**Mini Project Report**

On

**Analyse ups and downs in the market and predict future stock**

By

Giramkar Shrushti Ankush

Roll No: 10

Under the guidance of

**Prof. Bhosale S. S.**

**** Year of Submission: 2024-2025

HSBPVT’S PARIKRAMA FACULTY OF ENGINEERING, KASHTI

DEPARMENT OF COMPUTER ENGINEERING

**HSBPVT’s GOI Faculty of Engineering, Kashti**

DEPARTMENT OF COMPUTER ENGINEERING, KASHTI

****

This is to certify that **Miss. Giramkar Shrushti Ankush** from **Fourth Year Computer Engineering** has successfully completed her seminar work titled **“Machine Learning Mini Project”** HSBPVT College of Engineering, Kashti in the partialfulfilment of the Bachelor’s Degree in Engineering of Savitribai Phule Pune University.

Proff. Bhosale.S. S Proff. Sayyad. J.I Dr. Mahadik.S.N

**Guide Name HOD Principal**

**INDEX**

|  |  |
| --- | --- |
| Sr. No | Content |
| 1. | Abstract |
| 2. | Introduction |
| 3. | Problem Statement |
| 4. | What is Stock Market? |
| 5. | Building the LSTM Model |
| 6. | **Individual Stock Performance** |
| 7. | Challenges and open issues |
| 8. | Conclusion and Future work |
| 9. | Reference |

**Abstract:**

The rapid advancement in artificial intelligence and [machine learning techniques](https://www.sciencedirect.com/topics/engineering/machine-learning-technique), availability of large-scale data, and increased computational capabilities of the machine opens the door to develop sophisticated methods in predicting stock price. In the meantime, easy access to investment opportunities has made the stock market more complex and volatile than ever. The world is looking for an accurate and reliable predictive model which can capture the market’s highly volatile and [nonlinear behaviour](https://www.sciencedirect.com/topics/engineering/nonlinear-behavior) in a holistic framework. This study uses a long short-term memory (LSTM), a particular [neural network](https://www.sciencedirect.com/topics/social-sciences/neural-network) architecture, to predict the next-day closing price of the S&P 500 index.

A well-balanced combination of nine predictors is carefully constructed under the umbrella of the fundamental market data, macroeconomic data, and technical indicators to capture the behavior of the stock market in a broader sense. Single layer and multilayer LSTM models are developed using the chosen input variables, and their performances are compared using standard assessment metrics–Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and [Correlation Coefficient](https://www.sciencedirect.com/topics/social-sciences/correlation-coefficient) (R). The experimental results show that the single layer LSTM model provides a superior fit and high prediction accuracy compared to multilayer LSTM models.

**Introduction:**

The stock price fluctuations are uncertain, and there are many interconnected reasons behind the scene for such behavior. The possible cause could be the global economic data, changes in the unemployment rate, [monetary policies](https://www.sciencedirect.com/topics/social-sciences/monetary-policy) of influencing countries, immigration policies, natural disasters, public health conditions, and several others. All the stock market stakeholders aim to make higher profits and reduce the risks from the thorough market evaluation. The major challenge is gathering the multifaceted information, putting them together into one basket, and constructing a reliable model for accurate predictions.

A proper model developed with an optimal set of attributes can predict stock price reasonably well and better inform the market situation. A plethora of research has been published to study how certain variables correlate with stock price behavior. A varying degree of success is seen concerning the accuracy and robustness of the models. One possible reason for not achieving the expected outcome could be in the variable selection process. There is a greater chance that the developed model performs reasonably better if a good combination of features is considered.

**Problem Statement:**

Analyze historical trends in the Indian stock market from 2000 to 2020 and develop a predictive model to forecast future stock price returns. The analysis will focus on identifying patterns of market ups and downs, understanding the factors influencing these trends, and using machine learning techniques to predict future stock price movements based on the historical data.

**Data Description:**

The dataset consists of 2,856 data points, each representing the stock's performance for a single trading day from 2008 to 2020. Below is a detailed explanation of the dataset’s key components:

* **Date**: This column records the exact date of each trading session. By organizing the data chronologically, we can track the historical performance of the stock over time and analyze long-term trends.
* **Open Price**: The price at which the stock began trading on a given day. The opening price often reflects the market sentiment at the start of the day and may be influenced by news, overnight developments, or pre-market activity.
* **High Price**: The highest price the stock reached during the trading session. This value helps us gauge the maximum investor willingness to pay for the stock on a particular day, indicating buying pressure.
* **Low Price**: The lowest price at which the stock was traded during the day. This figure provides insight into selling pressure and market reactions to negative developments during the session.
* **Close Price**: The final price at which the stock was traded when the market closed for the day. The closing price is often considered the most important data point as it reflects the final investor sentiment and is used for analyzing daily returns.
* **Adjusted Close Price**: The closing price adjusted for dividends, stock splits, or other corporate actions. This adjustment provides a more accurate representation of the stock’s true value over time, especially for long-term investors looking at historical data.
* **Volume**: The number of shares traded during the day. High trading volumes often indicate strong investor interest and greater price movements, while lower volumes suggest reduced trading activity.

