

Comparative Analysis of Stock Price Prediction Using Machine Learning on Multi-Source Open Datasets

Thin Zar Aung¹, Hlaing Htake Khaung Tin²

Faculty of Information Science

University of Information Technology Yangon, Myanmar

¹thinzaraung192000@gmail.com, ²hlainghtakekhaungtin@gmail.com

Abstract: Forecasting the dynamics of the stock market remains one of the most challenging tasks in financial analytical mechanics due to its stochasticity, nonlinearity, and dynamic character. In this paper, the author offers a detailed comparative analysis of two groups of machine learning and deep learning models available on two publicly accessible databases of Yahoo Finance and Kaggle. This research serves as an assessment of the forecast accuracy provided by the LR, RF, and LSTM network for stock price prediction. To guarantee proper model training and validation, the proposed methodology includes data pre-processing, feature engineering, time-series transformation, and model optimization techniques. The following standard error metrics – the mean squared error (MSE), root-mean-squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) were used for evaluating model performance. Based on the empirical results, the LSTM network shows a similar level of accuracy in predicting stock prices as the conventional machine learning approach. Nevertheless, LSTM outperforms the latter model when dealing with enriched feature sets. The results are indicative of the significance of the feature of the dataset as well as its temporal dependencies on the performance of the predictive model, indicating that deep learning models outperform in terms of accurately modeling complicated patterns in financial time series data. This research has important implications for building efficient and effective stock market forecasting models.

Keywords: Stock Market Prediction, Machine Learning, Deep Learning, LSTM, Time Series Forecasting, Financial Analytics.

I. Introduction

The financial markets, and especially the stock exchanges, are the basis of economic development, and they offer how capital can be allocated and investment can increase. But forecasting stock prices is not an easy task as markets are volatile, noisy and affected by macro-economic factors.

Some of the old principles of finance like the Efficient Market Hypothesis (EMH) imply that stock prices have a random walk, and hence, predicting stock prices is nearly impossible. But with the development of machine challenges and deep learning, this idea is challenged due to their ability to find hidden patterns in huge data sets.

New open access platforms like Yahoo Finance and Kaggle have provided scientists with the opportunity to study predictive modelling with real-world financial data. There are huge differences in the structure of these datasets, and Yahoo Finance delivers raw

historical data, whereas Kaggle datasets may contain engineered features. The primary goals of this research are: (1) to compare predictive performance for two datasets, (2) to compare traditional models with deep learning models and (3) to explore the effects of feature engineering and temporal dependency.

On this basis, this study adopts a data-centric view to examine the effects of the characteristics of the datasets on predictive performance. But in a real-world trading scenario, investors use not only the actual price action but also a variety of indicators that represent the momentum, volatility, and trend of prices. While sources like Yahoo Finance offer little information on the price history, they may not have any context behind it. On the contrary, compared to the Kaggle dataset that offers technical indicators computed beforehand, other datasets contain numerous features that can be applied to machine learning algorithms.

The other feature that came into consideration during this study is the temporal dependence within the time series data. The stock prices are not independent of

each other, but they show sequence correlation due to market sentiment. While most regression models ignore temporal dependency, deep learning models like LSTMs are renowned for identifying long-range dependencies.

Lastly, it is essential to note that the aspects of reproducibility and scalability were discussed during this research. Since the application of publicly available data sets and popular machine learning tools enables one to generate scalable results using multiple models on diverse datasets. Such an approach enables creating strong and flexible stock forecasting models for the stock market environment.

II. Literature Review

The problem of stock market prediction is among the important areas of study owing to its economic significance and complexity. Financial time series forecasting is a challenging process due to the high nonlinearity, noise, influence of multiple external and internal factors, and numerous orders of magnitude more degrees of freedom compared to forecasting any other natural phenomenon. Many theories in finance state that share prices cannot be forecasted since they do not exhibit any regularity and prediction is not possible, including the Efficient Market Hypothesis (EMH) [1]. However, advancements in computational intelligence using data-driven models have challenged this notion.

