Comparative Analysis of Hybrid Feature Optimizaton on Deep Learning and Ensemble MachineLearning in Stock Price Prediction

[1]*Shobhita Singh, [2]Divya Khanna, [3]B.S. Bhatia

*[1]*Research Scholar, Department of School of Computing, RIMT University, Mandi Gobindgarh, Punjab, India*
[2]Assistant Professor, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[3]Professor, Department of Management, RIMT University, Mandi Gobindgarh, Punjab, India

Abstract:- Data mining technologies that have given significant results and is used by researchers in a variety of fields is the neural network in the past. In today's economy, stock market data analysis and prediction play a crucial role. In this proposed study, the Apple Inc. (AAPL) stock data that is listed on NASDAQ stock exchange is considered and the stock's day wise closing price has been predicted and analysed using two proposed systems. The Proposed System1 integrated key technical indicators with traditional retrieved dataset and uses a combination of SelectKBest-RF fusion method for important feature selection and apply the hyper tunned parameterised-data to train the five ensemble machine learning models- Random Forest, Gradient Boosting, XGBoost, Stacking, and Voting to forecast the stock closing price. In Proposed System2, various volume related, and other essential technical indicators are computed and used as features to improve performance. This work also focuses on reducing the complexity, so a hybrid feature optimization technique (C-R-L) combining Correlation Analysis, Recursive Feature Elimination (RFE) and L1 Regularization is proposed for relevant and optimum feature selection. To anticipate the company's stock price based on existing historical data and computed indicators, three different types of deep learning architectures were employed: Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks with LSTM (CNN-LSTM). It is observed that CNN-LSTM outperformed the rest of the two models and the five ensemble models. Moreover, comprehensive comparative analysis has been performed for the validation.

Keywords: neural networks, stock market, ensemble, deep learning, technical indicators, volume related indicator, hybrid feature optimization

1. Introduction

The stock market is serving as a barometer for the financial health of a nation and providing a platform for companies to acquire capital and for investors to achieve financial gains. Despite its significance, the market is characterized by volatility and uncertainty, driven by a complex interplay of economic indicators, investors sentiment, and global events among other factors.

Stock price forecasting is the process of determining the value of a company's stock for future and other financial assets that are traded on the exchange [1]. The accurate prediction of a stock's future price could yield significant profit. However, this task is complicated by the fact that stock prices are not only influenced by the internal dynamics of the company but also by market sentiment, global financial trends, macroeconomic indicators, which are themselves influenced by unpredictable social, political, and environmental events [2].

Stock market analysis is divided broadly into two main analysis techniques: Fundamental analysis and Technical analysis. A fundamentally strong company can be kept for long term investment; because it involves evaluating a company's financial statements, health, competitors, and markets. However, technical analysis involving technical indicators focuses on patterns in price movements, trading signals, and other analytical charting tools to gauge the stock's strength or weakness for trading purpose [3].

In current scenario, the advent of machine learning and deep learning algorithms has transformed the landscape of financial forecasting. Researchers and financial analysts are increasingly turning towards sophisticated computational methods to predict stock prices. Techniques like neural networks, decision trees, and ensemble methods have been deployed to capture the non-linear relationships inherent in market prices and the multitude of factors that drive them [4][5]. Deep learning [6] has shown particular promise in this area due to its ability to process large volumes of data and recognize complex patterns that are not apparent to human analysts or traditional algorithms. This capacity makes it highly suitable for the dynamic and data-rich environment of stock market prediction, where traditional models struggle to handle the noise and volatility inherent in the data [7].

This paper aims to study the efficiency and effectiveness of hybrid feature optimization and deep learning models in predicting stock prices by conducting and comparing two proposed systems. The Proposed System1, uses a fusion feature selection technique and hyper-parameter tuning to train five machine learning ensemble models. The Proposed System2, introduces a hybrid feature optimisation and three deep learning techniques. We hypothesize that by combining multiple feature selection techniques and state-of-the-art deep learning architectures [8], we can significantly enhance the results and accuracy of stock price predictions compared to using conventional models alone.

1.1. Significance of the study

This research substantially improves the accuracy of current stock market prediction systems in forecasting stock closing prices with precision. Deep Learning (DL) is capable of analysing complex correlations and patterns in financial data, enabling it to adapt to volatile market conditions. The incorporation of hybrid feature optimization facilitates the identification of germane and significant attributes for DL models [9]. In addition, hybrid model of DL can improve the performance and enhance the system's capabilities.

1.2. Contribution of the paper

This work's primary contribution is the creation of a reliable and effective technique for predicting stock market closing prices that may benefit individuals as well as businesses. The following is a list of this paper's other technical contributions:

1.2.1. In the preliminary stage of the research, an experiment is conducted focusing on calculating the day-wise closing prices for Apple Inc. (AAPL). This task is accomplished using a variety of machine learning ensemble techniques, enhanced by the incorporation of multiple critical technical indicators and a combined feature selection technique. These indicators, extracted from the stock data, are utilized to improve the predictive accuracy of the models, and the results are computed.

1.2.2. Later, in our proposed study, a new experimentation using Volume oscillators integrated with traditional technical indicators associated with the selected stock is constructed, to create a robust framework for analysing stock price fluctuations. It provides critical insights into the strength of price trends and trader sentiments, hence enhancing trading decisions.

1.2.3. After the integration, a novel hybrid feature optimization technique (C-R-L) is introduced by combining filter-based methods (Correlation Analysis), wrapper-based methods (RFE), and embedded methods (L1 Regularization). By integrating these methodologies into a hybrid framework, one can capitalise on the respective advantages of each approach to enhance the feature selection procedure and the stock price prediction model.

1.2.4. The fourth significant contribution is the strategic integration of CNN-LSTM, a hybrid model along with traditional CNN and RNN techniques. This approach not only improves prediction accuracy but also brings adaptability and scalability to handling stock data, effectively leveraging both spatial and temporal aspects of data processing. The results of the proposed deep learning models are computed.

1.2.5. Finally, the outcomes of both the initial and advanced modelling techniques were compared and thoroughly analysed to assess the improvements and overall performance of the predictive models.

The entire study is divided into structured and number of sections. The Section 2 is the literature review to understand the current approaches and their challenges. Section 3 details the methodology employed in the first experiment, including system configuration and all pertinent experimental results and analyses. Section 4 details the methodology employed in the second experiment, including system configuration and all pertinent experimental results and analyses. Section 5 presents a comparative analysis of the two conducted experiments and their respective outcomes. Section 6 offers a comparative discussion in relation to the foundational paper. Section 7 describes the limitations of the proposed study. The concluding remarks are presented in Section 8.

2. Literature Review

This section includes the study of various stock price prediction techniques adapted in the past years to understand the current gaps and the current opportunities in the field. This section also provides an insight of state-of-the-art techniques and can be evaluated against the proposed work.

