,IBM,MST | Accuracy | 76% |
| ( Bezan Lilauwala ,2022) | Linear Multivariate, Random Forest, K-Nearest Neighbor | Yahoo Finance, FINTA APIS. | RMSE<br>MDA | Regression 1.31 |

| | | | | |
|---|---|---|---|---|
| | and Long Short-Term Memory. | | | LSTM:8.39%<br>KNN:53%<br>RF:57% |
| (Sama hayder,2023) | long short-term memory (LSTM) , artificial neural network (ANN) algorithm and CNN | IRAQI STOCK | RMSE | LSTM:0.04<br>CNN:0.03 |
| (Nazif Ayyildiz etal.,2024) | Decision tree, RF,KNN,NB,LR,<br>SVM,ANN. | NYSE, NIKKEI,FTSE, CAC, DAX, TSX. | Accuracy | DT:56%<br>RF:59%<br>KNN:50%<br>NB:62%<br>LR:82%<br>SVM:79%<br>ANN:83% |