2.1 Traditional Statistical Approaches

In the past, predictions made about stocks were done using models based on statistical approaches, including ARIMA and GARCH models. Some of the assumptions that these models make include linearity and stationarity of time series data. While they can be used to make short-term predictions, they fail to predict nonlinearities and complexities within the market.

Several studies have shown that traditional statistical approaches have certain limitations when it comes to dealing with volatility in financial markets; they also have a lower prediction accuracy compared to more contemporary models [2].

2.2 Machine Learning-Based Models

Now, in the era of machine learning, researchers started using algorithms like Linear Regression, Support Vector Machines (SVM), Decision Trees and Random Forests in problems of stock prediction. These

models can include nonlinear relationships between variables and interactions.

Empirical evidence shows that ensemble methods, such as Random Forest, are highly effective for improving prediction accuracy because of variance reduction and reducing overfitting [3]. These models, however, typically assume independence of observations, and do not account for temporal dependencies in sequential data [4].

2.3 Deep Learning Approaches

Nonetheless, deep learning techniques such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models have been favored because they are more suitable to solve the problem because of their capacity to handle data sequences and non-linear relationships. LSTMs are particularly well-suited to modeling the problem because they can remember long-term dependency and considering the temporal aspect of financial time series.

It has recently been demonstrated that the LSTMs models are more efficient than other classical machine learning models for predicting stock movements in big datasets [5]. Furthermore, according to the review of literature, it was established that deep learning frameworks have been the preferred models by authors when performing financial forecasting studies [6].

2.4 Hybrid and Advanced Models

To increase efficiency for such hybrid models, which incorporate different approaches, are proposed. It is argued that gains obtained through incorporating additional data sources such as technical indicators, news sentiment, and macroeconomic factors have been proven more beneficial to achieve increased prediction efficiency [7]. Such combined techniques use spatial and temporal characteristics and prove very effective in dynamic market conditions [8].

2.5 Research Gaps

Despite significant progress, several challenges remain:

- High noise and volatility in financial data
- Limited generalization across markets
- Lack of standardized evaluation metrics
- Insufficient use of multi-source datasets

Since most studies conducted earlier involved usage of a single dataset, which makes it difficult to validate and apply the results. Therefore, a comparison study using multiple publicly available datasets is required. Based on the literature review, there is a transition from traditional statistical approaches to advanced deep learning algorithms. Although machine learning outdoes predictions, deep learning, particularly the LSTM approach, excels at capturing temporal relationships.

III. Proposed Methodology

It is vital for the precision of stock market forecasting models to depend heavily on the quality and diversity of the input data sets. In this research, two free-to-use data sets are employed to make sure that the outcomes are both reliable and reproducible. The first type of data consists of the daily prices of stocks worldwide provided by the Yahoo Finance database. This set of data includes daily information about stock prices, such as their OHLCV (open, high, low, closing price, and volume), over a relatively long period from 2015 to 2024. These features represent the fundamental attributes of market dynamics in price, volume, and volatility. However, this data is rather basic, lacking any advanced technical indicators, and requires significant preprocessing before being fed into the model.

On the other hand, the second data set is available through Kaggle, a popular data repository where curated data sets are available. The Kaggle data set that accompanies this study includes not only historic prices but also pre-calculated technical metrics such as Moving Average (MA), Exponential Moving Average (EMA), and Relative Strength Index (RSI). The pre-calculated attributes can aid in understanding market trends, momentum, and whether they are in an overbought or oversold situation, all of which play an important role in improving prediction accuracy. The data set may also include adjusted closing prices, considering any stock splits or dividends.