Two popular network models for stock market price prediction: Artificial Neural Network (ANN) and the Convolutional Neural Network (CNN), also known as the Deep Feed-forward NN [10][11]. Based on the parameter values from the previous days, the learning models are trained to forecast the stock's future price and trends. To optimize the stock prediction, implementation of deep learning has produced considerable outcomes. Other employed NN-based techniques (ANN, RNN, and LSTM) and tree-based techniques (Decision Tree, Gradient Boosting, Bagging, Adaboost, XGBoost and Random Forest) [12] [13]. An efficient CNN model based on 2D histograms was proposed that are generated from the quantized dataset within a specific time frame for stock price prediction[14], [15]. The day wise closing price data of iShares MSCI UK exchange-traded stock was utilised, and machine learning and deep learning models were employed on it [16]. [17] [18]Here, day-by-day closing prices of NSE and NYSE are being used. The ANN backpropagation and CNN model was trained using the NSE stock exchange price of an individual firm and was able to forecast the stock prices of five different NSE and NYSE companies. Two cutting-edge hybrid algorithms, CEEMD-CNN-LSTM and EMD-CNN-LSTM, that was used to extract time sequences and deep features before being applied to one-step-ahead prediction combined with LSTM [19][20]. Another investigation on how the network's overall capacity to forecast future market behaviour is affected by three unsupervised feature extraction techniques: limited Boltzmann machine, autoencoder, and principal component analysis and for input data, they used returns of high-frequency intraday stock [21][22]. [23] Proposed solution for predicting stock market price trends utilizing a customized deep learning-based system in conjunction with several feature engineering techniques and stock market dataset pre-processing. [24] provide an extremely reliable and effective methodology for predicting stock prices that combines deep learning, machine learning, and statistical models. [25], [26] applied four machine learning models on ten years of Apple Inc. stock data and RMSE was the evaluation parameter. The study is used for the validation purpose and is considered as the base paper. [27],[28] enhanced the algorithmic trading framework, by incorporating deep convolutional neural networks (CNN) and proposed planar feature representation methods. [29] applied machine learning techniques and evaluation demonstrates that the correlation between the trends of two months is minimal. [30] gives the importance of technical analysis, fundamental analysis, and combined analysis.

After reading through the research papers, it can be stated that: though the stock price prediction performance has improved through the current techniques and methodologies applied, problems like computational complexity and need for robust models to handle high-frequency trading data. The studies' published results haven't reached the necessary level of accuracy to be truly successful in the stock day trading.

3. Methodology of the Proposed System1

The purpose of the study is to explore and develop a system for predicting the day-wise closing price of the stock. This section will detail the methodologies employed in first experiments carried out. The Experiment1 involves predicting the day-wise closing price of the AAPL stock by applying various machine learning ensemble techniques like Random Forest, Gradient Boosting, XGBoost, Stacking and Voting, while incorporating the traditional historical data with some essential key technical indicators applying a (SelectKBest-RF) feature selection technique for retrieving relevant features. The primary stages of this system consist of the following: input, feature selection, training utilizing the five ensemble techniques, final predictions, and performance analysis phase.

3.1. Input Phase

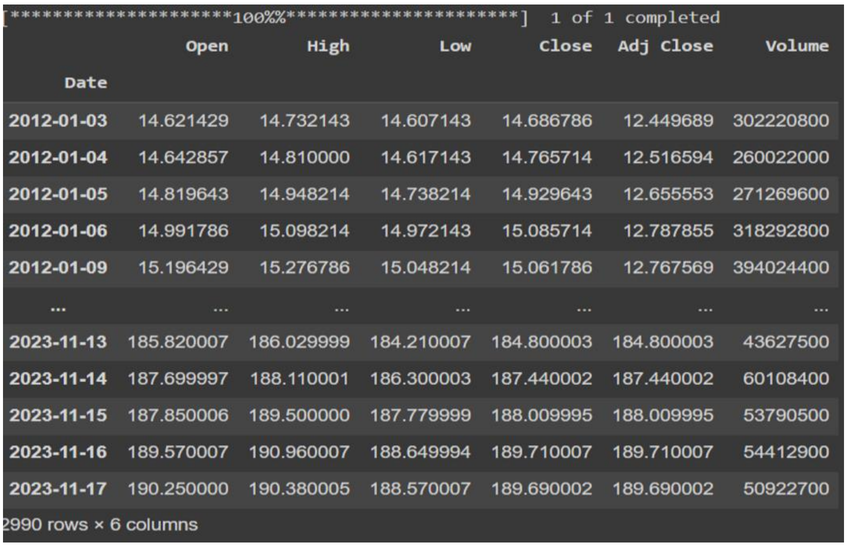
The dataset imported for this project consists of historical stock data for Apple Inc. (AAPL) spanning the past 10 years. It includes various attributes related to the stock's performance on different dates. Here's a description of the dataset:

- Date: Indicates the specific date of the stock record.
- Open: Reflects the price at which the stock began trading on that specific date.
- High: Indicates the maximum price that the stock touched during the trading day.
- Low: Tells the minimum price that the stock fell to during the trading day.
- Close: Reflects the price at which the stock closed trading on that specific date.
- Adj Close: Provides the closing price after adjustments of the stock, which accounts for corporate actions such as stock splits, dividends etc.
- Volume: Provides the count of shares that were traded on that specific date, indicating the level of interest and liquidity.

These attributes provide a detailed view of the stock's performance over time, capturing both price movements and trading volumes. The dataset is crucial for analysing historical trends, identifying patterns, and building predictive models to forecast future stock prices.

Experimental Dataset1

Historical Data
Eleven years data of Apple (AAPL), (a highly traded stock) is collected from yahoo finance starting from 01/1/2012 to 17/11/2023, glimpse is presented in Figure 1. The dataset contains 2990 rows and six columns including the information such as stock symbol, stock date, day-wise closing, opening, high, low, adjacent closing price and the volume traded. Stock day-wise adjacent closing price is considered and extracted from rest of the parameters because adjacent closing price gives the picture of last hour volatility movement of the stock. Thus, assists the investors in making timely decisions in terms of buying stocks.



```
[*********************100%%**********************]  1 of 1 completed
                  Open        High         Low       Close   Adj Close     Volume
      Date
2012-01-03   14.621429   14.732143   14.607143   14.686786   12.449689  302220800
2012-01-04   14.642857   14.810000   14.617143   14.765714   12.516594  260022000
2012-01-05   14.819643   14.948214   14.738214   14.929643   12.655553  271269600
2012-01-06   14.991786   15.098214   14.972143   15.085714   12.787855  318292800
2012-01-09   15.196429   15.276786   15.048214   15.061786   12.767569  394024400
       ...         ...         ...         ...         ...         ...        ...
2023-11-13  185.820007  186.029999  184.210007  184.800003  184.800003   43627500
2023-11-14  187.699997  188.110001  186.300003  187.440002  187.440002   60108400
2023-11-15  187.850006  189.500000  187.779999  188.009995  188.009995   53790500
2023-11-16  189.570007  190.960007  188.649994  189.710007  189.710007   54412900
2023-11-17  190.250000  190.380005  188.570007  189.690002  189.690002   50922700
2990 rows × 6 columns
```

Figure 1: Glimpse of AAPL stock dataset.

Technical Indicators:

Additionally, after feature engineering, the dataset will include additional technical indicators calculated from the original attributes, such as SMA, Exponential Moving Averages (EMAs), Moving Average Convergence Divergence (MACD), as shown in Table I.

Table I: Computed Technical Indicators

Technical Indicators	Description	Formula
Simple Moving Average (SMA)	SMA denotes the correlation between the volatility of the stock price and the corresponding movement of the moving average.	It is determined by averaging the closing prices of a security for the preceding "n" periods (number of days). SMA= $\frac{CP1+CP2+\cdots\ldots+CPn}{n}$ moving average with time period of 7 days. ma21- moving average with time period of 21 days.
Exponential Moving Average (EMA)	EMA is a price computation average that prioritizes the most recent price data over a specified time period.	EMA (current) = (multiplier of (Price(current) – EMA (previous)) + EMA (previous)) Constantly, the length of time has a significant effect on the weighing multiplier. 12ema- exponential moving average with 12-day time period. 26ema- exponential moving average with 26-day time period.
Moving Average Convergence-Divergence (MACD)	MACD is a leading momentum indicator that aims to predict stock market movements by	The MACD consists of three components: the MACD line (Fast Line), the Signal line (Slow Line), and the Histogram (showing

	analysing short- and long-term trends.	the difference between the MACD and Signal lines). MACD is determined by subtracting the 26-day Exponential Moving Average (EMA) from the 12-day EMA.

3.2. Feature Engineering and Selection

After data preprocessing, involving data cleaning, normalizing or scaling data, detecting and eliminating outliers. A combination of Random Forest and SelectKBest techniques are implemented for feature selection. This combination is applied as it allows the model to benefit from the robustness and depth of Random Forest while ensuring that individual features have statistically significant relationships with the target variable through SelectKBest, as shown in Figure 2.

