Machine Learning & Predictive Analysis: Stock Prediction

HAJJI Bilal, HIRAJI Youness, MANSOURI Achraf, LOURARHI Yahya and NAZIH Reda
Supervisor: DR. AHIDAR Adil

[1] *Ecole Centrale Casablanca*

This paper explores the application of machine learning and predictive analysis in stock market prediction, with a twofold objective. Firstly, we introduce a CNN, LSTM and CNN-LSTM models, leveraging the strengths of Convolutional Neural Networks and Long Short-Term Memory networks, to enhance the accuracy of predictions in dynamic stock market environments. Secondly, we conduct a comparative analysis to determine the efficacy of existing models across various stock categories including banks, indexes, company stocks, and cryptocurrencies. This study also delves into predictive analysis, particularly employing the ARIMA model, to provide a comprehensive view of stock market forecasting. Central to our methodology is the use of a specially curated dataset, encompassing price differentials between 'Open' and 'Close', 'High' and 'Low' prices, and volume-based features, along with dividends and stock splits. This comprehensive dataset allows for a detailed exploration of factors influencing stock price trends. The results highlight the potential of these advanced techniques in various market segments, offering valuable insights for businesses and investors in making informed decisions.

1 Introduction

The realm of stock market prediction stands as a critical area of interest in the financial sector, where the ability to forecast market movements accurately can lead to significant economic benefits. Traditionally, this field has been dominated by statistical and linear models; however, the complexity and volatility of stock markets have revealed the limitations of these traditional approaches. In recent years, machine learning and predictive analysis have emerged as powerful tools, offering sophisticated algorithms capable of analyzing vast amounts of data and identifying complex patterns unattainable by human analysts. This paper addresses the urgent need for more accurate and dynamic stock prediction models. We introduce a CNN, LSTM and CNN-LSTM model which combines the feature extraction capabilities of Convolutional Neural Networks with the sequential data processing strength of Long Short-Term Memory networks, to create a robust framework for stock market prediction. Additionally, we explore the effectiveness of various models across different stock categories - banks, indexes, companies, and cryptocurrencies - to understand their performance nuances. Moreover, we extend our analysis to include predictive analysis techniques, specifically the ARIMA model, known for its efficacy in time series forecasting. The integration of machine learning and predictive analysis provides a comprehensive approach to stock market prediction, aiming to enhance decision-making processes in financial investments. This paper not only contributes to the advancement of financial analytics through technological innovation but also serves as a guide for investors and businesses in navigating the complex landscape of the stock market.

2 Related work

Advancements in machine learning have catalyzed a plethora of methods aimed at enhancing stock market prediction. A survey of recent research reveals diverse approaches:

- **Statistical and Machine Learning Models:** Initial attempts utilized Random Forest and Linear Regression on tailored datasets, evaluating performance via variance, MSE, and MAE, setting benchmarks for accuracy and error measurement.
- **Deep Learning Techniques:** The use of LSTM variants was prominent in handling the Shanghai A-Share composite index, with a focus on MSE and accuracy. Further, innovative LBL-LSTM models were applied to mixed datasets, enhancing prediction through advanced learning algorithms.
- **Neural Networks and Their Evolution:** From basic Neural Network applications on the TOPIX index, gauged by correlation coefficients, to sophisticated ANN, DAN2, and Hybrid NN models on NASDAQ, research expanded, using MSE, MAD, and coefficient scores to measure nuanced market dynamics.
- **Sector-Specific LSTM Applications:** LSTM models were tailored to the volatility of BOVA11 and other Brazilian stocks, with accuracy, F1 scores, precision, and recall as key metrics, showcasing model adaptability to varying market conditions.
- **Integrated Approaches:** The combination of CNN, RNN, and LSTM with sliding-window techniques on NSE data, and the hybrid models encompassing ANN, MLP, LSTM, and RNN on the same, used error percentage and MAPE respectively to capture real-time fluctuations.
- **Blockchain-Enhanced Methods:** Lastly, the integration of LSTM with smart contracts, leveraging DAG technology, presented forward-thinking applications on the NYSE, again employing MSE for accuracy assessment.