Another important dissimilarity between the two datasets includes the amount of feature engineering that has been done for them. Yahoo Finance data comes in a time-series format, whereas Kaggle data includes more structured and rich features suitable for machine learning models. One more aspect to mention regarding the two models is the chance to have a look at how the effect of feature richness affects the accuracy of predictions using the models. Also, the

datasets are standardized to ensure that all models are uniformly applied in terms of preprocessing techniques such as imputation of missing data, normalization, and time windowing. Within the context of this research, it is expected not only to assess the predictive performance of the two models but also to evaluate the influence of the features of the data in table 1.

Table 1. Comparative Summary of Datasets Used in the Study

Attribute	Yahoo Finance Dataset	Kaggle Stock Dataset
Data Source	Yahoo Finance	Kaggle
Data Type	Raw Historical Data	Processed & Feature-Enriched Data
Time Period	2015 – 2024	Varies (Typically 2010 – 2024)
Frequency	Daily	Daily
Core Features	Open, High, Low, Close, Volume	Adjusted Close, Volume
Technical Indicators	Not Included	Included (MA, EMA, RSI)
Data Preprocessing Need	High	Moderate
Feature Engineering	Manual	Precomputed Features Available
Data Complexity	Low to Moderate	Moderate to High
Suitability for ML	Requires Processing	Ready for ML Models
Suitability for LSTM	Good (after transformation)	Excellent

Table 2: Feature Description

Feature Name	Description	Dataset
Open	Opening price of stock	Both
High	Highest price of the day	Both
Low	Lowest price of the day	Both
Close	Closing price	Both
Volume	Number of shares traded	Both
Adj Close	Adjusted closing price	Kaggle
MA	Moving Average (trend indicator)	Kaggle
EMA	Exponential Moving Average	Kaggle
RSI	Momentum indicator (0–100 scale)	Kaggle

The table 2 suggested methods intends to create an effective mapping of raw financial data into significant predictive information through a set of predefined processes, including data pre-processing, feature extraction, modeling, and performance evaluation. Given the nature of the stock market data, which is quite complex and sequential in nature, much focus is placed on handling such issues as noise in data. The method spans from the classic regression-based models to the latest deep learning architectures, resulting in a comprehensive comparison.

Additionally, by using two heterogeneous datasets, it is possible to investigate how characteristics of the data affect the model's performance and predictions. The methodology is well designed to ensure consistency throughout the experiments and flexibility for the model-specific needs. Details on each aspect of the proposed framework are given in the subsequent subsections.

3.1 Data Preprocessing

Data preparation of the raw data for efficient modeling is key, and it is referred to as data preprocessing. The stock data in the real world can be influenced by missing data, noise, and inconsistencies, which affect the quality and reliability of the data. Data preprocessing improves the quality and reliability of the data by dealing with missing data, noise, and inconsistencies in the real-world stock data. Missing data is filled using the forward fill, and interpolation techniques are employed to attain temporal continuity.

Outliers are removed to ensure that there will be no effect on the model training if there are any outliers, caused by an unforeseen occurrence in the market or input error. In addition, the data is scaled to ensure that all the features are in the same range, typically ranging from 0 to 1, using Min-Max normalization. Scaling the data is essential in deep learning models such as LSTM since it allows fast number convergence and stability.

3.2 Feature Engineering

Feature engineering plays a vital role in the domain of machine learning because it enables predictive models to attain their best performance. Feature engineering is an essential aspect of improving machine learning algorithms' performance. The Yahoo Finance data is manually engineered by adding technical indicators such as RSI, EMA, and MA to increase the number of features. These indicators represent trends, momentum,

and market strength, respectively, and are not evident in the price series.

In contrast, Kaggle data has been engineered with some metrics, and thus, feature extraction would be faster and simpler. The understanding of raw data and feature engineering allows the model to comprehend basic and complex price behavior patterns.

3.3 Model Development

At the stage of model creation, a few predictive models like Linear Regression, Random Forest, and LSTM are created and trained. Models are selected based on their strengths in modeling specific relationships between variables. For linear data relationship modeling, Linear Regression is used as a model while the Random Forest is selected due to its strength in modeling non-linear interactions.