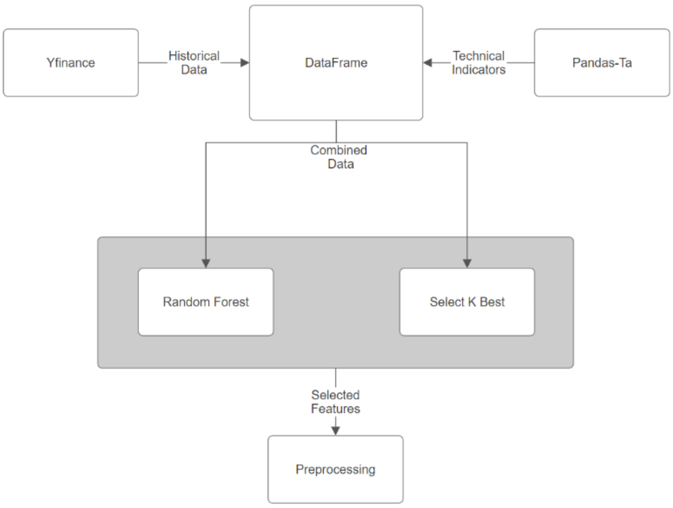


Figure 2: Feature Engineering and Selection

- **SelectKBest:** It evaluates each feature individually to determine the strength of the relationship of the feature with the response variable. In the proposed study, for the Proposed System1, Mutual Information is used as a statistical test to measure the amount of information that one variable contains about another variable. For two discrete variables X and Y, the formula is described using their joint probability distribution and their individual marginal probability distribution.
  I (X; Y) = $\int Y \int X\ p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)}\right) dx\ dy$
- **Random Forest for Feature Selection:** It provide the ranking of the important of features based on performance of the model. Here, accuracy parameter is considered to check the feature importance and select the most relevant features for model training.

Together, the combination significantly enhances the performance of the predictive model by capturing complex patterns and providing fast, univariate filter method.

3.3. Model Training

Once the crucial and relevant features are identified and pre-processed, the next step in stock price prediction involves training models using various machine learning ensemble techniques. The proposed techniques are Random Forest, Gradient Boosting, XGBoost, Stacking and Voting, as shown in Figure 3.

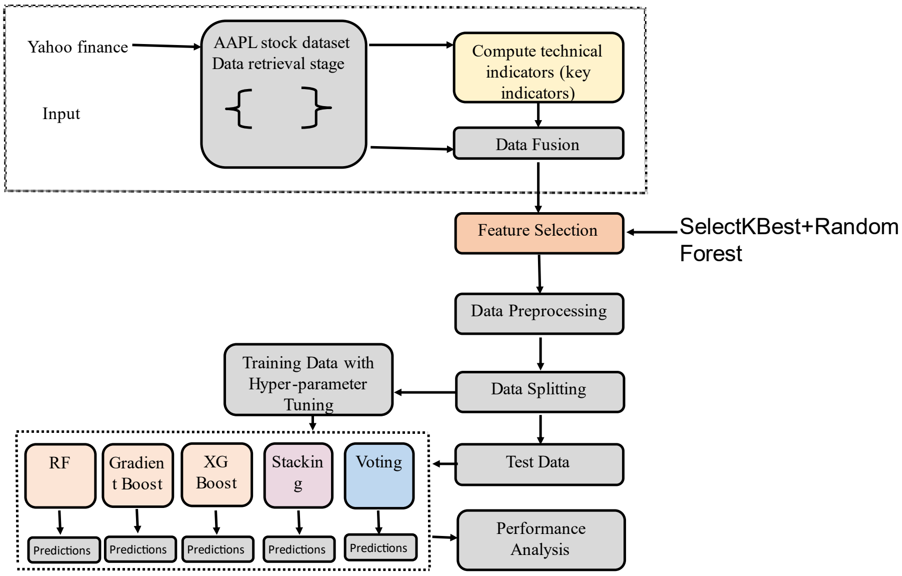


Figure 3: Methodology design of the proposed system1

3.4. Prediction Results and Analysis Phase

The experiment was conducted on a Windows system with an i7 processor and integrated GPU capabilities. Python programming language and the Jupyter Notebook platform are used to implement this proposed system.

The evaluation parameter that is used in understanding the different aspects of model performance is RMSE. The RMSE represents the root average squares residual and the result of RMSE is always non-negative (model whose results are near to zero shows better prediction quality) [21], [31]. The root of the average squared deviation between the predicted and observed values is computed. The RMSE formula is as follows:

$$\sqrt{1/\mathrm{N}\ \mathrm{n}\sum \mathrm{i} = 1\ (\mathrm{Ai} - \mathrm{Pi})2}$$

Table II displays the results containing RMSE value of proposed ensemble techniques applied on AAPL dataset. By integrating significant technical indicators with traditional historical data and applying fusion of the proposed feature selection techniques, it can be stated that Gradient Boosting technique has given the best output with minimum loss function.

Table II: Results of the Proposed System1

Ensemble Prediction Models	Proposed Parameters after Hyper- parameter tunning	RMSE

Random Forest	Max_depth:10, min_samples_split:2, n_estimators:100	1.9210
Gradient Boosting	Learning_rate:0.05, max_depth:5, n_estimators:300	1.5159
XGBoost	Learning_rate:0.1, max_depth:3, n_estimators:300	1.8537
Stacking	-	2.4017
Voting	-	1.8001

4. Methodology of the Proposed System2

The second experiment is done by applying three DL techniques- RNN, LSTM and a hybrid CNN-LSTM [32] and enhances its performance through a novel hybrid feature optimization. The overall design of the proposed system is illustrated in Figure 4. The primary stages of this system consist of the following: input, feature selection, training utilizing RNN and LSTM and CNN-LSTM, final predictions, and performance analysis phase.

4.1. Input Phase

The first phase is the input phase, in which the dataset of Apple Inc. stock, is extracted from Yahoo finance. The data retrieved from Yahoo Finances contains six attributes representing the price of opening, closing, high, low, adjacent close, and volume of the shares traded in an entire market during the given period.

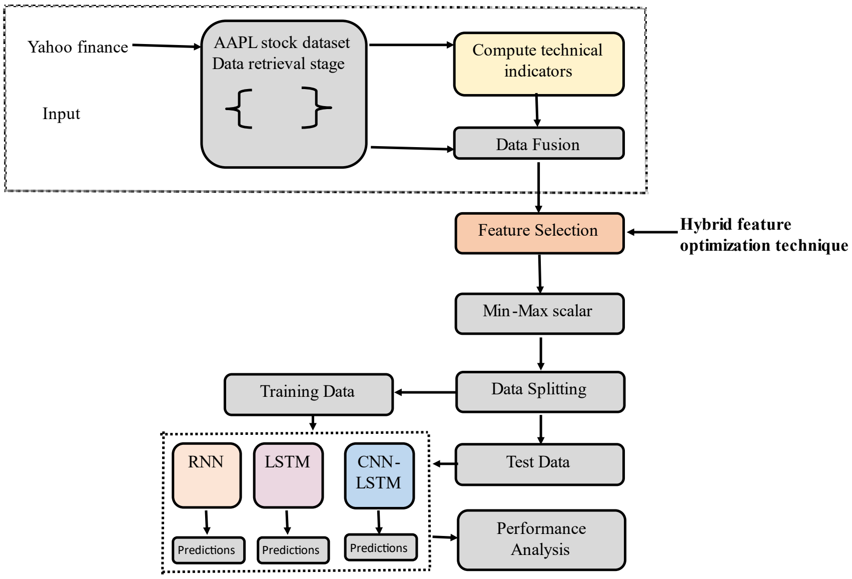


Figure 4: Methodology design of the Proposed System2

Experimental Dataset2

Historical Data

The dataset used is same as retrieved in Proposed System1, i.e. eleven years data of Apple (AAPL) from yahoo finance starting from 01/1/2012 to 17/11/2023. The dataset contains 2990 rows and six columns including the information such as stock symbol, stock date, day-wise closing, opening, high, low, adjacent closing price and the volume traded.

