**A comparative Analysis of Machine Learning Models for predicting S&P 500 Stock Market Returns**

**Abstract**

The stock market has always been a subject of great interest due to its unpredictable nature. In this study, the performance of two popular machine learning algorithms, Long Short-Term Memory (LSTM) and Support Vector Machine (SVM), are compared in predicting stock market returns. The primary objective of this research is to identify which algorithm demonstrates superior predictive capabilities.

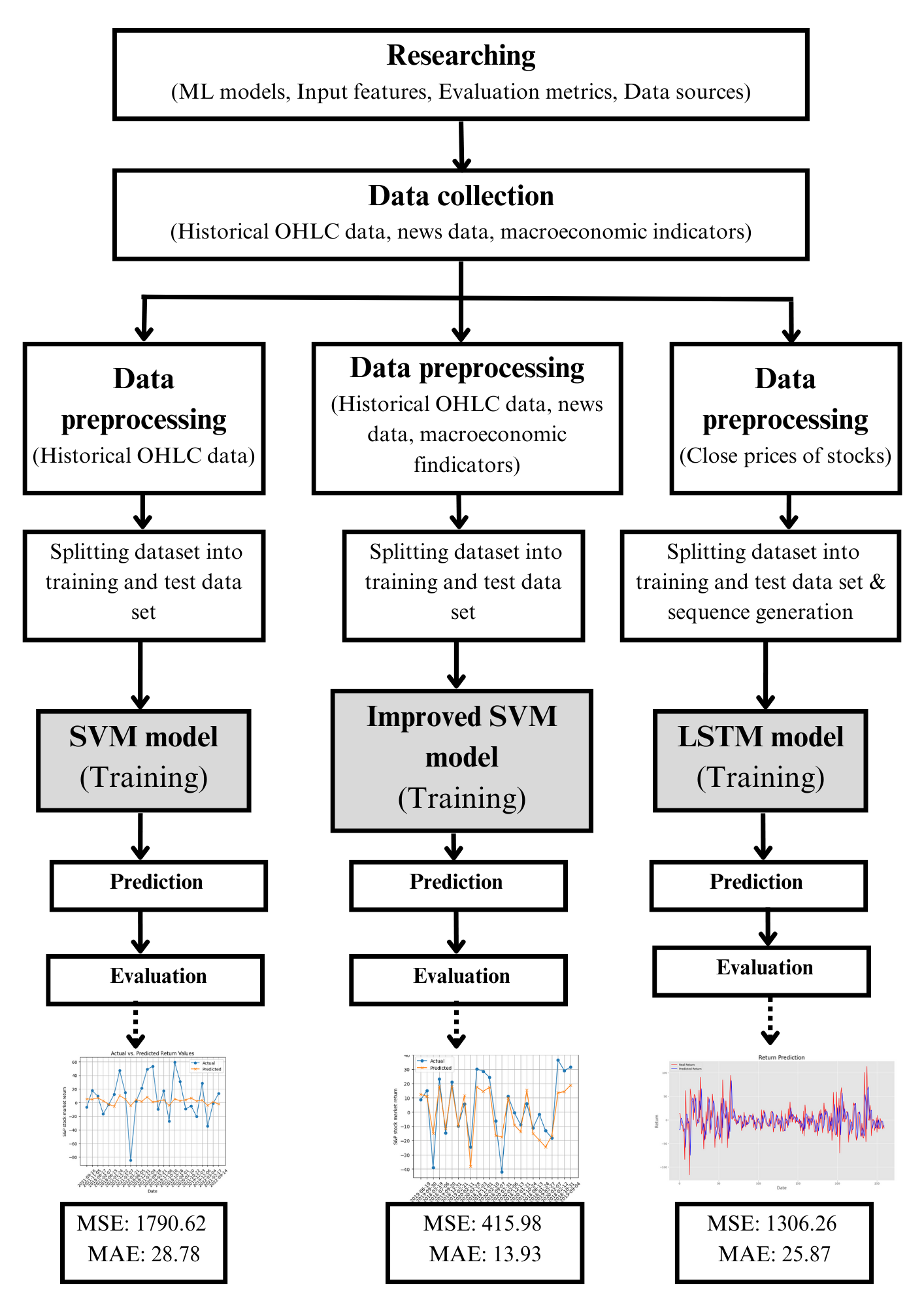
To achieve this objective, historical stock market data, including news sentiment data, were collected from various well-established financial markets. Then data were preprocessed to ensure the input's suitability for the LSTM and SVM models.