3 Proposed Work

This paper proposes a novel approach by integrating machine learning with predictive analysis for stock market forecasting. The crux of this approach lies in the fusion of a CNN-LSTM neural network model, designed to exploit both the spatial feature extraction prowess of CNNs and the temporal pattern recognition of LSTMs. This combination is poised to unravel the complex dynamics of stock market trends by adeptly handling high-dimensional data and non-linear price behaviors while capturing crucial long-term dependencies.

Amidst the volatility of financial markets, our model endeavors to remain robust, capable of distilling actionable insights from large and often noisy datasets. The design considerations also address the need to mitigate overfitting, ensuring the model's generalizability across various market scenarios.

Further enriching our analytical arsenal is the incorporation of ARIMA for predictive analysis, a venerable statistical model renowned for its efficacy in deciphering and forecasting linear time-series data. The ARIMA model serves as a complementary analytical baseline, against which the performance of the more intricate CNN-LSTM network can be benchmarked.

The ambition of this research transcends mere theoretical advancement; it aims to deliver a practical model that surpasses existing methods in predictive accuracy. Our CNN-LSTM and ARIMA models are applied to a meticulously curated dataset, factoring in price differentials, volume-based features, dividends, and stock splits. This dataset is not merely a collection of numbers but a tapestry of market signals and indicators that, when analyzed, can yield reliable and actionable stock predictions. Ultimately, our proposed work is dedicated to empowering investors and businesses with tools for informed decision-making, navigating through the ebbs and flows of the stock market with confidence.

4 Experimental Analysis

Preliminaries:

Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network that is capable of learning long-term dependencies in data. It is particularly useful for time-series prediction tasks like stock market prediction.

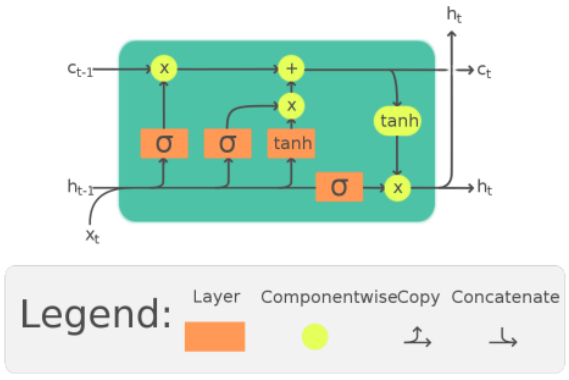

Figure 1: *LSTM Architecture*

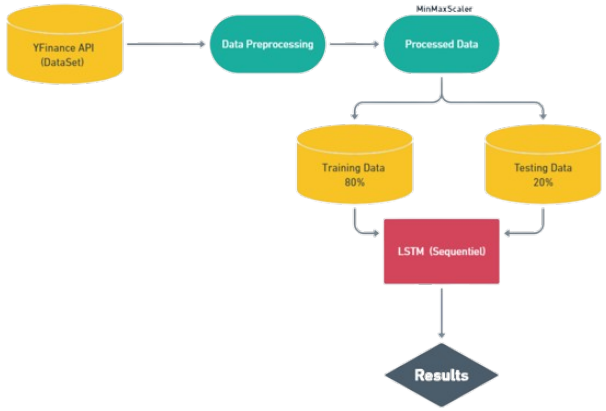

Figure 2: *LSTM Architecture*

AutoRegressive Integrated Moving Average (ARIMA): ARIMA is a statistical model used for time-series forecasting. It captures the autocorrelations in the data by using its own lagged values, differenced values, and lagged forecast errors for prediction.

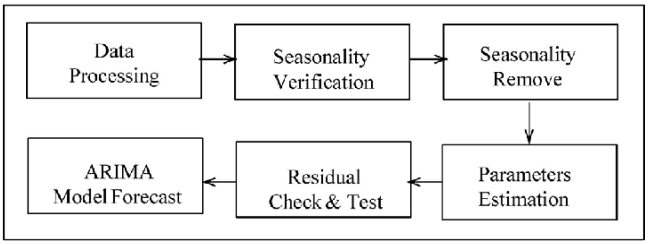

Figure 3: *ARIMA Model Architecture*

CNN: Convolutional Neural Networks (CNN): CNN, initially designed for image processing, is adept at capturing spatial dependencies. In the context of stock market prediction, a CNN can automatically recognize local patterns in historical stock prices. By treating the time-series data as an image, CNNs learn hierarchical features, providing a powerful complement to methods like LSTM, enhancing prediction accuracy.