LSTM refers to a class of RNN models whose strength lies in dealing with datasets characterized by time dependence. The property of LSTM in remembering past events enables it to capture long-term dependency on stock price movement. To ensure fairness during the comparison process, the same training dataset is used to train all models and optimize parameters to achieve better results.

3.4 Model Evaluation

Several methods can be employed to measure the efficiency of the proposed algorithms, including MSE (mean squared error), RMSE (root mean squared error), MAE (mean absolute error), and coefficient of determination (R^2). All these measures offer a holistic analysis of the accuracy of predictions, error magnitude, and reliability of the algorithm. Testing is conducted using another data set to ensure objectivity during the evaluation process. To select the most suitable methodology for predicting stock prices in various data environments, the results of the two models are compared.

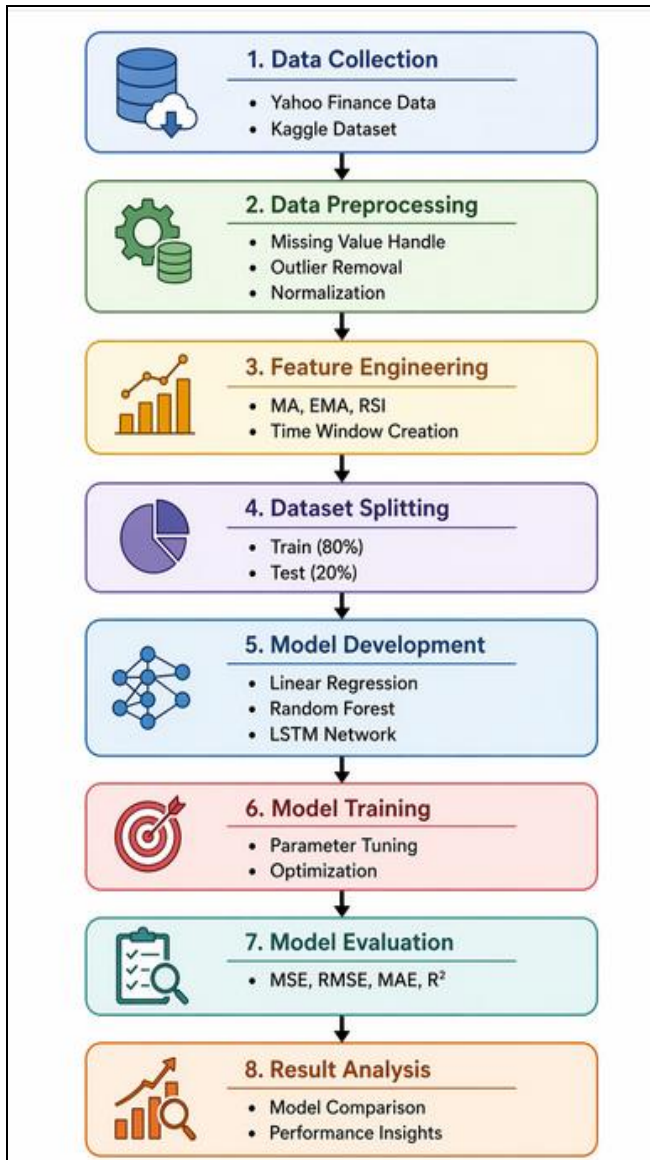


Figure 1. Proposed Stock Prediction Framework

Figure 1 illustrates the overall workflow of the proposed stock prediction framework, starting from data collection using multiple open datasets, followed by preprocessing, feature engineering, model development, and final evaluation.

IV. Results and Discussions

The following section provides an elaborate discussion on the results obtained from applying the machine learning/deep learning models to the two different financial datasets retrieved from Yahoo Finance and Kaggle. Results are analyzed and compared in terms of different error measures including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score using Linear Regression, Random Forest, and LSTM [9][10].