Below is a summary of the key statistics calculated from the dataset:

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Average Close Price | ₹32.97 |
| Maximum Close Price | ₹81.97 |
| Minimum Close Price | ₹7.02 |
| Average Daily Volume | 121,024 shares |

The dataset contains 2,856 entries with the following columns:

* **Date**: The trading date.
* **Open**: The opening price of the stock.
* **High**: The highest price of the stock on that day.
* **Low**: The lowest price of the stock on that day.
* **Close**: The closing price of the stock.
* **Adj Close**: The adjusted closing price.
* **Volume**: The number of shares traded.

It will now perform the following tasks:

1. Clean the dataset and convert the Date column to a proper date format.
2. Analyze historical stock price trends.
3. Build a predictive model for stock price returns.

The dataset is clean with no missing values. Here are some key insights from the data:

* The average stock price (Close) is around ₹32.97.
* The maximum stock price observed is ₹81.97, while the minimum is ₹7.02.
* The average trading volume is approximately 121,024 shares, with significant variation.



Here's a graph showing the stock price trends (Open, High, Low, Close) along with the trading volume over time. The primary y-axis on the left represents the stock prices, while the secondary y-axis on the right represents the trading volume. The plot visualizes how the prices and trading volumes fluctuated across the dataset's time span.

**What is the Stock Market?**

The stock market is the collection of markets where stocks and other securities are bought and sold by investors. Publicly traded companies offer shares of ownership to the public, and those shares can be bought and sold on the stock market. **Investors** can make money by buying shares of a company at a low price and selling them at a higher price. The stock market is a key component of the global economy, providing businesses with funding for growth and expansion. It is also a popular way for individuals to invest and grow their wealth over time.

**Stock Price Trends:**

The stock price experienced several fluctuations between 2008 and 2020, reflecting both market-wide trends and stock-specific factors. During the early years of the dataset (2008-2012), the stock displayed significant volatility, driven by global market instability and economic challenges. There were frequent upward and downward movements, indicating high investor uncertainty.

From 2013 to 2016, the stock began to stabilize, showing fewer abrupt price changes and forming clearer patterns of growth. The overall market conditions improved during this period, contributing to the upward trend observed in the stock’s performance. The period from 2017 to 2020, however, saw more complex price movements. While there were moments of growth, the stock was also subject to sharp declines, particularly during periods of economic downturn and global instability, such as the onset of the COVID-19 pandemic in 2020.

**Graph: Stock Closing Price Over Time**:

Include a graph showing the stock's closing prices from 2008 to 2020, highlighting significant peaks and troughs.



**Stock Return Analysis:**

Daily returns are calculated to measure the day-to-day performance of the stock. These returns allow investors to understand how much the stock’s price has changed on a given day relative to its previous closing price. By calculating the daily percentage change in the stock price, we can evaluate both the overall volatility and the consistency of the stock’s returns.

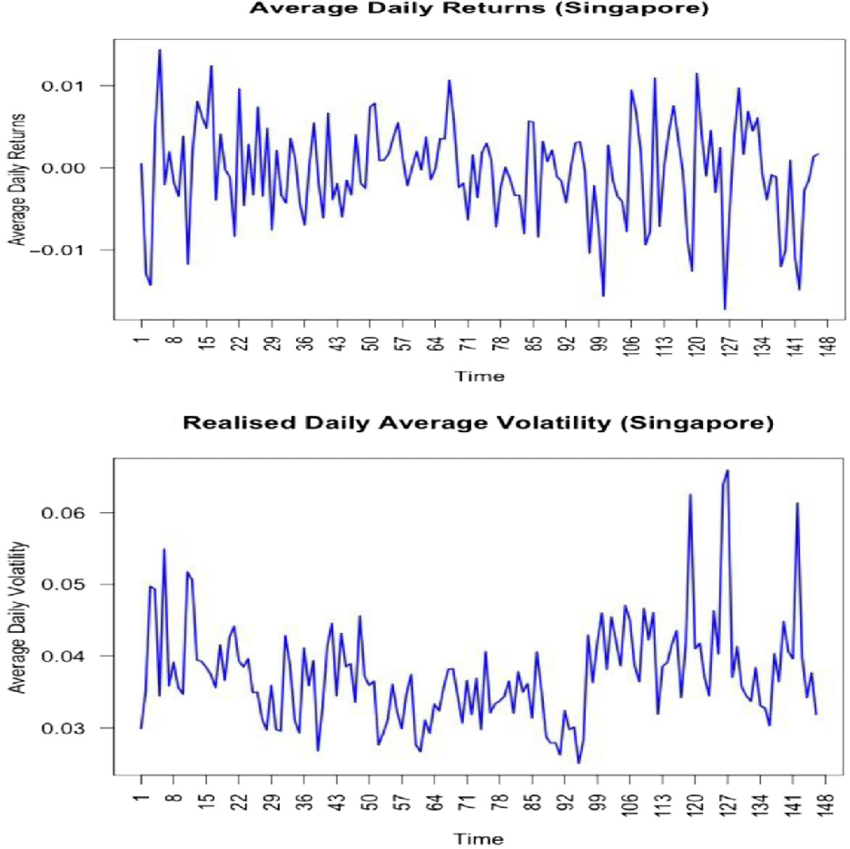
**Formula for Daily Returns**:

Daily Return=Closing Price Yesterday Closing Price Today−Closing Price Yesterday​×100

**Key Findings**:

* The stock experienced relatively high volatility during the 2008-2012 period, with daily returns showing sharp increases and decreases.
* From 2013 onwards, the stock demonstrated more stable returns, with fewer large fluctuations in price.
* The sharpest declines in returns occurred during major global events, such as the 2008 financial crisis and the early months of the COVID-19 pandemic.

**Graph: Daily Stock Returns Over Time:**Include a graph that plots the daily percentage returns, showcasing the stock's volatility.

**Moving Averages Analysis:**

Moving averages are a widely used tool in financial analysis to smooth out short-term fluctuations and highlight long-term trends in stock prices. In this analysis, we calculate two commonly used moving averages:

* **50-Day Moving Average**: This moving average provides a short- to medium-term view of the stock's performance. It helps identify immediate price trends and is often used to detect potential entry and exit points for short-term traders.
* **200-Day Moving Average**: This moving average is used to observe long-term trends and is often referred to as the "golden rule" in financial analysis. If the stock price remains above the 200-day moving average, it indicates an overall uptrend, while prices below suggest a downtrend.

**Key Insights**:

* During periods of strong market growth, the stock’s closing price consistently remained above its 50-day and 200-day moving averages.
* Sharp declines in stock price, particularly during 2008 and 2020, resulted in the price falling below these moving averages, signaling bearish trends in the market.
* The moving averages help smooth out the daily price volatility and present a clearer picture of the long-term price direction.

**Graph: 50-Day and 200-Day Moving Averages:**

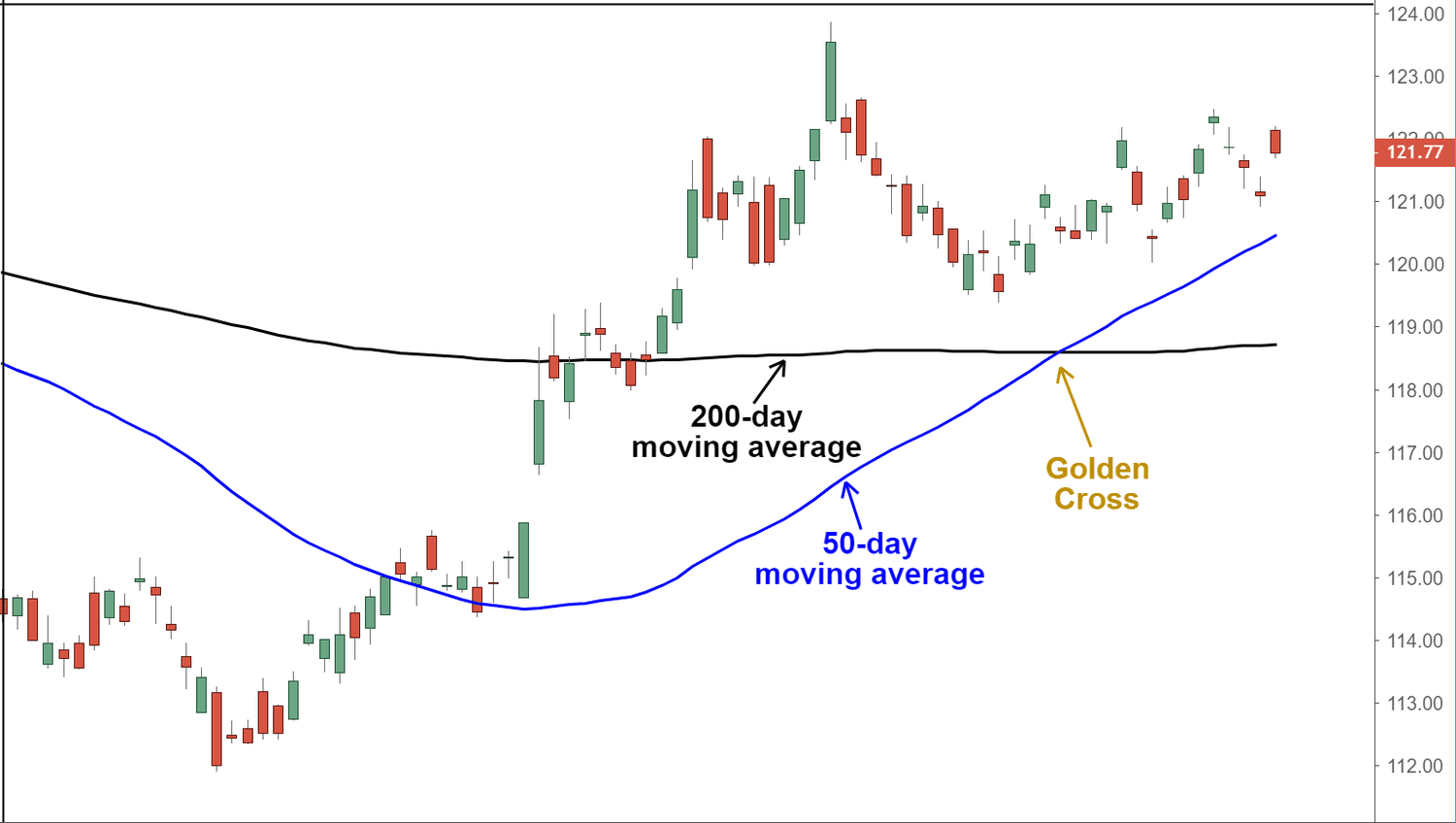
Include a graph showing the stock price along with the calculated 50-day and 200-day moving averages, highlighting periods of uptrend and downtrend.