For the LSTM model, a deep learning architecture was constructed with multiple LSTM layers to capture temporal dependencies within the data. SVM was employed with a kernel trick to transform the input data into a higher-dimensional space, enabling the algorithm to predict stock market returns.

Key metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used to evaluate the performance of each model. The experimental results revealed that both LSTM and SVM demonstrated promising capabilities in predicting stock market returns. However, the LSTM model exhibited slightly superior performance.

The findings from this study show the potential of LSTM and SVM in predicting stock market returns and provide valuable insights for investors and financial analysts. This research lays the groundwork for further investigations into advanced hybrid models that leverage the strengths of both algorithms, leading to even more accurate and robust predictions.

**Graphical Abstract:**

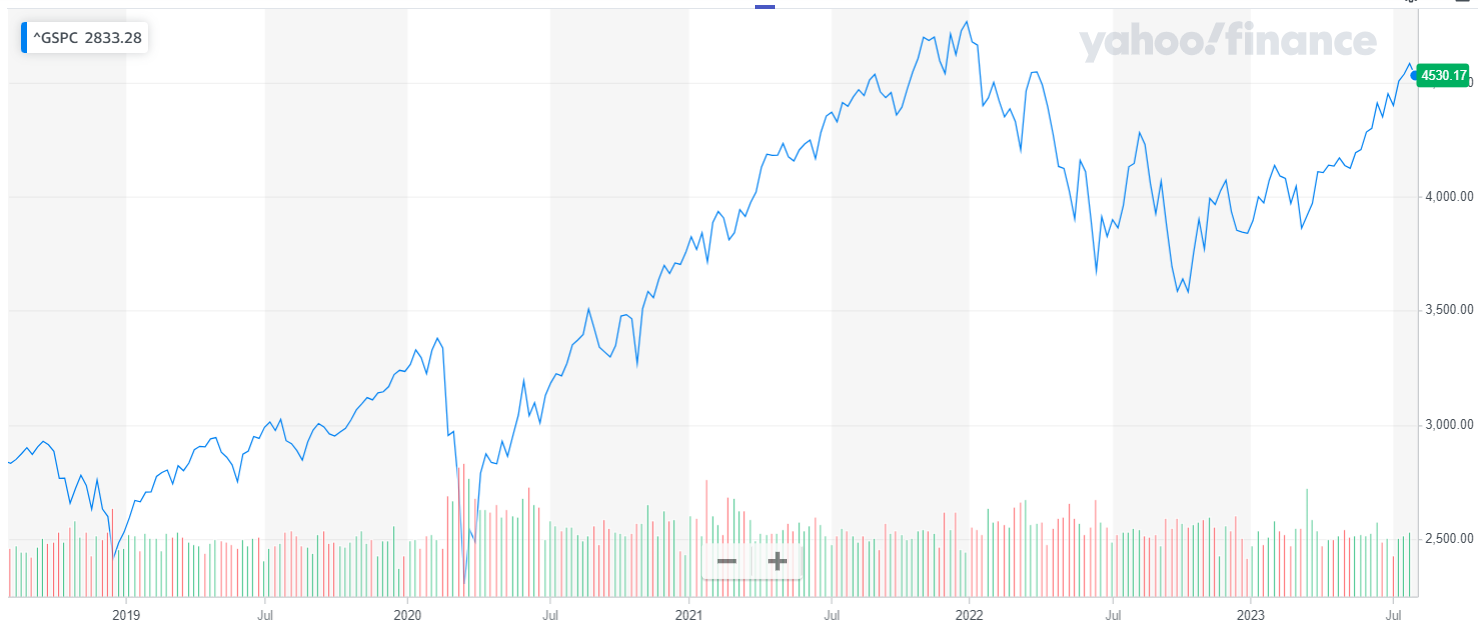


**Figure 1.** Graphical abstract of the project. The figure shows the graphical abstract of project

1. **Introduction**

The Standard and Poor’s 500, or simply the S&P 500 is a stock market index to track the value of 500 corporations that have listed their stocks on the New York Stock Exchange (NYSE). This collection of stocks represents most of the composition of the U.S. economy. Therefore, this index is closely watched by market participants such as investors, traders, financial analysts, business leaders, etc. For them, being able to predict the returns of the S&P 500 is significant since it allows them to optimize their marketing strategies.