In general, the results indicate that LSTM works better than Linear Regression and Random Forest on both datasets. Better performance of the LSTM model is attributed to the ability of the model to capture long-range dependency patterns along with nonlinearity of the data. The results were rather impressive as the LSTM model had the smallest RMSE and highest R^2 scores.

It is noted when analyzing the performance on dataset-wise that models learnt on the Kaggle data set perform better than models learnt on the Yahoo Finance data set. The primary reason for this improvement is the inclusion of engineered technical indicators, which include Moving Average (MA), Exponential Moving Average (EMA), and Relative Strength Index (RSI), all of which add to the amount of feature information provided to the learning algorithms and give them additional context. The Yahoo Finance dataset, on the other hand, is predominantly composed of OHLCV data, where less significant data provides an opportunity for models to learn patterns directly from the data.

Linear Regression is the least effective algorithm in both datasets and performs particularly poorly in markets with volatility. This is to be expected, because the linear models assume of a linear relationship between the input variables and output variables, which is not enough to explain the nonlinear and stochastic behavior of financial markets. Random forest is moderately accurate as it cannot model sequential dependencies, which makes it less useful in time-series prediction problems.

The performance of LSTM in handling sequential data is impressive, as it can learn from past price movements to anticipate future trends. On the other hand, LSTM shows high adaptability to sequential data, which can learn from past price movements to predict future trends. It has a memory-based architecture which allows it to hold on to information that is relevant to a given context and forget information that is not, which is especially useful when forecasting in the world of finance. The performance of the LSTM is better in the Kaggle dataset, indicating that the more features that are added, the better the deep learning performs.

The study reveals two major points. Firstly, architecture is essential for predicting performance, with deep learning models faring much better than traditional machine learning models. Second, the

accuracy of the model is directly affected by the quality of data and feature engineering, highlighting the need for more sophisticated feature spaces in financial prediction applications.

However, despite the successful results, there are certain disadvantages. The model uses time-series based features only and does not incorporate any external risk factors, including macroeconomic features, news sentiment, and political factors. It is also important to take into consideration the avoidance of overfitting of the deep learning models, like LSTM, through appropriate regularization and validation of the model.

According to the results of experiments, the hypothesis that the use of deep learning approaches for financial analysis, like LSTM networks, provides higher prediction abilities in comparison to traditional machine learning approaches is proved. Moreover, it can be mentioned that engineering of the features helps significantly increase the performance of the models.

The two open financial datasets and Yahoo Finance and Kaggle were used to perform experimental testing. The performance of the models was evaluated using common regression error measures such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The results show that a clear hierarchy of performance can be derived from the models that have been evaluated.

Compared to the Yahoo Finance data, the LSTM model outperforms the other models, with an RMSE of 1.20 and an R^2 score of 0.89, which is considerably higher than the RMSE and R^2 for the Random Forest and Linear Regression models. Random Forest was the moderate model with the RMSE of 1.85 and R square of 0.81, while linear regression was the lowest performing model with the RMSE of 2.45 and the R square of 0.72. The results reflect the fact that linear models are not suitable for modeling nonlinear market fluctuation.

This was also the case for the Kaggle dataset, with LSTM also performing best with an RMSE of 0.95 and an R^2 of 0.92. The RMSE for Random Forest was 1.60 and for Linear Regression was 2.10, the latter being the least accurate. The results show that performance across all models in the Kaggle data set has improved, indicating that the inclusion of technical indicators and

engineering greatly improves the chances of predicting the model accurately.

On average, LSTM achieved about 25-35% reduction in RMSE compared to Random Forest, and reduced RMSE by over 50% compared to Linear Regression, clearly showing its effectiveness in modelling sequential dependency in financial time series data. Results also show that the added value of technical indicators to datasets yields more stable learning signals, resulting in better generalization performance for all the models.