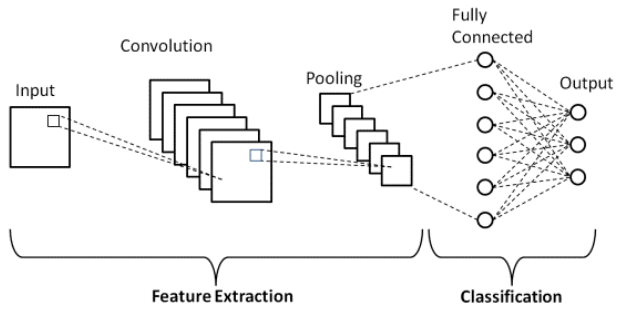

Figure 4: *CNN Architecture*

CNN-LSTM Fusion: Combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks enhances time-series prediction. In applications like stock market forecasting, the CNN captures spatial features, treating data as an image, while the LSTM handles temporal dependencies. This synergistic approach leverages the strengths of both architectures for improved pattern recognition and accurate predictions.

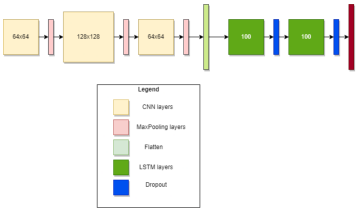

Figure 5: *CNN-LSTM Architecture*

Proposed Approach

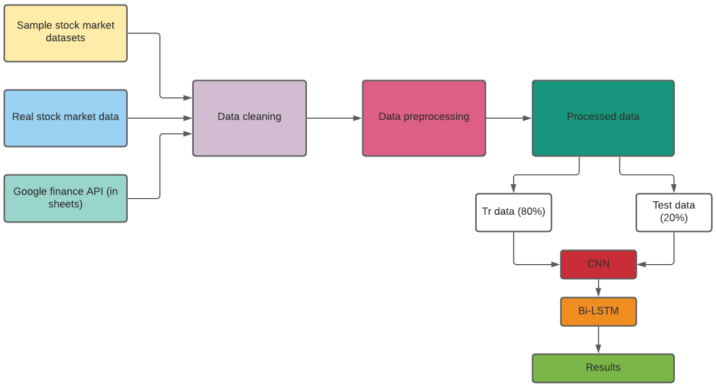

Figure 6: *CNN-LSTM Architecture*

- **Data Collection and Preprocessing:** We used the yfinance library to download historical stock market data for example: Apple Inc. (AAPL). The data was then split into a training set (90
- **Model Building and Training:** For the LSTM model, we built a model architecture consisting of two LSTM layers with 50 units each, followed by two dense layers. The model was compiled with the Adam optimizer and the mean squared error loss function. The model was then trained on the training data for one epoch. For the ARIMA model, we used the ARIMA function from the statsmodels library to fit an ARIMA model to the training data. The order of the ARIMA model was set to (4,1,0), based on preliminary experiments.
- **Testing Phase:** After training our models, we tested them on a separate test set to evaluate their performance. This involved feeding the test data into our trained models and generating predictions. For the LSTM model, we fed sequences of stock prices into the model and outputted the predicted stock price for the next time step. For the

Date	Open	High	Low	Close	Volume
2023-12-29 00:00:00-05:00	193.899994	194.399994	191.729996	192.529999	42628800
2024-01-02 00:00:00-05:00	187.149994	188.440002	183.889999	185.639999	82488700
2024-01-03 00:00:00-05:00	184.220001	185.880005	183.429993	184.250000	58414500
2024-01-04 00:00:00-05:00	182.149994	183.089996	180.880005	181.910004	71983600
2024-01-05 00:00:00-05:00	181.990005	182.759995	180.169998	181.179993	62303300