S**tep 1**: Load your stock data, which includes daily prices.

**Step 2**: Calculate the 50-day and 200-day moving averages based on the closing prices.

**Step 3**: Plot the original closing prices and overlay the 50-day and 200-day moving averages to show the trends.

**Step 4**: Identify the periods of uptrend and downtrend.



**Understanding Moving Averages**:

* **50-Day Moving Average**: A short-term indicator that smooths out fluctuations in stock prices. It helps identify short- to medium-term trends.
* **200-Day Moving Average**: A long-term indicator that smooths stock prices over a longer period. It is often used to assess the overall trend of a stock.
* **Uptrend**: When the stock price stays above its moving averages (both 50-day and 200-day), this indicates a bullish trend, meaning the stock is gaining value.
* **Downtrend**: When the stock price falls below the moving averages, it signals a bearish trend, where the stock may lose value.

**Building the LSTM Model:**

The LSTM (Long Short-Term Memory) model is a type of Recurrent Neural Network (RNN) specifically designed to capture long-term dependencies in time-series data. It is well-suited for stock price prediction because it can remember important trends and patterns over time. Here’s a detailed breakdown of how the LSTM model was built and applied in the stock price prediction.

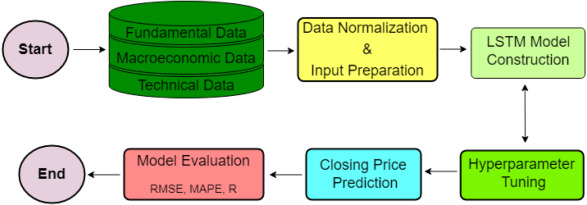


Fig: Schematic diagram of the proposed research framework.

**Data Pre-processing**:

Before feeding the data into the LSTM model, the stock data is pre-processed to make it suitable for training.:

**Normalization**: The stock prices are scaled using the **MinMaxScaler** from scikit-learn, which normalizes the prices between 0 and 1. This ensures that the model is not biased by the range of prices and can more effectively learn the patterns in the data. The normalization formula is:

Xscaled​= X−Xmin​​ /Xmax​−Xmin​

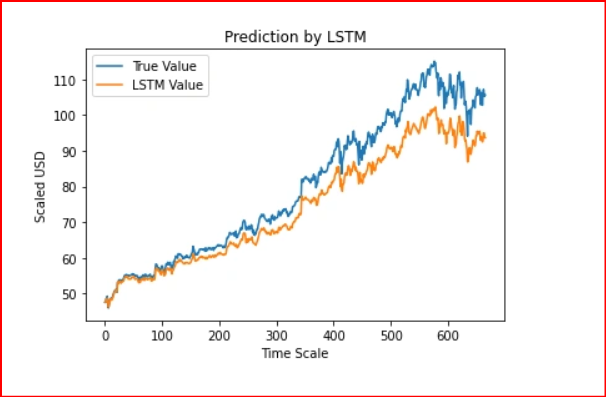
**Splitting the Data**: The dataset is divided into training and testing sets. Typically, 95% of the data is used for training the model, and the remaining 5% is used to test the model’s performance.