S&P 500 stock market return is referring to the percentage change in the value of the S&P 500 index over a specific period, typically daily, weekly, monthly, or annually. Machine Learning (ML) and Deep Learning (DL) models’ ability to forecast S&P 500 market returns are tested and compared as those techniques have gained popularity in financial forecasting. This comparison based on the evaluation metrics and overall performance can be used to determine the advantages and disadvantages of each model.



**Figure 2.** Historical data chart of S&P (^GSPC). The figure shows how the S & P stock market index has varied over the last few years.

Even with those cutting-edge technologies, predicting S&P 500 returns is quite challenging due to the complex and dynamic nature of financial markets including macroeconomic factors and sentiment data.

1. **Methodology/ Methods/ Development of Models**
2. *Data Collection*

Then need to collect the historical data, macroeconomic data and news data related S&P 500 stock market from the beginning of 2018 to the beginning of 2023 with a daily frequency. These data can be obtained from trustworthy resources such as Yahoo Finance [1]. It is crucial to have a good quality and reliable data set for the success and effectiveness of a model. The relevance of data significantly affects the model’s performance, generalization ability, and prediction accuracy.

OHLC (stands for Open, High, Low, and Close) are the four key data points used to describe the price movements of a stock. The stock market is influenced by multiple factors including macroeconomic indicators, geopolitical events are market sentiment which make it dynamic. In order to predict S&P stock market returns accurately, consideration of those factors is a must.



**Figure 3.** OHLC data chart. Figure shows the dataset provided by Yahoo Finance which was used to model.

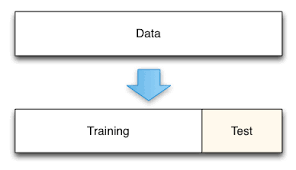
Moving Average (MA) is one of many indicators used in financial analysis, which calculates the average price of a stock over a specific period of time. It helps to smooth out short-term price fluctuations and identify trends in price movements over a specific period. The relative Strength Index (RSI) is used to assess the strength and speed of price movements in a stock. These two indicators are also used in addition to the OHLC to predict stock market returns.

In addition to all those factors, positive and negative news can impact stock prices. Therefore, including news sentiment in the dataset is a valuable approach to account for the impact of news on stocks.

1. *Data pre-processing*

Data preprocessing is an important step that involves preparing raw data to suit training by handling any anomalies. It includes handling missing values, encoding categorical variables, feature selection, feature scaling, and data splitting. Feature Engineering comes in when creating informative features that can help the model extract meaningful patterns from the data.

If there are missing data in a dataset can choose to remove those particular rows or fill them with default values. It is needed to convert categorical variables to numerical values as ML algorithms need numerical inputs. In order to reduce computational complexities and avoid overfitting need to select the most relevant features for the model. Feature scaling ensures that no single feature dominates and scales numerical features to a comparable range.



**Figure 4.** Dataset splitting. The figure shows how a dataset is divided into a training dataset and a test dataset.

Finally, these preprocessed data are divided into training and test sets. The training set is used to train the model while the test set is used to evaluate its performance on unseen data.

1. *Model selection*

Support Vector Machines (SVM) and Long Short Term Memory (LSTM) models were selected from each domain of Machine Learning and Deep Learning respectively in order to perform the prediction and comparative analysis.

* Support Vector Machines: SVM is an ML model which performs well in high-dimensional feature spaces and can efficiently capture complex and non-linear relationships between input features and the output feature. SVMs reduce the risk of overfitting which is highly beneficial when dealing with historical financial data.
* Long Short Term Memory: LSTM is a type of Recurrent Neural Network that is designed to handle sequential data with time dependencies and can capture long-term dependencies and non-linear patterns. LSTMs can handle data with irregular time intervals well and this feature comes in handy as stock market returns are recorded at specific trading days.

1. *Model training*

SVM can be used for both regression and classification tasks, in order to predict stock market returns Support Vector Regression (SVR) is used [2]. The training data set is used to train the model with the best hyperparameters such as kernel type.

Sequence generation should be done when feeding training data into the LSTM model where it converts the training data into time windows and each sequence contains the set of input features (OHLC, indicators, news sentiments) and the target variable (S&P 500 return). It is also important to tune hyperparameters such as neurons, hidden layers, learning rate, and batch size to train well.