In this study, a small experiment over a short period of time (01/2023 to 11/2023) was also done comparing the actual adjacent close price with the predicted closing price by applying MA (Moving Average) for the look back period of 10 days, 20 days, and 50 days. Figure 5 displays the line chart showing the comparison.

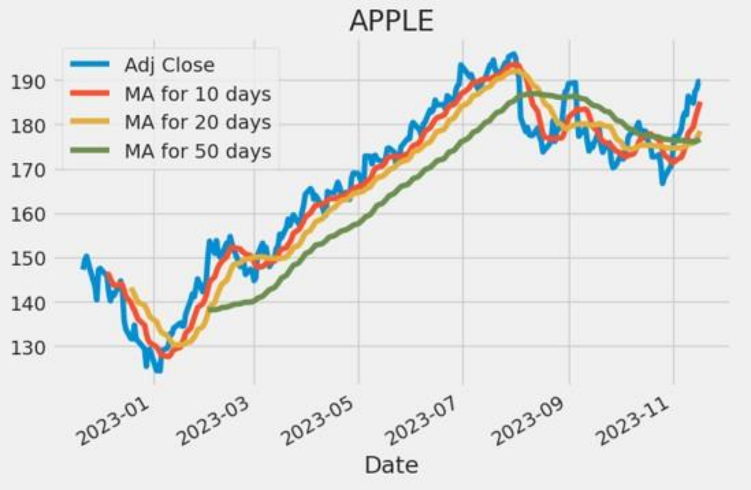


Figure 5: Graph presenting comparison of actual Adj close price with predicted close price using MA of 10 days, 20 days, and 50 days on Apple stock dataset.

Technical Indicators:

For Proposed System2, along with historical data of Apple stock, volume-related indicators as shown in Table III including 50-day Average Volume, 200-day Volume Moving Averages, Volume Changes, Distribution Line, On-Balance Volume, Chaikin Money Flow (CMF), Volume Oscillators are utilized to train deep learning models [33]. Volume oscillators are utilized with price charts and additional technical indicators in the proposed study to improve the study of stock price fluctuations and detection of possible trend reversals or continuations; and enhance the trading decisions and the prediction results. Hybrid technique utilizing a filter-based method, followed by a wrapper-based method and an embedded method is employed to select the most relevant features and eliminating redundancy caused by highly correlated features.

Volume-Related Features:

Table *III*: Computed Volume Indicators, presents a list of volume indicators that are computed to incorporate volume fluctuations as a characteristic in predictive models, to improve the model's capacity to recognize significant market dynamics. Examining fluctuations in volume in conjunction with price fluctuations offers deeper insights into market dynamics and can enhance the quality of trading decisions.

Table III: Computed Volume Indicators

Volume Indicators	Description	Formula
Volume Moving Averages (VMAs)	VMAs are a statistical tool utilized in stock market analysis to	By considering each day's new data and discarding the previous day's data, a moving average is created that

	reduce the impact of changes in trade volume by averaging them over a set time frame.	adjusts promptly to recent fluctuations in volume. VMA$_t$ = $\frac{Vt+Vt-1+\cdots+Vt-n+1}{N}$ where: VMA$_t$ is Volume Moving Average for day t, v$_t$ is volume traded for day t and n is the time period chosen. Here, we have taken 200 periods.
Volume Change	Volume changes indicate the percentage or absolute variance of trade volume within a specific timeframe. In this research, absolute change in volume is taken as one of the features.	Calculate the absolute volume difference between the current day and the previous day or a given baseline day. Absolute Volume Change = $Current\ Volume - Previous\ Volume$
Accumulation/Distribution Line (A/D Line)	The Close Location Value (CLV) quantifies the correlation between the closing price and the daily price range.	CLV= $\frac{Close-Low}{High-Low}$ CLV ranges from 0 to 1, showing the proximity of the closing price to the high (CLV ≈ 1) or low (CLV ≈ 0) of the day.
On-Balance Volume (OBV)	aims to quantify the total purchasing or selling force by adjusting the volume of a security according to its price movement.	OBV$_t$= $OBV_{t-1}+Volume_t*Sign(Close_t\text{-}Close_{t-1})$ where, OBV$_t$ is the On-Balance Volume at time t, Volume t is the volume at time t, and Close$_t$ represents the closing price at time t, Close$_{t-1}$ represents the closing price at time t−1, and Sign is the sign function (positive if Close$_t$ > Close$_{t-1}$, negative if Close$_t$ < Close$_{t-1}$, and zero if they are equal).

Chaikin Money Flow	CMF is a momentum indicator that evaluates the inflow and outflow of capital into and out of a stock by combining price and volume data. It is utilized to validate trends, detect possible reversals, and evaluate the level of purchasing or selling demand within the market.	Apply the following formula to determine the Money Flow Multiplier (MF Multiplier): MF Multiplier = $\frac{Close-Low-(High-Close)}{High-Low} * Volume$ Perform the Money Flow Volume (MFV) calculation: MFV = $MF\ Multiplier * Volume$ Determine the 10-day CMF by: CMF = $\frac{Sum\ of\ 10\ day\ MFV}{Sum\ of\ 10\ day\ Volume}$
Volume oscillators indicators	the volume can offer valuable insights regarding the robustness or feebleness of a given price trend. Traders can utilize volume oscillators to corroborate trends, identify potential trend reversals, and generate buy or sell signals.	Volume Rate Of Change calculates the volumetric percentage change over a given time period of 20 days: VROC= $\frac{Volume_t - Volume_{t-20}}{Volume_{t-20}} * 100$ In this context, "volume_t" denotes the volume for the current day, "volume_{t-20}" represents the volume 20 days ago, and "20" represents the selected time period.

4.2. Feature Selection Phase

The second phase of the Experimentation2 prediction system is the feature selection phase, which helps select the relevant and correlated attributes and reduces the computations of the system, hence improving its performance[34]. Different feature selection methods, such as correlation analysis, machine learning approaches, and optimization algorithms, are used to select attributes so that only relevant attributes can be chosen. This work proposes a hybrid feature optimization approach (C-R-L) combining Filter-based Correlation Analysis, followed by Wrapper-based RFE method using random forest technique and L1 Regularization embedded method. Figure 6 represents the block diagram of Hybrid Feature Optimization approach.

4.2.1. Filter-based Correlation analysis identifies a potential set of features that are relevant and remove less significant features based on correlation with the target variable. Positive or negative features, where correlation coefficients are high, are regarded as more significant and selected for more analysis. This method is computationally efficient and helps in reducing the dimensionality of the dataset without extensive training requirements.

4.2.2. Using the RFE Wrapper-Based approach, to evaluate feature subsets based on their predictive power. This involves using a predictive model to assess the efficacy of different combinations of features. In this work, random forest model is trained for the purpose to refine the feature set and hence, minimising the risk of overfitting and allowing iterative improvement of the feature set to optimize model performance. It also adds dynamic adaptability, which ensures better and optimal performance.

4.2.3. Lastly, to fine-tune the feature selection process, employ L1 regularization (Lasso regression) is integrated during the learning process of the machine learning model, which penalizes irrelevant features and encourages sparsity in the coefficient vector. By essentially decreasing certain coefficients to zero and removing superfluous features, it promotes sparsity in the coefficient vector.