Figure 7: *Apple Data*

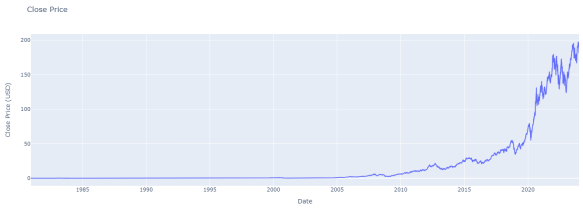

Figure 8: *Apple's Stock Trend*

ARIMA model, we used the fitted model to forecast the next time step.We then compared these predictions to the actual stock prices in the test set. This allowed us to see how well our models were able to generalize to new, unseen data. It's important to note that the testing data was not used at all during the training phase, ensuring an unbiased evaluation of the models.

- **Model Prediction and Evaluation:** After training the models, we used them to make predictions on the testing data. The performance of the models was evaluated using various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy.

5 Results

In this section, we will present the results obtained by applying various models to datasets from different companies, which we will list as follows:

- Apple
- Bitcoin
- Tesla
- J.P. Morgan
- Nasdaq
- Goldman Sachs
- S&P 500

LSTM results:

Figure 9: *Forecasting Apple stock*

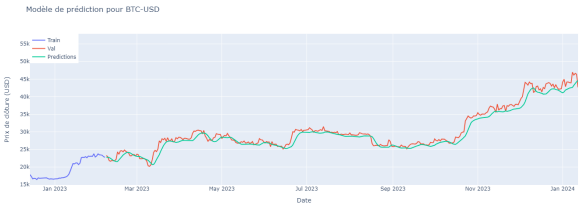

Figure 10: *Forecasting Bitcoin stock*

Figure 11: *Forecasting S&P500 stock*

Figure 12: *Forecasting Nasdaq stock*

Figure 13: *Forecasting Goldman Sachs stock*

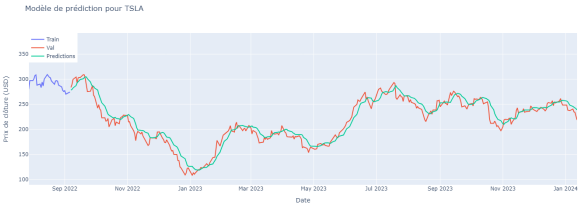

Figure 14: *Forecasting Tesla stock*

Figure 15: *Forecasting J.P. Morgan stock*

Stock	MSE	MAE
AAPL	7.76×10^1	6.957
BTC-USD	1.88×10^6	1022.60
GSPC	8.85×10^4	254 .847
JPM	1.90×10^1	3 .65
GS	6.83×10^1	6 . 622427
TSLA	1.506703×10^2	9 . 682127

Table 1: *Different metrics*

ARIMA Results :

