A Proposal for the Final Project

A Hybrid Deep Learning Approach to Multivariate and Multiple-Step-Ahead Forecasting of Sector-Specific Stock Market Trends

Submitted by:

Hou Nam Ip (aka Ip Hou Nam Davis)

Submission Date: 26-June-2023
Word Count:2,529 (Excluding References)

Submitted in partial fulfilment of the requirements for the Final Project (DSM500-2023-APR), as part of the MSc Data Science and Artificial Intelligence programme, University of London

1. Introduction

The financial market is a complex web of interrelated variables such as macroeconomic indicators, geopolitical risks, oil prices and even health policies (Sharif, Aloui, & Yarovaya, 2020). To understand this complexity, sector-specific stock market forecasts are essential, as they not only provide important insights into investment strategies, but also reflect the economic health of key sectors. For example, the energy, IT and real estate sectors, each of which are key economic pillars, show different dynamics and different responses to economic indicators.

Considerable efforts have been made to understand the impact of past stock market trends, anticipate developments, and uncover the correlations between stock market trends and macroeconomic indicators for business interests and policy decisions. While various AI-based forecasting models have been developed, much of the potential for forecasting sector-specific stock market trends using Deep Learning (DL) models remains untapped (Ozbayoglu, Gudelek, & Sezer, 2020).

This project proposes a novel approach using a hybrid DL model for multivariate and multiple-step-ahead forecasting to improve the accuracy of sector-specific stock market forecasts. The hybrid model will combine the capabilities of the following three network architectures:

A. Convolutional Neural Networks (CNNs)
B. Recurrent Neural Networks (RNNs)
C. Long Short-Term Memory (LSTM) networks

The need of a complex model stems from the limitations of conventional forecasting models, which are often unable to capture the complex relationships in financial markets.

In the context of this project, the term 'stock market trends; refers to the following three aspects:

A. Price, as indicated by the daily adjusted closing prices
B. Volatility, measured through the daily high-low price range
C. Trading volume, as reflected by the daily trading volumes

of each of the three selected large-cap index funds traded on the US stock exchange:

A. S&P 500 Information Technology Index (VGT), as an indicator of the IT sector
B. S&P 500 Energy Index (VDE), as an indicator of the energy sector
C. S&P 500 Real Estate Index (VNQ), as an indicator of the real estate sector

This project also seeks to explore the capability of forecasting stock market trends by employing four crucial macroeconomic indicators from the US:

A. Interest Rates
B. Gross Domestic Product (GDP) Growth Rates
C. Inflation Rates
D. Unemployment Rates

2. Objectives & Research Questions

This project has the following five objectives:

A. Collect historical data for the selected three sectors and identify relevant macroeconomic indicators.
B. Develop and implement the hybrid DL model, which combines the strengths of CNNs, RNNs and LSTMs for multivariate forecasting with multiple steps in advance.
C. Evaluate the forecasting power of the four selected macroeconomic indicators using Granger causality tests.
D. Compare the performance of the hybrid model with stand-alone CNN, RNN and LSTM models.
E. Interpret the results and provide insights into sector-specific stock market trends.

Based on these objectives, the project aims to answer the following research questions:

A. Can a hybrid DL model outperform stand-alone CNN, typical RNN and LSTM models in forecasting sector-specific stock market trends?
B. Do the forecasting accuracy and model performance differ among the three selected sectors?
C. Can the hybrid DL model effectively capture and forecast market volatility in these three sectors?

3. Literature Review

A. Application of Deep Learning Techniques in Financial Forecasting

The application of DL models in financial forecasting has seen a remarkable upsurge. The study by Sahu, Mokhade, and Bokde (2023) highlights the potential of DL techniques, such as CNNs, RNNs and LSTM networks, in capturing various types of data patterns for financial forecasting.

Meanwhile, Torres et al. (2021) provide an overview of how CNNs, originally designed for image recognition tasks, have been applied to time-series data, showcasing their capability in identifying temporal dependencies and complex patterns. More recent studies have started to investigate the implementation of CNNs in stock market forecasting and show promising results (Shah et al., 2022).