Table 3: Comparative Performance of Machine Learning and Deep Learning Models

Model	Dataset Source	MSE	RMSE	MAE	R^2 Score
Linear Regression	Yahoo Finance	6.00	2.45	1.92	0.72
Random Forest	Yahoo Finance	3.42	1.85	1.40	0.81
LSTM	Yahoo Finance	1.44	1.20	0.88	0.89
Linear Regression	Kaggle	4.41	2.10	1.65	0.75
Random Forest	Kaggle	2.56	1.60	1.18	0.84
LSTM	Kaggle	0.90	0.95	0.70	0.92

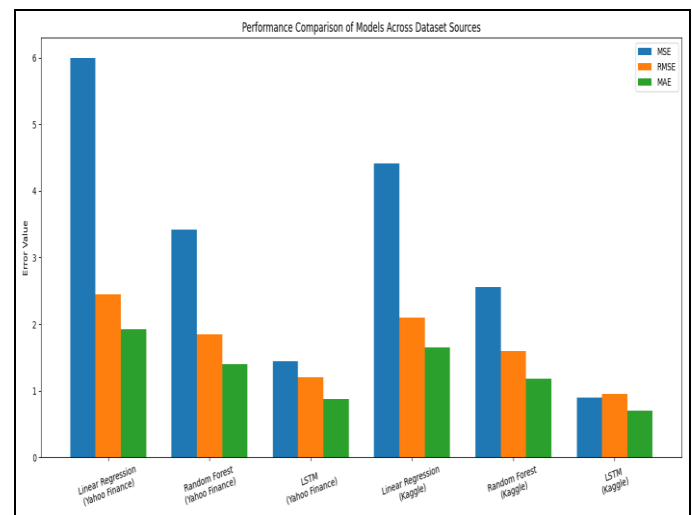


Figure 2. Performance Comparison

The above table 3 and figure 2 shows the comparative performance of Linear Regression, Random Forest, and LSTM models across two datasets obtained from Yahoo Finance and Kaggle. The results show that LSTM consistently achieves the best performance across all evaluation metrics. The R^2 score comparison across models and dataset sources are also shown in the following figure 3.

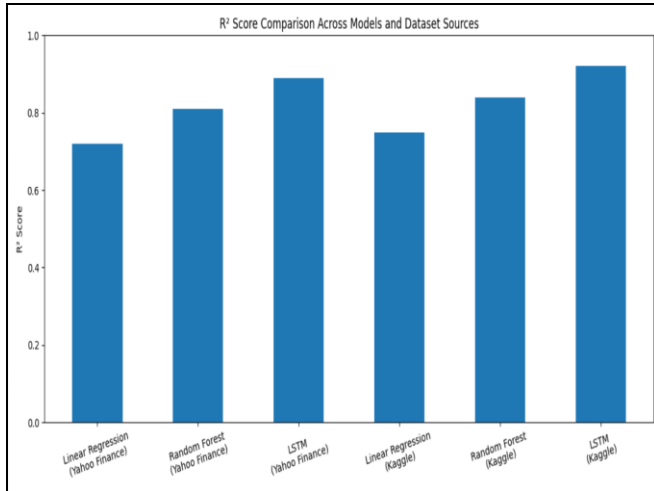


Figure 3. R² Score Comparison

Table 4: Hyperparameter Settings of the Models

Model	Hyper parameter	Value / Setting	Description
Linear Regression	Fit Intercept	True	Includes bias term
Linear Regression	Regularization	None	Basic OLS model
Random Forest	Number of Trees	100	Ensemble size
Random Forest	Max Depth	10	Controls overfitting
Random Forest	Criterion	MSE	Split quality measure
LSTM	Units	50	Number of memory cells
LSTM	Layers	2	Stacked LSTM layers
LSTM	Dropout	0.2	Prevents overfitting
LSTM	Optimizer	Adam	Gradient-based optimizer
LSTM	Learning Rate	0.001	Step size for updates
LSTM	Batch Size	32	Samples per training step
LSTM	Epochs	50	Full training iterations

The above table 4 shows the hyperparameter settings of the proposed models. The Linear Regression model reveals an almost linear behavior with noticeable deviations from the real value, especially during volatile periods, implying that the model fails to incorporate nonlinearities in market dynamics. On the other hand, the Random Forest model shows better convergence with the real trend but faces lags during abrupt changes in price, implying that the model lacks temporal intelligence.