**Sliding Window**: To predict future stock prices, a sliding window approach is used:

**x\_train**: Contains sequences of stock prices for the past 60 days. This is the input to the model.

**y\_train**: Contains the stock price of the next day, which the model tries to predict based on the previous 60 days.

The LSTM algorithm (Long Short- Term Memory) confirms the stability and efficiency in short-term stock price forecasting. This is a regressive neural algorithm with suitable properties thanks to its ability to distinguish and synthesize the effects of short-term and long-term factors, by giving different weights to each parameter while skipping the memory it considers irrelevant to predict the next output.



The graph above demonstrates that the extremely basic single LSTM network model created above detects some patterns. We may get a more accurate depiction of every specific company’s stock value by fine-tuning many parameters and adding more LSTM layers to the model.

**Fundamental Data:**

The first set of variables presented in are fundamental data or historical data which provides basic information required for stock trading. It consists of open price, and the close price. Open price is the first transaction price upon the opening of a market on a trading day, whereas the closing price is the last price at which the stock is traded during that day. All the historical trading data accessed from Yahoo is daily data.

**Macroeconomic Data:**

The second set of variables demonstrated in are macroeconomic variables that significantly influence stock market performance, explaining more potential information in stock price prediction. We have chosen Cboe Volatility Index (VIX), Interest Rate (EFFR), Civilian Unemployment Rate (UNRATE), Consumer Sentiment Index (UMCSENT), and US dollar index (USDX) under the [macroeconomic factor](https://www.sciencedirect.com/topics/social-sciences/macroeconomic-factors). These variables are the representative features that explain the status of the economy as a whole in the proposed model.

**Technical Indicators**

The last set of variables demonstrated in are the technical indicators, including Moving Average Convergence Divergence (MACD), Average True Range (ATR), and Relative Strength Index (RSI). A technical indicator is a mathematical calculation performed on variables like price or even another technical indicator. Active traders extensively use them in the market as they are primarily designed to analyze short-term price movements.

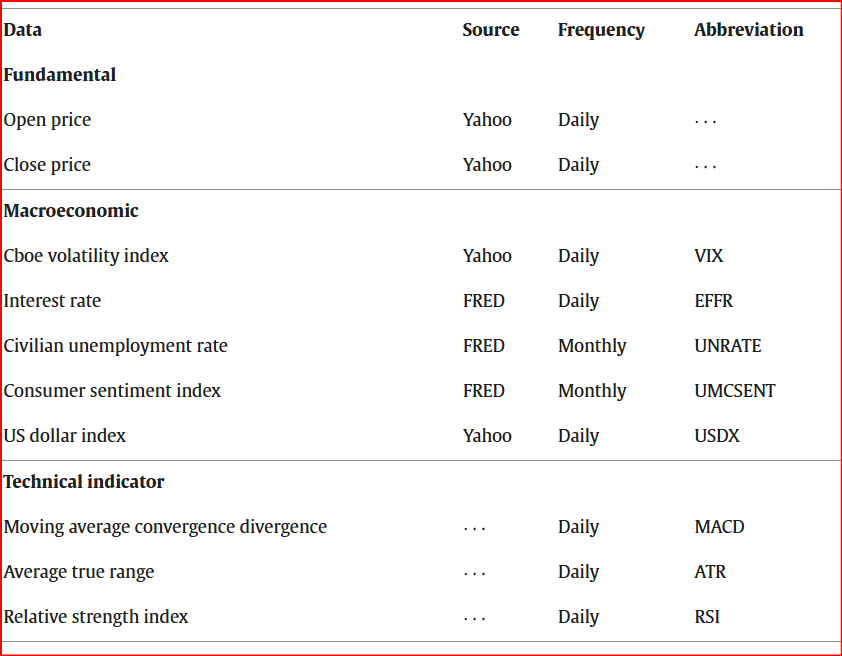


Fig: Snapshot of the dataset.

**Data Denoising, Normalization and Input Preparation:**

Stock price data are noisy and are in sequential discrete format. The discrete Wavelet transformation is common to denoise time-series data. Haar wavelets are the most suitable and popular in stock price data. We have applied the soft mode of the Haar wavelets using python library scikit-image to denoise the close price of the index. The values of input variables vary from one to another; thus, it leads to a high level of variation. For instance, the stock index close price is much higher than the interest rate. More specifically, the standard deviation of the close price is 695.33, which is significantly higher than the standard deviation 1.664 of the interest rate. If the range of one feature varies more widely than the others, most ML algorithms might not perform well. We have implemented a min–max normalization technique for the feature scaling to address this concern.