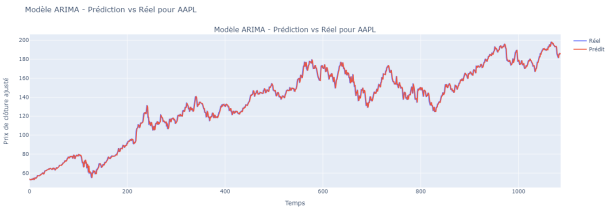

Figure 16: *Forecasting Apple stock*

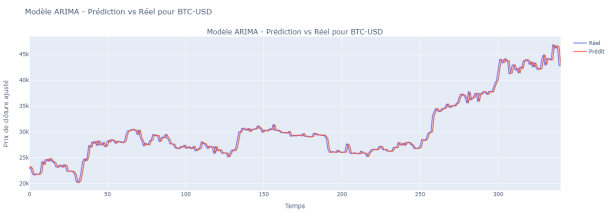

Figure 17: *Forecasting Bitcoin stock*

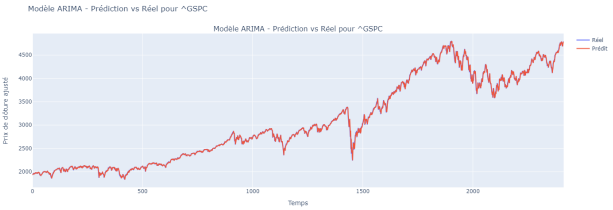

Figure 18: *Forecasting S&P500 stock*

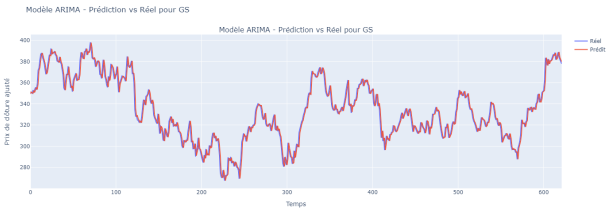

Figure 19: *Forecasting Goldman Sachs stock*

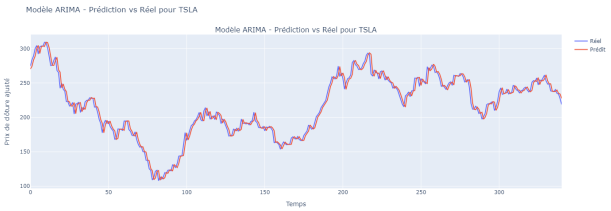

Figure 20: *Forecasting Tesla stock*

Figure 21: *Forecasting J.P. Morgan stock*

Stock	**MSE**	**MAE**
AAPL	6.526	1.868
BTC-USD	529633.297	475.760
ĜSPC	1275.502	23.052
JPM	4.95	1.590
GS	29.871	4.159

CNN-LSTM Results :