1. *Model evaluation*

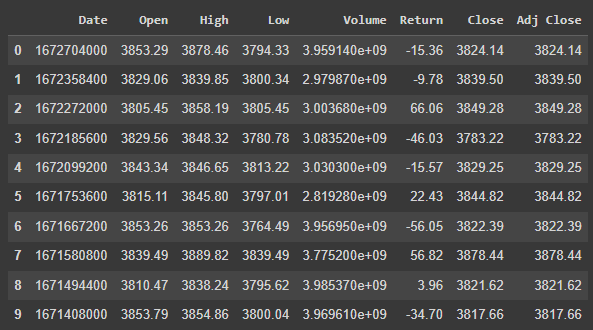
The model’s performance is evaluated on the separate test dataset to understand how well it predicts the stock market return using evaluation metrics. Mean Squared Error (MSE), and mean absolute error (MAE) are the evaluation metrics that can be used to evaluate the particular model’s performance and also the performance comparison between two models. MAE can be represented as follows when , and N is actual output, predicted output, and the total number of data points respectively [3].

MSE takes the difference between the model’s prediction and the original value and averages it across the whole dataset. MSE is great to learn outliers as it puts a larger weight on these errors. MAE takes the absolute value of the difference between the model’s prediction and the original value and averages it across the whole dataset. It will provide a generic measure of how well the model is performing as all the errors will be weighted on the same linear scale unlike in MSE.

Evaluation metrics help to assess how an ML model is performing as they provide quantitative measures.

1. **Results and Discussion**
2. *SVM model with OHLC data*

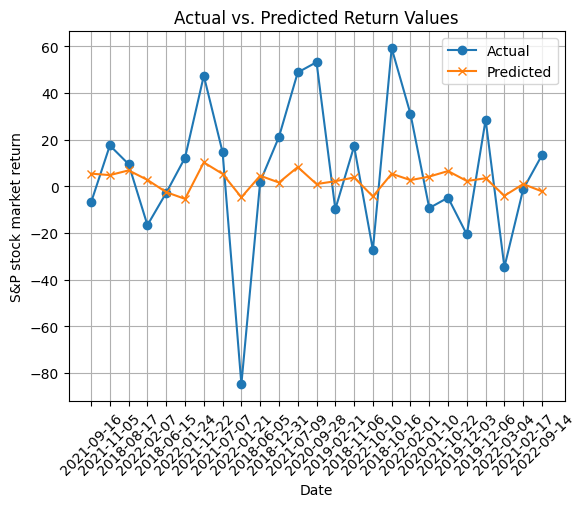
The dataset was acquired and pre-processed prior to the model training. The dataset contains 1260 samples for each trading day in the past few years with their respective Open, High, Low, Volume, Close, and Adj close data. The dataset also includes the target variable which is the daily return.

****

**Figure 5.** Preprocessed dataset for the SVM model. The figure shows the features which were used to train the model.

|  |  |
| --- | --- |

**Figure 6.** Actual vs Predicted returns of the SVM model. The figure shows the actual returns and predicted returns for each day.



**Figure 7.** Actual vs Predicted returns of the SVM model. Figure shows the actual returns and predicted returns for each day.



**Figure 8.** Evaluation metrics for the SVM model. The figure shows the MAE and MSE when trained with the SVM model (with only the OHLC dataset)

1. *SVM model with OHLC data, MA20, and news sentiment data*

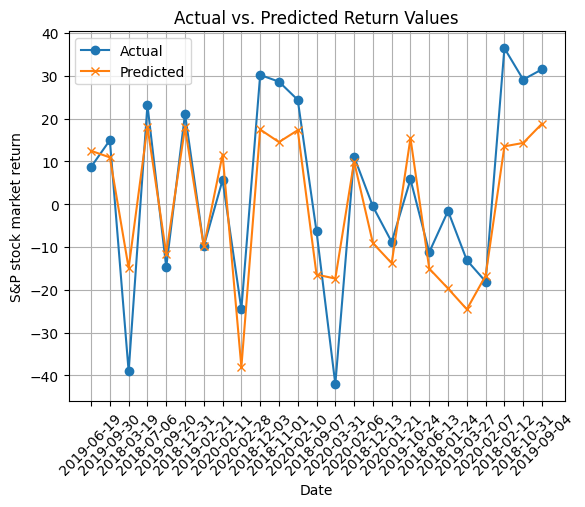
The dataset was acquired and pre-processed prior to the model training. Moving Average for 20 days are also included in the dataset. Additionally, historical market news data was obtained from Kaggle and merged with the current dataset in order to consider those factors in the model training. Positive and negative news can have an impact on stock prices.