In parallel, RNNs have gained prominence due to their ability to manage sequential data and capture the temporal dependencies of a series (Hewamalage et al., 2021). This property is particularly important for stock market forecasting, where future values depend heavily on the past. However, RNNs often suffer from vanishing or exploding gradients during training, which poses a challenge to their implementation (Surakhi et al., 2020). To address this problem, LSTM networks, a variant of RNNs, have been developed. They are equipped with a unique architecture for learning long-term dependencies, which solves the problems of conventional RNNs and shows strong performance in forecasting time series (Sak et al., 2014).

Over the past decade, the concept of hybrid models has also been extensively introduced in forecasting stock prices, which combine different types of networks to exploit the strengths of each model (Kim & Won, 2018).

Regardless of the types of DL models, multiple-step-ahead forecasting is a crucial aspect in designing such models for stock market forecasts due to the inherent volatility of financial markets. Models typically forecast at various horizons, including 1-, 5-, 10-, 15-, 20-, and 30-day-ahead prices, with shorter forecasts targeting immediate market reactions and longer ones capturing broader market trends influenced by macroeconomic factors (Htun, Biehl, & Petkov, 2023).

B. Importance of Sector-Specific Analysis in Stock Market Forecasting

There have been in recent years many hybrid models for forecasting stock market trends, such as the one by Xiao et al. (2019) on forecasting the daily return direction of the stock market using hybrid machine learning algorithms. Their project forecasts the daily return direction of the SPDR ETF tracking the S&P 500, which is regarded as 'the best single gauge of large-cap U.S. equities' (S&P Dow Jones Indices, n.d.).

However, it remains to be investigated whether a sector-specific model would be more effective. It is this project's assumption that considering sector-specific stock market trends is critical to developing effective forecasting models. In their study about Chinese stock prices across different sectors during the COVID-19 pandemic, He et al. (2020) found that different sectors exhibit unique trends that are influenced by a variety of factors. This study serves as an important reference in this project's forecasting of the energy, IT, and real estate sectors' stock prices.

Research has also shown that sector-specific variables significantly affect stock prices in each sector. In Shah et al.'s (2018) study, they concluded that oil prices and renewable energy policies can strongly influence the energy sector, while interest rates and inflation influence trends in the real estate sector. Similarly, Caner et al.'s (2018) investigation in the US biopharmaceutical industry found that the development of the biopharmaceutical industry sector could be influenced by technological breakthroughs, patent approvals, and regulatory changes in the industry.

Comparative studies, such as the one done by Nabipour et al. (2020), which compares nine machine learning models and two DL methods (RNN and LSTM) using ten years of historical data, have found that the performance of forecasting models varies across sectors. This means that a model that is good at forecasting trends in one sector may not perform the same in another.

C. Forecasting Power of Economic Indicators in Stock Market Forecasting

Macroeconomic indicators such as the four selected for this project (interest rates, GDP growth rates, inflation rates, and unemployment rates) play a crucial role in forecasting stock market trends. These indicators, through their **absolute values** and **variations**, reflect the state of the economy, influence investor sentiment, and thus affect stock prices (Ma et al., 2022).

Existing literature has also shown that changes in these economic indicators have significant forecasting power for stock returns. Moreover, Huang et al. (2020) used Google Trends search data as an indicator of investor attention and found it to be a strong predictor of the directional movements in the S&P 500. Also in the same project, statistical techniques such as Granger causality tests were used to assess the forecasting power of various economic indicators. As pointed out by Croux and Reusens (2013), such tests have long been used to examine whether changes in one variable can help forecast changes in another variable.

4. Data

This project will rely on two primary data sources:

A. Vanguard index fund data (from the API of Alpha Vantage)
B. Macroeconomic indicators (from the API of Federal Reserve Economic Data (FRED)).