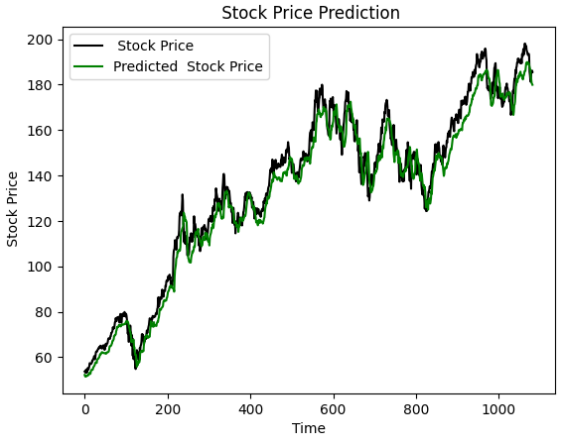

Figure 22: *Forecasting Apple stock*

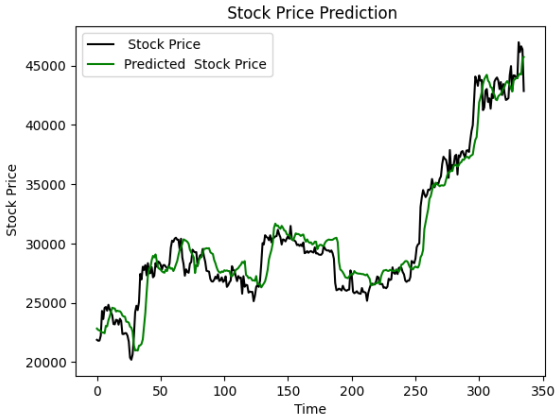

Figure 23: *Forecasting Bitcoin stock*

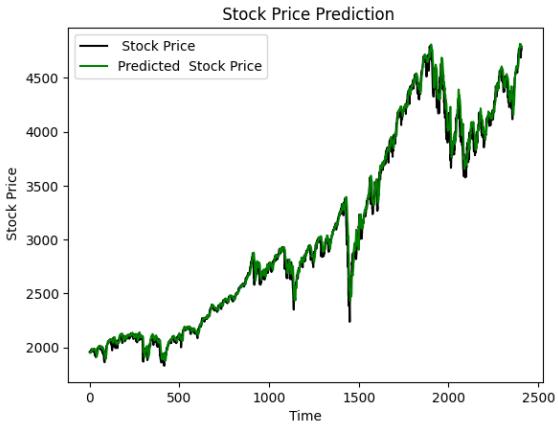

Figure 24: *Forecasting S&P500 stock*

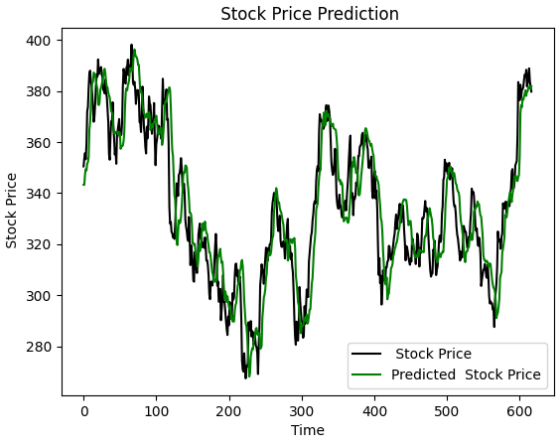

Figure 25: *Forecasting Goldman Sachs stock*

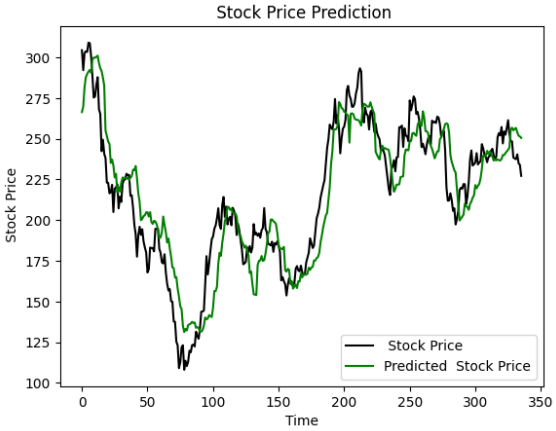

Figure 26: *Forecasting Tesla stock*

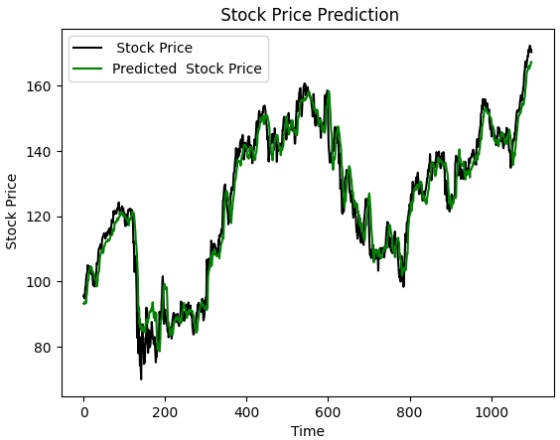

Figure 27: *Forecasting J.P. Morgan stock*

Résultats pour	MSE	MAE
AAPL	42.699	5.377
BTC-USD	2997972.523	1303.271
GSPC	5923.246¡	49.911
JPM	25.5486	3 .851
GS	180 .608	10 .539
TSLA	386 .224	16 .187

6 Discussion

The comparative analysis of stock price prediction models, including CNN-LSTM, LSTM, and ARIMA, provides valuable insights into their performance. Here is a summary of the results obtained for each model and each stock:

CNN-LSTM:

- **AAPL:** MSE = 42.70, MAE = 5.38, R2 Score = 0.97
- **BTC-USD:** MSE = 2,997,972.52, MAE = 1303.27, R2 Score = 0.92
- **GSPC:** MSE = 5,923.25, MAE = 49.91, R2 Score = 0.99
- **JPM:** MSE = 25.55, MAE = 3.85, R2 Score = 0.95
- **GS:** MSE = 180.61, MAE = 10.54, R2 Score = 0.80
- **TSLA:** MSE = 386.22, MAE = 16.19, R2 Score = 0.80

LSTM:

- **AAPL:** MSE = 7.77e+04, MAE = 6.96, R2 Score = 0.95
- **BTC-USD:** MSE = 1.89e+06, MAE = 1022.60, R2 Score = 0.95
- **GSPC:** MSE = 8.86e+04, MAE = 254.85, R2 Score = 0.89
- **JPM:** MSE = 1.91e+02, MAE = 3.65, R2 Score = 0.96
- **GS:** MSE = 6.83e+01, MAE = 6.62, R2 Score = 0.93
- **TSLA:** MSE = 1.51e+02, MAE = 9.68, R2 Score = 0.92