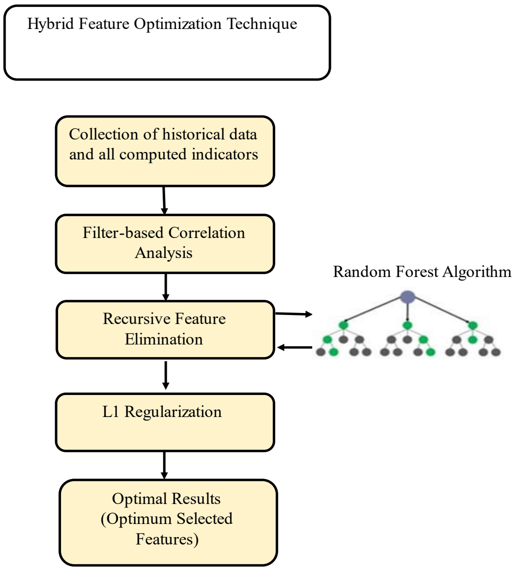

Hybrid Feature Optimization Technique
Collection of historical data and all computed indicators
Filter-based Correlation Analysis
Recursive Feature Elimination
Random Forest Algorithm
L1 Regularization
Optimal Results (Optimum Selected Features)

Figure 6: Hybrid Feature Optimization approach.

The proposed hybrid optimization uses Correlation Analysis, where the correlation coefficients are calculated among features and the target variables (stock closing price). The degree of linear relationship is identified between the variables through the correlation matrix. Hence, a variable subset is built and multicollinear variables are removed. The selected variables are then used for Recursive Feature Elimination technique used to train Random Forest and Mean Square Error (MSE) metric is used for performance evaluation. Lower the MSE, better the performance. The selected feature set keeps on updating based on the lowest MSE. In the final step of feature optimization, the StandardScaler technique is applied for L1 Regularization to retain relevant attributes to improve prediction accuracy. The proposed comprehensive feature selection strategy enhances the predictive accuracy of our models and supports more sophisticated trading strategies.

4.3. Training Preparation Phase

This phase prepares the data before feeding it to the model. First, the data is normalized using the min-max scaling method[35]. Min-max scalar is a commonly used normalization method that adjusts the scale of the data without losing any information. It transforms the feature value between 0 and 1, using that feature's minimum and maximum values. After feature selection phase, the optimal relevant selected features are transformed using a min-max scalar. The data is then converted into sequences of specific length to capture temporal dependencies in the data. The fixed size of 60 days means 02 months is chosen, which is long enough to capture trends, cyclic behavior, and patterns. It helps the model understand the changes day by day and even for longer periods, increasing the accuracy of the predicted results. After this, the data is separated into training and test data with an 80:20 ratio.

4.4. Deep Learning Model Selection Phase

Even while deep neural networks have demonstrated great capability, optimizing a network is a difficult undertaking. The network's feature set, training technique, activation function, input data, width (neurons per layer), depth (the number of hidden layers in the neural network), and training algorithm all have a substantial impact on its performance [36] [37]. The selected and scaled stock features are used as input to train the mentioned below deep learning models.

4.4.1. Recurrent Neural Networks (RNN)

The neural networks in which the connections between the units occur again is known as recurrent neural networks. This enables them to process the incoming sequence using their internal memory with the help of feedback loop [38], [39]. This indicates the presence of memory in the recurrent neural network. There is a lot of information in each input sequence, and recurrent networks store this information in a hidden state. The network uses this concealed knowledge recursively as it moves ahead to handle a fresh sample. In this study, recurrent neural networks are used since the stock data requires consideration of long-term relationships in the data. Furthermore, RNN's recursive formulae are shown in the following equations [12] and Figure 7 represents the RNN in pictorial form.

$h_t = \tanh (W_t h_{t-1} + W_x x_t)$,

$y_t = W_y h_t$,

where y_t, h_t, x_t, and W_x indicates output vector, hidden layer vector, input vector, and weighting matrix, respectively.

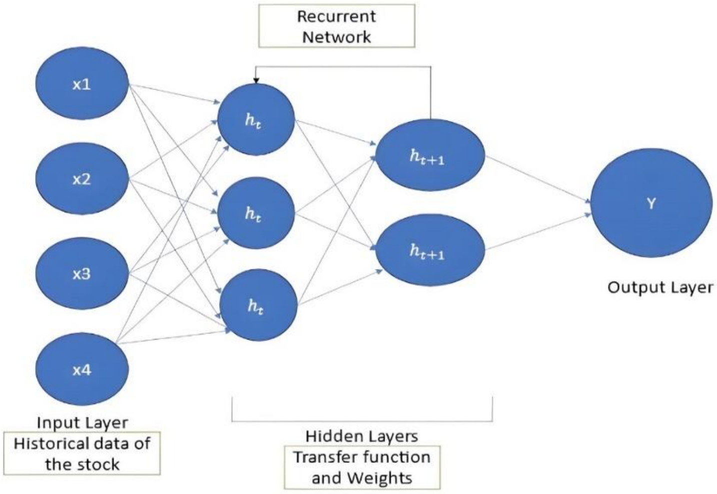


Figure 7: Recurrent Neural Network

4.4.2. Long-Short Term Memory (LSTM)

LSTM is an enhanced type of RNN, that learns long-term dependencies using sigmoid activation functions. LSTM retains the data from the prior state, whereas RNN operates on establishing the connection between current and recent information.

As stock price prediction depends on large amount of sequential historical data, LSTM model is expected to provide better results, [40] by pertaining information of various older stages and controlling errors through it. Long-Term Memories are devoid of biases and weights; thus, they can traverse a sequence of unrolled units without inducing a gradient explosion or disappearance [41]. Remembering cells (which retain the value for long-term propagation), an input gate, a forget gate (which regulates the value), and an output gate are the modules of LSTM model. The percentage of the previous stage's long-term memory that is retained is determined by the forget gate and the input value and the sigmoid and tanh activation functions is used by the input gate on Short-Term Memory to establish how to update the Long-Term Memory. The concluding output of the complete LSTM unit is the updated Short-Term Memory, which serves as the output gate [42]. Figure 8 is the pictorial representation of LSTM network.

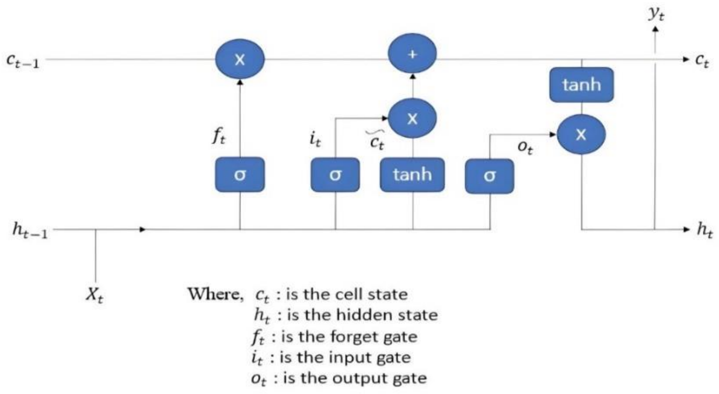

c_{t-1}
y_t
c_t
x
+
tanh
x
x
f_t
i_t
$\tilde{c}_t$
o_t
σ
σ
tanh
σ
h_{t-1}
h_t
X_t
Where, c_t : is the cell state
h_t : is the hidden state
f_t : is the forget gate
i_t : is the input gate
o_t : is the output gate

Figure 8: Long Short-Term Memory

4.4.3. CNN-LSTM

Convolutional Neural Network (CNN) is a type of artificial neural network that works good for identifying and recognising patterns used for image analysis [24]. A CNN mainly consist of two parts: convolutional layer (hidden layers for filtering) and polling layer (to reduce the extracted feature dimensions and cost of training). CNN can perform prediction of time series data effectively through weight sharing and local perception of CNN [38]. As per the properties of LSTM and CNN, a CNN-LSTM model for forecasting stock price is built. Figure 9 represents the block diagram of CNN-LSTM.