The min–max normalization technique is expressed in the following equation,

**Z= X-Xmin / Xmax-Xmin.** where, z, x are scaled and the original input respectively. Similarly, xmin and xmax are the minimum and the maximum values of the input respectively.

**Hyperparameter tuning:**

The final optimal model architecture is selected by exploring a wide range of possibilities. The overall [model selection procedure](https://www.sciencedirect.com/topics/engineering/model-selection-procedure) is divided into two broad categories.

Every model is executed 10 times for each combination of hyperparameters and the average RMSE score is calculated. The best possible combination of hyperparameters for the given model is chosen based on the lowest average RMSE score on the validation data.

In the first phase, the values of hyperparameters optimizer, initial learning rate, and batch size are optimized using the validation data. For the single layer LSTM architecture.

Data is initially collected below, including the following indexes: closing price, opening price, highest price, lowest price and trading volume corresponding to each trading session of the stocks in the list. This historical price data is processed through the following specific steps:

* Step 1: check the data, handle the defects of the data such as: empty data, data deviation. Instances with defective data will be checked and supplemented.
* Step 2: Calculate the corresponding technical analysis indicators for each stock, including: simple moving average (SMA), convergence divergence moving average (MACD), and relative strength index (RSI).
* Step 3: Historical price data is aggregated with the corresponding technical analysis indicator, observations that lack data due to differences in the calculation process of technical analysis indicators will be eliminated.
* Step 4: Aggregate data including price history and technical analysis indicators are used as input data for the Long Short-Term Memory (LSTM) model to make stock price forecasts.

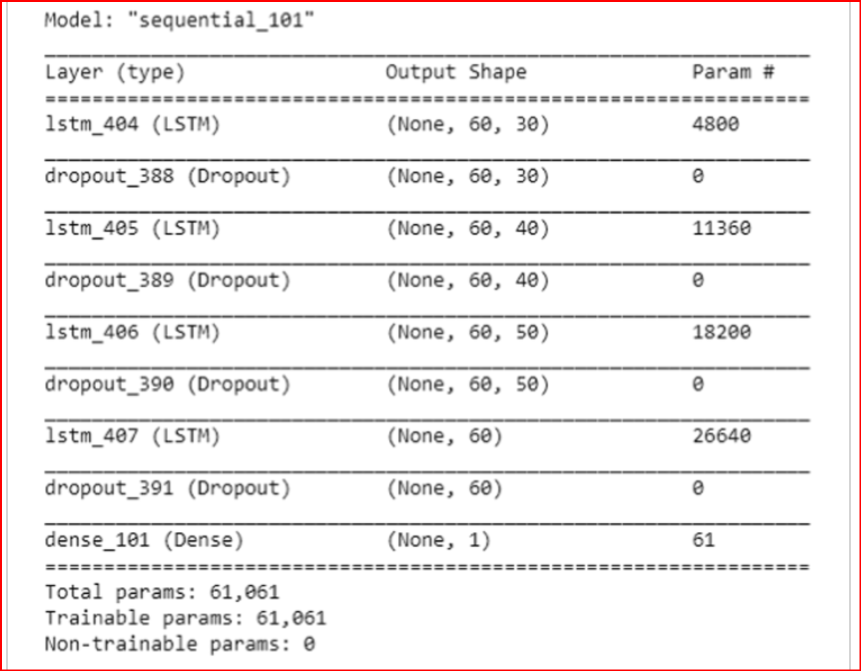
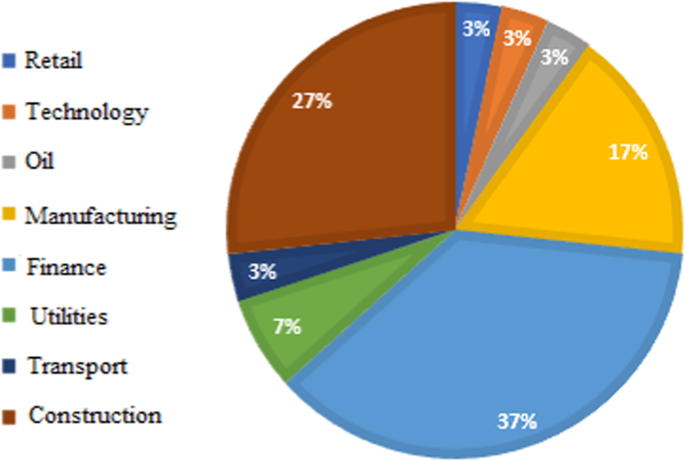


Fig: LSTM Model

When apply LSTM algorithm and technical analysis indicators to forecast price trends on the Vietnamese stock market. In this section, authors will present the results of the data after performing the analysis according to the research process and method, as well as make comments and discuss the research results.