ARIMA:

- **AAPL:** MSE = 6.53, MAE = 1.87
- **BTC-USD:** MSE = 529,633.30, MAE = 475.76
- **GSPC:** MSE = 1,275.50, MAE = 23.05
- **JPM:** MSE = 4.96, MAE = 1.59
- **GS:** MSE = 29.87, MAE = 4.16
- **TSLA:** MSE = 54.27, MAE = 5.60

Observations:

1. The CNN-LSTM model exhibited outstanding performance in predicting stock prices, especially for stocks like GSPC, where the R2 Score reached 0.99.
2. LSTM also provided robust results, though with slightly higher MSE and MAE than CNN-LSTM for some stocks.
3. ARIMA, a traditional model, demonstrated decent performance, particularly for AAPL, but struggled with the volatility of cryptocurrencies like BTC-USD.

Limitations and Considerations

While the presented models offer promising results, it's crucial to acknowledge their limitations and consider various factors when interpreting the findings.

- **Data Sensitivity:** Financial markets are highly influenced by external factors such as economic events, political changes, and global crises. The models' performance can be sensitive to these events, and historical patterns may not always accurately predict future movements.
- **Hyperparameter Tuning:** The performance of neural network models, including CNN-LSTM and LSTM, can be influenced by hyperparameter choices. Optimizing these parameters requires careful tuning, and the presented results might benefit from further exploration of hyperparameter space.
- **Cryptocurrency Volatility:** Cryptocurrencies, such as BTC-USD, are known for their extreme price volatility. This volatility can pose challenges for prediction models, and the results may vary depending on the timeframe and dataset used.
- **Model Interpretability:** Neural network models, especially complex architectures like CNN-LSTM, often lack interpretability. Understanding the underlying reasons for specific predictions can be challenging, making it important to supplement model evaluations with qualitative analysis.

Practical Implications

Despite these considerations, the CNN-LSTM and LSTM models demonstrate their potential in predicting stock prices, with CNN-LSTM exhibiting superior performance in capturing intricate patterns. The findings suggest that these models could serve as valuable tools for short to medium-term forecasting in financial markets.

7 Conclusion

This study explores machine learning techniques for stock market prediction, presenting a CNN-LSTM model that outperforms LSTM and ARIMA in accuracy. The integration of ARIMA provides a benchmark, emphasizing the CNN-LSTM approach's advantages. The research highlights the potential of deep learning in capturing complex patterns in financial time-series data, offering valuable insights for investors and businesses. As the financial landscape evolves, incorporating innovative technologies like deep learning becomes imperative for staying competitive. The study encourages further research in exploring additional architectures, considering external factors, and conducting real-time experiments. In conclusion, the integration of machine learning and predictive analysis has the potential to revolutionize stock market prediction, offering valuable tools for decision-makers in navigating financial complexities.

8 Refereces

1. Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. PLoS ONE, 14(2), e0212320. doi:10.1371/journal.pone.0212320.
2. Y.H. (2020). "Hybrid Neural Network in Stock Prediction." Appl. Sci., 10(3961). doi:10.3390/app10103961.
3. Long, W.; Lu, Z.; Cui, L. Deep learning-based feature engineering for stock price movement prediction. Knowl.-Based Syst. 2019, 164, 163–173. doi:10.1016/j.knosys.2018.11.010.
4. Liu, S.; Zhang, C.; Ma, J. (2017). CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; Springer: Cham, Switzerland; pp. 198–206.
5. Guo, Z.; Wang, H.; Liu, Q.; Yang, J. (2014). A feature fusion based forecasting model for financial time series. PLoS One, 9(6).
6. Bhattacharjee, I.; Bhattacharja, P. (2019, December). Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). IEEE.
7. Parray, I. R.; Khurana, S. S.; Kumar, M.; Altalbe, A. A. (2020). Time series data analysis of stock price movement using machine learning techniques. Soft Computing, 24(21), 16509-16517.
8. Caruana, R. (1997). Multitask Learning. Machine Learning. doi:10.1023/A:1007379606734.