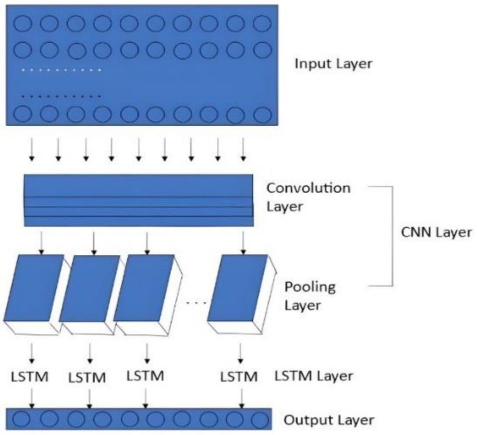


Figure 9: CNN-LSTM neural network structural diagram.

In the proposed study, the deep learning models used are LSTM, RNN and CNN-LSTM for predicting the day-wise closing price of the selected stock[43]. The models are trained over 80% of the data (from 2012 to 2022 end) and the testing is done on the rest 20% of the data from the dataset. For RNN and LSTM, the value of parameters employed are: two layers with 100 neurons per layer, with learning rate optimiser ADAM, ReLU as activation function. For CNN-LSTM hybrid model: 1D convolutional layer with 64 filters with kernel size of 2x2 and LSTM layer of 100 units is used after convolutional layer. Each time the number of epochs is 50 and 32 is the batch size applied for all the proposed models. The proposed study utilizes Root Mean Square Error (RMSE) as the evaluation metric [44].

4.5. Prediction Results and Analysis Phase

The experiment was conducted on a Windows system with an i7 processor and integrated GPU capabilities. Python programming language and the Jupyter Notebook platform are used to implement this proposed system.

The study comprises of implementing two single deep learning models i.e. Long Short-Term Memory Neural Network and Recurrent Neural Network and one combined model CNN-LSTM each using 50 number of epochs and batch-size of 32, trained on the highly traded stock of Apple company. The evaluation parameter used is RMSE. The RMSE represents the root average squares residual and the result of RMSE is always non-negative (model whose results are near to zero shows better prediction quality) [21], [31]. The root of the average squared deviation between the predicted and observed values is computed. The RMSE formula is as follows:

$$\sqrt{1/\mathrm{N}\ \mathrm{n}\textstyle\sum \mathrm{i} = 1\ (\mathrm{Ai} - \mathrm{Pi})2}$$

Figure 10 represents the line chart showing the trained data, actual price (Val), and the predicted price (Predictions) after applying neural network model LSTM. **Error! Reference source not found.** represents the l ine chart showing the trained data, actual stock price, and the predicted price after applying RNN model. Figure 12 represents the line chart showing the trained data, actual price, and the predicted stock price after applying the combinational CNN-LSTM neural network model. Table V conveys the RMSE value of various models applied on the dataset. It is observed that CNN-LSTM combined model with the RMSE score of approximately 0.63, performed better as compared to RNN (2.26) and even LSTM (1.05) alone.

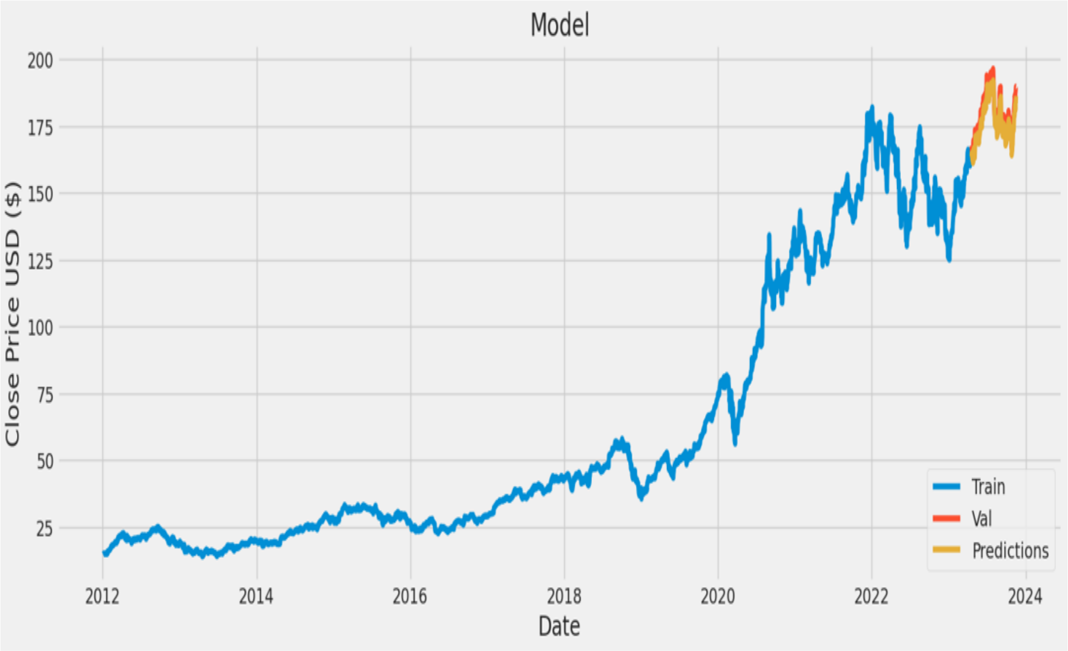


Figure 10: Line chart depicting Trained data, Actual value (Val) and Predicted data (Predictions) using LSTM.

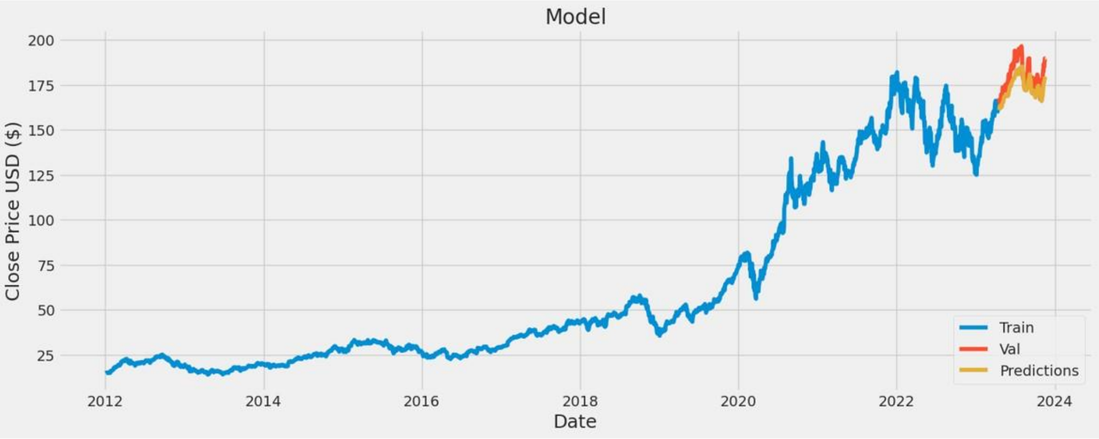


Figure 11: Line chart depicting Trained data, Actual value (Val) and Predicted data (Predictions) using RNN.

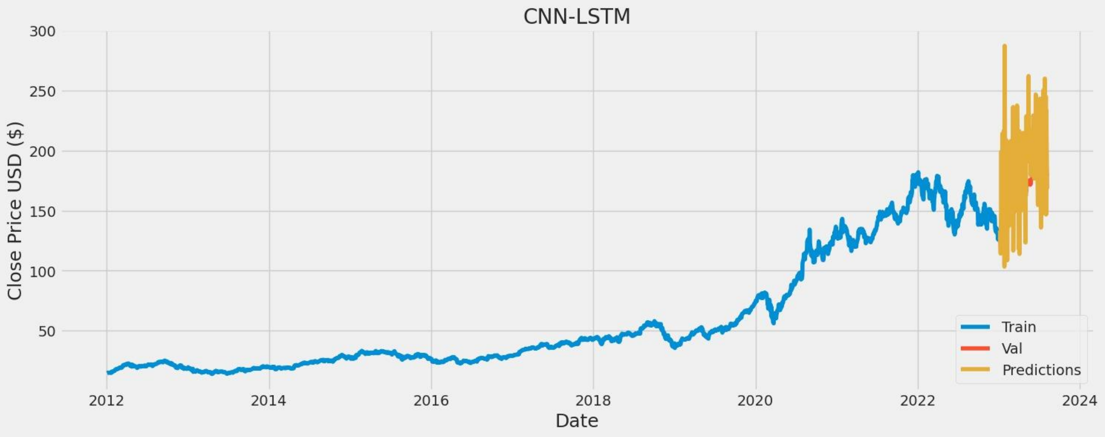


Figure 12: Line chart depicting Trained data, Actual value (Val) and Predicted data (Predictions) using CNN-LSTM.