Phase	Task	Details
A: Processing Vanguard Index Funds Data	1.Data Extraction	Extract: - **Daily adjusted closing prices** - **Daily high-low price range** - **Daily trading volumes** for the Vanguard index funds tracking: - S&P 500 **Information Technology Index** (VGT) - S&P 500 **Energy Index** (VDE) - S&P 500 **Real Estate Index** (VNQ) from the Alpha Vantage API, spanning a 10-year period (Jan 2012 - Dec 2021).
	2.Data Cleaning & Preprocessing	Clean and preprocess the raw data to ensure accuracy and relevance for subsequent analysis.
	3.Data Export	Save the preprocessed data into CSV files.
B: Processing Macroeconomic Indicators	1.Data Extraction	Extract the absolute values of: - **Monthly interest rates,** - **Monthly GDP growth rates,** - **Monthly inflation rates** - **Monthly unemployment rate** from the Federal Reserve Economic Data (FRED) API, spanning the same 10-year period.
	2.Data Export	Save the collected data in CSV format.
	3.Data Frequency Harmonization	Perform linear interpolation using the pandas library in Python to estimate the missing daily values of these monthly indicators
C: Further Data Preparation	Create Training and Test Sets	Split the dataset into a temporary set (85% of the data) and a test set (15% of the data)
	Create Validation Set	Further split the temporary set into a training set (70% of the original data) and a validation set (15% of the original data)

5. Methodology

A. Software & Hardware

The project will use Visual Studio Code and Python 3.11.1. The hardware used will be an iMac (24-inch, M1, 2021) running macOS Big Sur version 11.5.1. The iMac has 8GB of unified memory.

This project involves a combination of neural network techniques and statistical analysis to forecast sector-specific stock market trends using a number of Python libraries, including:

Library	Purposes
Pandas	Handling time series data, linear interpolation
NumPy	Array operations
Matplotlib	Data visualisation
Scikit-learn	Using metrics like MAE and MSE, and conducting model comparison
TensorFlow	Creation and training of the hybrid DL model and the standalone models

B. Steps to Build and Test the Models

The core of this project is the development and implementation of the hybrid DL model, which integrates CNNs, RNNs and LSTM networks to perform multiple-step-ahead forecasting.

The CNN layers of the model first will process the input data. LSTM layers will then be implemented to capture long-term dependencies in the data. This takes advantage of the LSTMs' ability to selectively remember and forget information over longer sequences.

Once the model architecture is established, the model will be trained with historical stock market data and the four macroeconomic indicators for the three selected sectors. This project aims to produce forecasts in several steps for both the next 15 days and the next 30 days. These are common forecast horizons in financial studies for daily data, and they allow for the capture of bi-monthly and monthly trends.

In the training process, the weights of the model will be iteratively adjusted to minimise a loss function that measures the difference between the model's forecasts and the actual observed data.

The performance of the model will be evaluated using the mean absolute error (MAE) and the root mean square error (RMSE), which are measures of the model's forecasting accuracy.

In addition, the performance of the hybrid model will be compared to that of single architecture models to determine if it provides better forecasting performance.

Finally, Granger causality tests, a statistical technique, will be used to find out whether changes in the macroeconomic indicators can forecast shifts in stock volumes, volatility and adjusted closing prices.

6. Ethical Considerations

The data used in this project comes exclusively from publicly available datasets and does not contain any personal, private or sensitive information. Therefore, there are no obvious ethical concerns related to the data.

7. Anticipated Outcomes

The project envisages three outcomes:

A. **A novel hybrid DL model for forecasting sector-specific stock market trends**. This model will integrate the capabilities of CNNs, RNNs, and LSTMs to capture both local and long-term dependencies in stock market data. If successful, this model could potentially outperform single-architecture models.

B. **Insights into whether various macroeconomic indicators for sector-specific stock market trends possess forecasting power over stock market trends**. This will be assessed by examining whether considering macroeconomic indicators can enhance the forecasting power of the hybrid DL model. By using Granger causality tests, the project aims to uncover potential causal relationships between these economic indicators and stock market performance.

C. **A better understanding of the differences in forecasting accuracy and model performance across the three sectors.** By comparing the model's forecasting performance in the energy, IT, and real estate sectors, the project aims to uncover sector-specific nuances and trends that could enhance the forecasting power of future models. By examining two forecast horizons (15 days and 30 days), this project also aims to assess the consistency and robustness of the model across different forecast intervals.