**Figure 9.** Preprocessed dataset for the improved SVM model. Figure shows the features which were used to train the model.

|  |  |
| --- | --- |

**Figure 10.** Actual vs Predicted returns of the improved SVM model. Figure shows the actual returns and predicted returns for each day.

****

**Figure 11.** Actual vs Predicted returns of the improved SVM model. Figure shows the actual returns and predicted returns for each day.



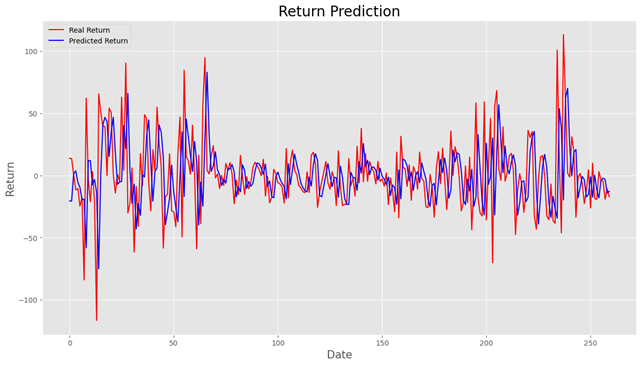
**Figure 12.** Evaluation metrics for the improved SVM model. The figure shows the MAE and MSE values when trained with an improved SVM model.

1. *LSTM model (Univariate Time Series)*

LSTMs are a type of RNN that remember information over long periods of time making them better suited for predicting stock prices.

Time series forecasting is about estimating the future value of a time series on the basis of historical data. In a univariate model, the focus is on a single dependent variable. The basic assumption behind the univariate prediction approach is that the value of a time series at time-step t is closely related to the values at the previous time steps t-1, t-2, and so on. Univariate models are useful when the target variable's historical data is the most crucial factor for prediction and there is little evidence of strong interdependencies between multiple variables.

LSTM is commonly used for stock price prediction as it has strong time series predictive capabilities [4]. The dependent variable used to train the LSTM is the closing price stock.



**Figure 13.** Actual vs Predicted returns of the LSTM model. Figure shows the actual returns and predicted returns.



**Figure 14.** Evaluation metrics for the LSTM model. Figure shows the MAE and MSE values when trained with the LSTM model.

The evaluation metrics for each model are shown below.

**Table 1**. Comparison of MSE and MAE for each model

|  | **Mean Squared Error** | **Mean Absolute Error** |
| --- | --- | --- |
| **SVM model (with OHLC data)** | 1790.62 | 28.78 |
| **Improved SVM model (with OHLC, ma20, news data)** | 415.98 | 13.93 |
| **LSTM model - Univariate** | 1306.26 | 25.87 |

1. **Conclusion**

The conclusions drawn from the above results can be listed as follows.

* Improved SVM is the best-performing model in terms of both metrics as it has the smallest overall errors in its predictions compared to the other models.
* SVM model performance can be improved by considering external factors such as news events, and economic indicators.
* The improved SVM model has exhibited superior predictive performance in identifying the positive and negative aspects of return values compared to the SVM model.
* LSTM model has performed well even when considering the univariate time series (without considering additional features).

By considering additional features, multivariate models may improve forecasting accuracy, especially in situations where the target variable's behavior depends on other variables' interactions [5].

1. **References**

[1]. “S&P 500 (^GSPC) Historical Data.” *Yahoo! Finance*, 6 Aug. 2023, finance.yahoo.com/quote/%5EGSPC/history?period1=1514764800&period2=1672531200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true.

[1]. Meesad, Phayung, and Risul Islam Rasel. “Predicting Stock Market Price Using Support Vector Regression.” *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, 2013, https://doi.org/10.1109/iciev.2013.6572570.

[3]. Chugh, Akshita. “Mae, MSE, RMSE, Coefficient of Determination, Adjusted R Squared - Which Metric Is Better?” *Medium*, 16 Mar. 2022, medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

[4]. Lian, Yulong Eric. “Deep Learning on Spy Stock Prices Using Improved Multi-Head LSTM.” *International Journal of Finance Research*, vol. 3, no. 2, 2022, pp. 135–146, https://doi.org/10.47747/ijfr.v3i2.785.

[5]. Jarrah, Mutasem, and Morshed Derbali. *Predicting Saudi Stock Market Index by Using Multivariate Time Series Based on Deep Learning*, 2023, https://doi.org/10.20944/preprints202306.1537.v1.