Table IV: Evaluation parameter RMSE value of various deep models applied on the AAPL dataset.

Prediction Models	Parameters	RMSE
LSTM	Epochs=50, batch-size=32	1.05099
RNN	Epochs=50, batch-size=32	2.26875
CNN-LSTM	Epochs=50, batch-size=32	0.63475

5. Comparative Analysis of the two Proposed Systems

This section discusses the comparative analysis between the two proposed systems for predicting AAPL stock prices using machine learning and deep learning techniques.

The Proposed System1 uses a combination of feature selection techniques (SelectKBest and Random Forest) and applies five different ensemble machine learning models. Gradient Boosting and XGBoost shows significant results, capturing the complex patterns. The voting ensemble, which aggregates predictions from multiple models, shows relatively better performance, demonstrating the benefit of leveraging diverse models.

The Proposed System2 incorporates a hybrid feature optimization technique (C-R-L) and applies it to three advanced deep learning models. CNN-LSTM outperforms suggesting the combinational architecture, which integrates convolutional neural layers to capture spatial dependencies and LSTM layers to capture long-term temporal dependencies, suited well for the times-series forecasting. The comparison of both the proposed systems is shown in Figure 13 and Figure 14.

The sophisticated feature selection method (C-R-L) appears to provide a more robust basis for training for training deep learning models compared to the relatively simpler method used in Proposed System1.

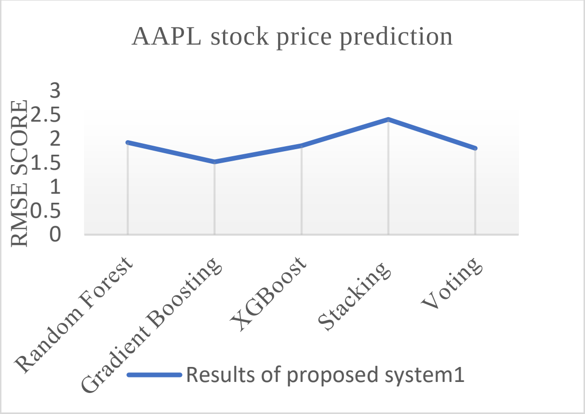


Figure 13: Line chart depicting the results of proposed system1

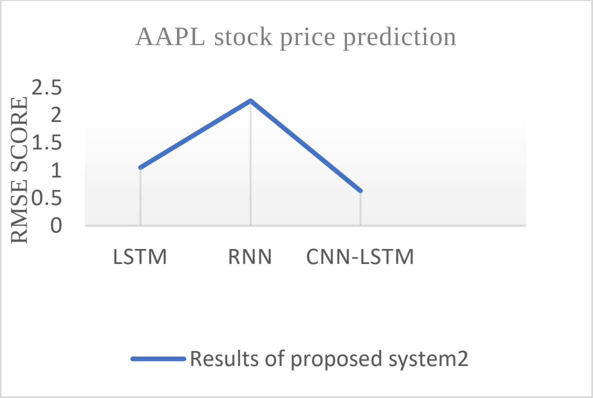


Figure 14: Line chart depicting the results of proposed system2

6. Comparative Analysis

The research paper [25] is considered for comparative analysis as it has done the prediction on the same company's dataset. The paper took ten year of Apple Inc. stock data from 2008 to 2018 and four machine learning models- k- Nearest Neighbors model, Moving Average model, Linear Regression model, Prophet model, and Long Short-Term model. The evaluation parameter RMSE was used for comparing the model's performance. The LSTM model gave better results with RMSE of 2.39 without applying any technical indicator, whereas LSTM model implemented in the proposed study with technical indicators outperformed with RMSE of 1.05 approximately. Also, it can be stated that proposed hybrid model CNN-LSTM along with hybrid feature optimization technique applied on technical indicators like- MA for 10 days, 20 days and 50 days, MACD and various volume related indicators; gives the better result with RMSE of 0.633 approximately. Furthermore, the system setup for the deep learning models implemented in the proposed study involved using 50 epochs and a batch size of 32 to achieve significant results. Table *V* is the tabular representation showing the comparative analysis between the base paper and the proposed study by comparing the obtained evaluation parameters of the best performed model respectively.

Table V: Comparative analysis of the proposed study

	Parameters					
	Dataset	**Technical Indicators**	**Feature optimization or selection technique**	**Models**	**Best performed model**	**Evaluation Parameter**
[25]	10 years of AAPL stock data (2008 to 2018)	_	_	Linear Regression model, Prophet model, and Long Short-Term Memory (LSTM)	LSTM	RMSE= 2.39
Proposed Model	11 years of AAPL stock data (2012 to 2023)	10 days, 20 days, and 50 days Moving Average, MACD and various other volume related indicators	Hybrid Feature Optimization approach combining Correlation Analysis, Recursive Feature Elimination (RFE) and L1 Regularization	RNN, LSTM and CNN-LSTM hybrid	CNN-LSTM	RMSE= 0.633

7. Conclusion and Discussion

The stock prices that are constantly fluctuating, depending on several factors and hence, generate complex patterns, due to which stock market trends prediction or market returns become a tough task. Firstly, when data is collected from Yahoo finances, the number of attributes is much smaller, so different technical and volume related indicators are computed and fused to increase the number of attributes in the data. Secondly, this work proposed two systems- the first system utilizes basic technical indicators integrated with the retrieved dataset and used a fusion based feature selection technique with hyper-parameter tuning technique. The results are used to train five ensemble machine learning models. The second system employs an optimized hybrid feature selection approach (C-R-L) combining Correlation Analysis, Recursive Feature Elimination (RFE) and L1 Regularization with random forest regression. In this study, Apple stock eleven-year dataset was extracted, and predictions were done using machine learning and deep learning models such as LSTM, RNN and CNN-LSTM fusion model. The CNN-LSTM fusion model outperforms the rest of the learning models applied with the RMSE of 0.633. For further work, deep learning models might be created by taking financial news items as well as financial metrics like a price per earnings ratio, ROCE, debt by equity ratio, traded volume, profit and loss statements, etc. into account for potentially improved outcomes.

Reinforcement learning can be explored for dynamic adaptation to market conditions. Additionally, experimenting with real-time data streaming and online learning could make the model more robust and responsive to market conditions.

Author Contribution

All authors have contributed to the study conception and design. Conceptualization, Implementation and Writing original draft was done by Shobhita Singh. Data collection, Implementation was done by Divya Khanna and Writing-review and Editing was done by B.S. Bhatia. All authors have read and approved the final manuscript.

Ethical Statement

The work does not include any personal data relatable to identifiable living persons.

Conflict of Interest

The authors affirm that they do not possess any identifiable personal relationships or any competing financial interests that might have appeared to exert an influence on the research presented in this article. The authors affirm that there is no disagreement concerning the acknowledgment of their authorship. Therefore, the authorship acknowledgement is consistent with the information provided in the manuscript and the details.

Data Availability Statement

The dataset used for training is publicly available dataset and the data including some test samples, codes, etc. can be made available on request.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Research Involving Human and /or Animal

This research does not involve any kind of animal and human participation.

Informed Consent

Not Applicable.