8. Anticipated Challenges & Limitations

Some anticipated challenges this project will encounter are as follows:

Challenge	**Details**
Data Availability and Quality	Data quality issues, missing values and inconsistencies could affect the performance of the hybrid and standalone models
Model Complexity	In the absence of a more powerful computer cluster, DL models can be time-consuming and resource-intensive to build and train.
Model Performance	The inherent volatility and unpredictability of stock markets can make accurate forecasts difficult.
Overfitting	DL models may overfit training data and perform poorly on unseen data.
Time Constraints	The timeframe of the project may limit the depth of investigation and optimisation possible.
Technical Difficulties	Unforeseen technical problems or software glitches

Moreover, significant changes in economic conditions during the ten-year period under study, triggered by major events such as the Covid-19 pandemic or geopolitical developments such as the war in Ukraine, could affect the fit or accuracy of the same time series model over time (Franses, Dijk, & Opschoor, 2014). Given the complexity of this challenge, it falls outside the scope of this project.

This project assumes that the four selected US macroeconomic indicators have a correlation with the US stock market and the three selected index funds, but it is important to note the inherent limitation of this assumption. In reality, the US stock market is a globalised and diverse entity that includes many companies that are not based in the US. Therefore, relying solely on US macroeconomic indicators cannot fully capture the complex nature of stock market trends.

9. Timeline

The project plan consists of nine key phases.

Week	Start Date	Task
11-12	26th June	Building further literature review based on the one in the project proposal
13-14	10th July	Extract data from the selected API
15-16	24th July	Data cleaning and exploratory data analysis
17-18	7th August	Data preprocessing including performing linear interpolation
19-20	21st August	Writing codes to build the Deep Learning (DL) models
21-22	4th September	Writing codes to run the DL models
23	18th September	Comparing the performance of different models on the three different datasets and writing results
24	25th September	Submit the first draft and receive feedback
25	2nd October	Incorporate feedback, revise and submit the final version

References

Caner, T., Bruyaka, O., & Prescott, J. E. (2018). Flow signals: Evidence from patent and alliance portfolios in the US biopharmaceutical industry. *Journal of Management Studies*, 55(2), 232-264.

Croux, C., & Reusens, P. (2013). Do stock prices contain predictive power for the future economic activity? A Granger causality analysis in the frequency domain. *Journal of Macroeconomics*, 35, 93-103.

Franses, Dijk, D. van, & Opschoor, A. (2014). *Time series models for business and economic forecasting* (Second edition.). Cambridge University Press.

He, P., Sun, Y., Zhang, Y., & Li, T. (2020). COVID–19's impact on stock prices across different sectors—An event study based on the Chinese stock market. *Emerging Markets Finance and Trade*, 56(10), 2198-2212.

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388-427.

Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, *9*(1), 26.

Huang, M. Y., Rojas, R. R., & Convery, P. D. (2020). Forecasting stock market movements using Google Trend searches. *Empirical Economics*, 59, 2821-2839.

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, 103, 25-37.

Ma, F., Lu, X., Liu, J., & Huang, D. (2022). Macroeconomic attention and stock market return predictability. *Journal of International Financial Markets, Institutions and Money*, 79, 101603.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access*, 8, 150199-150212.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, 93, 106384.

S&P Dow Jones Indices. (n.d.). U.S. Core. S&P Dow Jones Indices. Retrieved May 23, 2023, from https://www.spglobal.com/spdji/en/landing/investment-themes/us-core/

Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. *Applied Sciences*, 13(3), 1956.

Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

Shah, I. H., Hiles, C., & Morley, B. (2018). How do oil prices, macroeconomic factors and policies affect the market for renewable energy? *Applied energy*, 215, 87-97.

Shah, J., Vaidya, D., & Shah, M. (2022). A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications*, 200111.

Sharif, A., Aloui, C., & Yarovaya, L. (2020). COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: Fresh evidence from the wavelet-based approach. *International Review of Financial Analysis*, 70, 101496.

Surakhi, O., Serhan, S., & Salah, I. (2020). On the ensemble of recurrent neural network for air pollution forecasting: Issues and challenges. *Advances in Science, Technology and Engineering Systems Journal*. J, 5, 512-526.

Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep learning for time series forecasting: a survey. *Big Data*, 9(1), 3-21.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, 5(1), 1-20.