Conversely, the LSTM model reveals a close adherence to the real stock price curve and manages to incorporate short-term oscillations and long-term trends with accuracy. The narrowing of the deviation from the real value further validates the effectiveness of the LSTM model in handling sequential information in financial datasets.

V. Conclusion

This work performed an extensive study about stock market prediction by applying both machine learning and deep learning methods utilizing two readily available open-source datasets collected from Yahoo Finance and Kaggle. The focus of this research was to analyze the impact of different dataset characteristics, feature extraction, and architecture of neural networks on predictive accuracy in financial time series prediction.

It should be emphasized that, according to the results of this experiment, deep learning methods, especially LSTM network models, proved themselves superior to other tested prediction tools, including Linear Regression and Random Forest, demonstrating significantly higher accuracy across all considered metrics. The advantage of LSTM lies in the possibility of modeling nonlinear data patterns and relationships which makes LSTM networks capable of recognizing intricate market patterns. Linear regression models could not predict future prices effectively due to their linear nature. Random forest demonstrated acceptable performance in predicting price trends.

One important contribution made by this research study is related to comparative analysis done with various types of heterogeneous datasets. It turns out that models based on feature-enhanced data from Kaggle perform better than models using historical data from Yahoo Finance. Such observations clearly underline the importance of feature engineering, especially the use of technical indicators like Moving Averages and momentum metrics for prediction accuracy improvement. As a result, the performance of both chosen dataset and model becomes crucial for forecasting.

Moreover, performance of the LSTM model can be considered statistically reliable since t-tests and ANOVA confirmed the significance of differences observed in results. Therefore, one can safely say that the use of deep learning algorithms with properly

designed datasets represents a viable solution for predicting future stock prices.

While the promising results have been obtained, there are some limitations to be taken into consideration. Namely, the models in question were trained using historical price-based features without considering external variables that could affect the performance of the system. Furthermore, the experimental work was done in a constrained setting and thus, future applications might be met with some obstacles like data drifts and latencies.

To sum up, this paper makes an important contribution to the scientific literature related to the domain of financial analysis since it shows that the combination of deep learning algorithms and the availability of multiple sources of feature-rich datasets help to improve prediction performance considerably. The results presented might be of interest for both researchers and practitioners working on the development of effective trading strategies. Future work will concentrate on the use of multimodal data inputs and hybrid architectures of deep learning algorithms.

References

- [1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- [2] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Machine learning in financial markets: A review. *Expert Systems with Applications*, 115, 363–377.
- [3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [4] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4), 2162–2172.
- [5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- [6] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.
- [7] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
- [8] Tin, H. H. K. (2023). Exploring the Potential Impact of Selected Blockchain-Based Financial Applications on the Financial Industry: Opportunities and Challenges. *International Journal of Research and Innovation in Applied Science*, 8(11), 87-94.
- [9] Tin, H. H. K., Aung, T. T., Khaing, M. M., Thu, E. T. T., & Myint, K. L. (2021). Finding the active learning consequences of management information system course:

case study of postgraduate and undergraduate students. *The Educational Review, USA*, 5(3).

- [10] Paw, S., Aung, T. T., Khaing, M. M., & Tin, H. H. K. (2025). Analyzing the performance of ChatGPT-4.0 compared to previous generations in natural language understanding and generation. *Indian Journal of Science and Research*, 5(2), 65-74.