References

[1] D. and D. R. A. Burnie, “Do Stock Prices Conform to an Absolute Price Level?,” *47th Annual Meeting of European Finance Association (EFA), April 13-16, 2011, Savannah, Georgia, USA*, 2011.

[2] D. A. Burnie and A. De Ridder, “The Nominal Stock Price Puzzle-A Long View-Sweden-03-2010.” [Online]. Available: https://www.researchgate.net/publication/280625271

[3] B. T. (2022). Khoa, “ Forecasting stock price movement direction by machine learning algorithm.,” *International Journal of Electrical and Computer Engineering.*, 2022.

[4] M. and A. I. AlKandari, “Solar power generation forecasting using ensemble approach based on deep learning and statistical methods,” *Applied Computing and Informatics*.

[5] S. a. Mehtab, . “Stock price prediction using machine learning and LSTM-based deep learning models.,” *Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14--17, 2020, Revised Selected Papers 2, 88--106.*, 2021.

[6] E. and G. D. and B. P. Pechlivanidis, “Can intangible assets predict future performance? A deep learning approach,” *International Journal of Accounting \& Information Management*.

[7] M. M. (2022) Akhtar, “Stock market prediction based on statistical data using machine learning algorithms.,” *Journal of King Saud University-Science*, 2022.

[8] Y. and C. Z. Zhao, “Forecasting stock price movement: New evidence from a novel hybrid deep learning model,” *Journal of Asian Business and Economic Studies*.

[9] Y. and A. P. H. M. and K. H. and S. C. A. P. B. Peng, “Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators,” *Machine Learning with Application*.

[10] S. a. (2023) Mukherjee, . “Stock market prediction using deep learning algorithms.,” *CAAI Transactions on Intelligence Technology, 82--94.*, 2023.

[11] S. and C. N. N. Maheshwari, “Applications of artificial intelligence and Machine learning-based supervisory technology in financial Markets surveillance: A review of literature,” *FIIB Business Review*.

[12] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. Shahab, “Deep learning for stock market prediction,” *Entropy*, vol. 22, no. 8, Aug. 2020, doi: 10.3390/E22080840.

[13] J. and C. Q. and D. Y. Wang, “An XGBoost-based multivariate deep learning framework for stock index futures price forecasting,” *Kybernetes*.

[14] S. Mukherjee, B. Sadhukhan, N. Sarkar, D. Roy, and S. De, “Stock market prediction using deep learning algorithms,” *CAAI Trans Intell Technol*, vol. 8, no. 1, pp. 82–94, Mar. 2023, doi: 10.1049/cit2.12059.

[15] N. K. and G. V. and A. A. and A. H. M. and V. S. G. and A. D. and G. N. and K. S. Trivedi, “Early detection and classification of tomato leaf disease using high-performance deep neural network,” *sensors*.

[16] M. Nikou, G. Mansourfar, and J. Bagherzadeh, “Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms,” *Intelligent Systems in Accounting, Finance and Management*, vol. 26, no. 4, pp. 164–174, Oct. 2019, doi: 10.1002/isaf.1459.

[17] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, “NSE Stock Market Prediction Using Deep-Learning Models,” in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 1351–1362. doi: 10.1016/j.procs.2018.05.050.

[18] N. Dahiya, S. Gupta, and S. Singh, “Qualitative and quantitative analysis of artificial neural network-based post-classification comparison to detect the earth surface variations using hyperspectral and multispectral datasets,” *J Appl Remote Sens*, vol. 17, no. 03, Jun. 2023, doi: 10.1117/1.jrs.17.032403.

[19] H. Rezaei, H. Faaljou, and G. Mansourfar, “Stock price prediction using deep learning and frequency decomposition,” *Expert Syst Appl*, vol. 169, May 2021, doi: 10.1016/j.eswa.2020.114332.

[20] L. U. K. , P. M. , S. S. , K. A. , H. M. Ramesh T.R., “PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES,” *Malaysian Journal of Computer Science*.

[21] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Syst Appl*, vol. 83, pp. 187–205, Oct. 2017, doi: 10.1016/j.eswa.2017.04.030.

[22] M. A. and S. J. Hamid, "EEG Signal Processing for Detection of Colour Vision Deficiencies," *ECS Trans*, vol. 107, 2022.

[23] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00333-6.

[24] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/6622927.

[25] X. Zhao, "The Prediction of Apple Inc. Stock Price with Machine Learning Models," *2021 3rd International Conference on Applied Machine Learning (ICAML)*.

[26] K. N. and M. P. and J. J. J. Gopinathan, "Stock price prediction using a novel approach in Gaussian mixture model-hidden Markov model," *International Journal of Intelligent Computing and Cybernetics*.

[27] J.-F. and C. W.-L. and H. C.-P. and H. S.-H. and C. A.-P. Chen, "Financial time-series data analysis using deep convolutional neural networks," *2016 7th International conference on cloud computing and big data (CCBD*.

[28] B. and B. S. and A. V. and Y. M. and K. P. Dhingra, "Stock market volatility: a systematic review," *Journal of Modelling in Management*.

[29] A. and P. M. M. and P. R. M. Nayak, "Prediction models for Indian stock market}," *Procedia Comput Sci*.

[30] S. Singh and D. Divya Khanna, "A survey on fundamental and technical analysis used in stock market prediction," 2022, [Online]. Available: www.ijsdr.org

[31] K. Amzile, "Artificial Intelligence and Stock Market}," *Advances in Emerging Financial Technology and Digital Money*.

[32] O. and B. E. and E. E. Karahasan, "New deep recurrent hybrid artificial neural network for forecasting seasonal time series," *Granular Computing*.

[33] M. R. and D. A. C. E. and B. G. L. and E. A. G. Vargas, "Deep leaming for stock market prediction using technical indicators and financial news articles," *2018 international joint conference on neural networks (IJCNN)*.

[34] F. and B. M. A. and T. D. Konak, "Feature Selection and Hyperparameters Optimization Employing a Hybrid Model Based on Genetic Algorithm and Artificial Neural Network: Forecasting Dividend Payout Ratio," *Comput Econ*.

[35] Z. and P. L. Tian, "Stocks price prediction based on optimized echo state network by sparrow search algorithm," *Int J Dyn Control*.

[36] H. and S. N. P. Goel, "Dynamic prediction of Indian stock market: An artificial neural network approach," *International Journal of Ethics and Systems*.

[37] K. and D. O. and G. P. and A. V. Khare, "Short Term Stock Price Prediction Using Deep Learning ," *2017 2nd IEEE international conference on recent trends in electronics, information \& communication technology (RTEICT)*.

[38] K. Pawar, R. S. Jalem, and V. Tiwari, “Stock Market Price Prediction Using LSTM RNN,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2019, pp. 493–503. doi: 10.1007/978-981-13-2285-3_58.

[39] D. P. Kukreja V., “A Deep Neural Network based disease detection scheme for Citrus fruits,” *Proceedings - International Conference on Smart Electronics and Communication, ICOSEC 2020*.

[40] K. and S. S. M. Sujatha, “Stock index prediction using regression and neural network models under non normal conditions,” *IEEE*.

[41] A. Q. Md, “ Novel optimization approach for stock price forecasting using multi-layered sequential LSTM.,” *Applied Soft Computing.*, 2023.

[42] P. S. (2022). Sisodia, “ Stock market analysis and prediction for NIFTY50 using LSTM Deep Learning Approach. ,” *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), 156--161.*, 2022.

[43] M. T. and A. I. and A. M. T. and A. K. B. and A. H. M. and A. A. Q. M. Nuseir, “Stock Market Price Prediction Using Machine Learning Techniques,” *Cyber Security Impact on Digitalization and Business Intelligence: Big Cyber Security for Information Management: Opportunities and Challenges*.

[44] S. K. Chandar, “Convolutional neural network for stock trading using technical indicators,” *Automated Software Engineering*.