The Figure shows the list of stocks in the VN-30 group has mainly in the industry groups: finance, construction and manufacturing. Figure indicates that companies in these industry groups have high corporate capitalization, hich is why many companies in these industry groups appear in the list of stocks selected for the study



The LSTM model predicts stock prices corresponding to the trading sessions in the test set. The test set data length includes observations from January 1, 2021 to April 1, 2021. Thus, there are all 78 trading sessions observed in the test set. For each different stock ticker, the forecast performance of the built model is also different.

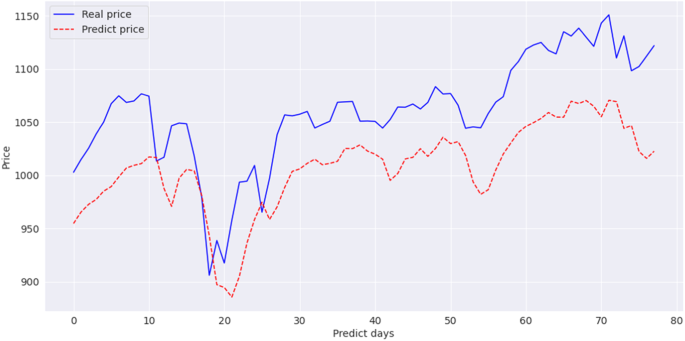
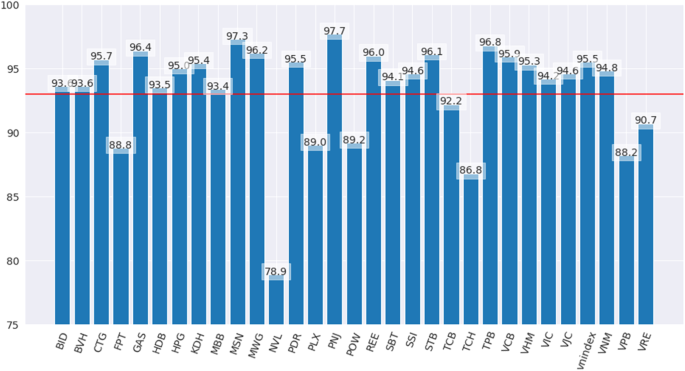


Fig: Forecast results and actual prices in the test set

It can be seen that the price forecast from the LSTM model tends to be very similar to the variation trend of the actual price on the data of the test set. In addition, the difference between the forecast price and the actual price is not significant. Note that for each data of different stocks, the accuracy of the LSTM model’s forecast will have a difference, a specific comparison chart between the forecast price and the actual price of all stocks. Stock codes carried out in the study are presented in the appendix.



(PRICE)

(STOCKS)

Fig: Accuracy level of LSTM Model.

Figure shows the level of accuracy in stock price forecast on the test set of the LSTM model corresponding to each stock in the research list. The red horizontal line represents the baseline at 93%. It can be seen that the LSTM model achieves a forecast accuracy higher than 93% for most of the stocks used in the study.

LSTM algorithm and technical analysis indicators to forecast price trends in Vietnam’s stock market, adding conclusions about forecasting performance and application level of LSTM. LSTM model for analyzing and forecasting stock price trends in Vietnam’s stock market in terms of both academic and practical applications.

**Individual Stock Performance:**

Here are the notable performances of various stocks:

* **Top Performers (Above 95)**:
  + **PAX** stands out with the highest value of **97.7**, making it the best performer.
  + Other strong performers include **BTH (95.7)**, **MNW (97.3)**, and **POW (96.1)**, showing values higher than the benchmark of 95.
  + These stocks could indicate companies performing well above industry standards, possibly reflecting high growth or returns.
* **Moderate Performers (Around 95)**:
  + **BID (93.6)**, **CTP (93.5)**, **NBT (93.4)**, and **NBB (95.5)** show close-to-benchmark performance, suggesting that they are meeting expectations but not exceeding them significantly.
  + These stocks might represent companies that are stable or growing at an average pace.
* **Low Performers (Below 90)**:
  + **MGN (78.9)** stands out with the lowest value, indicating significant underperformance.
  + **PVL (89.2)**, **VNC (88.8)**, and **VNE (88.2)** also fall below the 90 mark, which could indicate financial struggles or lagging performance in their respective sectors.

**CHALLENGES AND OPEN ISSUES:**

Financial market analysis and prediction continues to be a fascinating and challenging problem. Nowadays, data access is becoming easier, but difficulties are increasing in the acquisition and processing of data to extract valuable insights and analyze their impact on stock prices. Feature extraction from the financial data is a challenging task, as it is essential to observe the diversity of the variables that are used for the prediction. The Financial datasets are usually noisy. The quality of the data significantly affects SMPs.

Most literature on stock prediction regarding live testing affirms that the previously proposed methodologies can be utilized in real time. However, these methods may work in controlled circumstances. Still, a big challenge will be the live testing for the prediction. The live testing comes up with challenging factors, such as variations in prices, noise, and unpredicted events. One such example is the Knight Capital Tragedy, in which the loss of 440 million dollars was endured by the company

Market volatility is the severity with which the market price of an investment fluctuates. The main reasons for the volatility are uncertainty and inflation, and the risk increases when the market is volatile. The influence of volatility on our emotions is ceaseless. The prediction of stock prices is challenging when the market is volatile. One of the reasons for market volatility is algorithmic trading. One such example is the flash crash, which expunged $860 billion within 30 min from US stock markets. International politics also plays a dramatic role in stock market volatility.

Events in which panic selling is triggered are nowadays becoming more common, and they result in market overreaction. Panic selling is the reaction to fear and loss, which leads to the wide-scale selling of investments. The leading causes which result in panic selling are high speculation in the market, political issues, and economic instability. It becomes more difficult for a researcher to evaluate market behavior in such situations.

New algorithms are proceeding to flood the markets consistently at a pace, and it is challenging to compare the adequacy and exactness of these algorithms. A fascinating part of this research area is its self-defeating nature. In simple terms, sharing the methodologies that generate high profits with market competitors will render the methodologies useless. In this way, best-class algorithm exchanging in the markets is restricted, and is private. The procedure or strategy behind such algorithms is never published.

The data on social media platforms can either be generated by humans or bots. The sentiments of bots can sometimes result in inaccurate predictions. As such, there arises a need for social bot detection to obtain better predictions. Investigators, analysts, and researchers are continuously reporting the potential dangers brought about by social bots. Market investors actively participate in and react to social media sentiments. As such, it can be said that the data from social platforms play a significant role in stock prediction.

**CONCLUSION AND FUTURE WORK:**

Stock price prediction is the area of high interest for equity traders, individual investors, and portfolio managers. However, precise and consistent stock price prediction is a difficult task due to its noisy and [nonlinear behavior](https://www.sciencedirect.com/topics/engineering/nonlinear-behavior). There are several factors that can impact the prediction such as fundamental market data, macroeconomic data, technical indicators, and others. This study focuses on developing LSTM based models to predict S&P 500 index’s closing price by extracting a well-balanced combination of input variables capturing the multiple aspects of the economy and broader markets. Both single and multilayer LSTM architectures have been implemented and their performances are analyzed by using various evaluation metrics to identify the best model. The experimental results show that single layer LSTM model with around 150 [hidden neurons](https://www.sciencedirect.com/topics/engineering/hidden-neuron) can provide a superior fit and high prediction accuracy compared to multilayer LSTM. The proposed model can be easily customized to apply in other broad market indexes where the data exhibits a similar behavior. Interested stakeholders can use the proposed model to better inform the market situation before making their investment decisions.

In the near future, we plan to explore the possibility of incorporating unstructured textual information in the model such as investor’s sentiment from social media, earning reports of underlying companies, the immediate policy-related news, and research reports from market analysts. Another potential direction of the future work can be developing hybrid predictive models by combining the LSTM with some other [neural networks architectures](https://www.sciencedirect.com/topics/engineering/neural-network-architecture).

**REFERENCES:**

Here are some references and resources that can be useful for understanding and implementing the stock market analysis and prediction project with an LSTM model:

**1**. **Stock Market Data and Financial Indicators**

* **Yahoo Finance**: Yahoo Finance API and data (for historical stock price data).
* **Investopedia**:
  + Moving Averages - An explanation of SMA, EMA, and their significance in stock analysis.
  + Bollinger Bands - A tool for understanding market volatility.
  + Relative Strength Index – A momentum indicator to identify overbought conditions.
* **Quandl**: Quandl provides various datasets for financial and economic data, including stock prices and other financial indicators.

**2**. **Data Manipulation and Analysis in Python**

* **Pandas Documentation**: Pandas User Guide - Detailed documentation on how to use Pandas for data manipulation, handling missing data, and calculating moving averages.
* **Matplotlib**: Matplotlib Documentation - Guide to using Matplotlib for plotting and visualizing data in Python.
* **Seaborn**: Seaborn Documentation - For advanced statistical data visualization, useful for creating visually appealing plots.