# Applied Machine Learning Coursework 2

**Chapter 1. Introduction**

* 1. Introduction.

Predicting future stock values and identifying trends may be accomplished using machine learning, a challenging endeavour for humans. This aids algorithmic trading. Various techniques like “supervised learning”, “unsupervised learning”, “reinforcement learning”, and time series analysis are utilised to create trading techniques. overfittings benefits, challenges such as market unpredictability and overfitting exist. Traders can mitigate risks by integrating machine learning with financial knowledge, back testing, and risk management for adaptable, profitable strategies.

* 1. Main goals of the project.

Technological advancements in global digitization have revolutionized stock market prediction, reshaping traditional trading methods. Increased market capitalisation has made stock trading a key investment avenue for financial investors. Advanced machine learning algorithms have shown their efficacy in forecasting stock market patterns, as substantiated by the analysis of data spanning from 2011 to 2021 [1]. It highlights the challenges posed by dynamic stock market data and offers valuable insights.

The primary goal of this research is to create a lucrative trading strategy utilising machine learning methods to forecast future stock prices using historical daily OHLC (Open, High, Low, Close) price data. This involves utilising machine learning techniques to examine large datasets of historical stock market information, aiming to detect patterns, trends, and potential predictive signals that could impact stock prices [2]. The ultimate goal is to use these insights to design trading strategy that could lead to profitable outcomes in volatile market.

* 1. Summary.

This project consists of five chapters. Chapter one covers a general introduction, outlines the main

objectives, and summarises the roadmap of the study. Chapter two summarises the literature relevant to the study. Chapter three explains the sources of data, how it is cleaned and prepared for analysis, the methods employed to analyse them, and the advantages and disadvantages of those methods. Chapter four represents the results of the project and their analysis. Finally, the findings are summarised and recommendation.

# Chapter 2. Literature Review .

* 1. Introduction.

This review delves into the integration of machine learning technologies with financial market operations, focusing on their role in developing innovative algorithmic trading strategies. The paper examines the use of machine learning methods in analysing and forecasting stock prices, with a focus on their significance in improving the efficiency of these tactics. The review covers various topics, including methodologies, challenges and the importance of case studies and applications in demonstrating the potential of machine learning to refine trading strategies. It concludes by discussing future directions and the potential for technological synergies, aiming to contribute to the debate on merging machine learning with algorithmic trading. The review highlights the gaps in existing literature and proposes areas for further research.

* 1. Machine Learning Algorithms in Financial Markets.

Machine learning revolutionizes financial services by improving decision-making, accuracy, and predictive analytics. Its impact is profound across several key domains, including algorithmic trading, credit scoring, risk management, fraud detection, personalized banking, portfolio management, regulatory compliance, sentiment analysis, and predictive analysis.

2.3.Key Areas of Impact.

The financial sector is undergoing a transformation thanks to machine learning (ML), which is also transforming investment management, credit scoring, fraud detection, personalised banking, algorithmic trading, and regulatory compliance. Machine learning algorithms have the ability to evaluate large amounts of data, quickly spot hidden patterns and trade opportunities, increase the precision of risk assessments, and personalise financial products. Furthermore, machine learning (ML) has greatly increased the efficacy and efficiency of sentiment analysis, financial advice, and fraud detection. The use of ML in finance is anticipated to grow as data volume and technology progress.

According to [2], deep learning can significantly improve financial market predictions. The paper presents a one-dimensional Convolutional Neural Network (CNN) as an effective tool for predicting and detecting trends in financial markets. The model uses a shared-parameter approach to analyse financial trading data across time, including prices and volume, reducing biases. The study analysed historical financial data from 2010 to 2017 for six types of products and found significant enhancements in the extraction of generalizable features and financial performance compared to traditional indicators and existing machine learning methods.

The CNN model shown improved efficacy in detecting crucial characteristics as compared to conventional technical indicators, resulting in a higher mean yearly return and enhanced dependability[2]. However, the study encountered three obstacles: its dependence on historical data may not cover all future market situations or unforeseen developments, and the model's performance could be significantly impacted by data quality and completeness. The study explores the use of one-dimensional CNN models for predicting financial market trends.

The study[1] explores the impact of advanced technologies and digital platforms on stock market forecasting, highlighting the need for innovative prediction models and tools using non-traditional data sources like social media. Machine learning techniques have significantly enhanced predictions in volatile stock markets. The research analyzes market trend data from 2011 to 2021 and identifies three crucial phases in data preparation: extracting features from text, reducing dimensionality, and converting data into a machine learning format. The findings suggest that merging market data with textual information via machine learning algorithms significantly improves stock market forecasts.

The study evaluates stock market predictions using Random Forest and Deep Neural Network techniques, comparing their effectiveness to traditional methods during significant price fluctuations. Results show AR-DNN and AR-RF models outperform ARIMA, aiding investors and policymakers in making informed financial decisions during uncertain times.

The study examined daily closing prices of the KSE-100 index from January 2001 to August 2021, focusing on the initial COVID-19 case in Pakistan. The data was partitioned into separate training and testing sets to facilitate the construction and assessment of the model. Results showed superior accuracy of AR-DNN and AR-RF models over traditional ARIMA models, with AR-DNN showing exceptional accuracy and low error rates. AR-RF showed strengths during the COVID-19 era. However, the study had limitations, including a limited selection of machine learning models and its applicability to Asian markets. Future research should focus on reliance on index data and computational demands of deep learning models.

The study evaluates the accuracy of stock market predictions using advanced approaches like Random Forest and Deep Neural Network, focusing on the COVID-19 pandemic. All studies highlight the challenges associated with market volatility and data complexity, comparing advanced techniques with traditional methods. However, there are significant differences between the studies, with the first focusing on CNN models and the second using textual data from social media.

* 1. Research Gaps.

The studies lack exploration of combining multiple models for improved prediction accuracy, regional specificity, scalability, and efficiency, particularly in Asian markets. They also lack ethical considerations regarding social media data use for financial forecasting, indicating the need for further research. The researcher uses Long Short-Term Memory models.

* 1. Summary of Literature.

The research highlights the potential of advanced machine learning techniques like CNNs, Random Forests, and Deep Neural Networks in forecasting financial market patterns. However, gaps exist in exploring combined models, regional specificity, scalability, and ethical implications.

# Chapter 3: Methodology

This research utilises Long Short-Term Memory (LSTM) networks for stock price forecasting, overcoming limitations of traditional RNNs by effectively managing long-term dependencies.

* 1. Reasons For selecting LSTM Models.

LSTM networks are chosen over Convolutional Neural Networks (CNN) due to their capability to handle sequential data that is chronologically ordered, a common characteristic of stock market data. Unlike CNNs, which excel at identifying spatial hierarchies in data, LSTMs are designed to process time series data by preserving input history through their memory cells. This allows LSTMs to maintain crucial information throughout the training process, thereby addressing issues like the vanishing gradient problem often encountered in standard RNNs.

* 1. Data Collection and Preparation.

We gathered financial information for four significant corporations—Apple, Microsoft, Google, and Amazon—using the Yahoo Finance API, concentrating on indicators like stock prices and trading volumes. Three separate sets of data, spanning the period from January 1, 2018, to March 20, 2024, are identified within the data:

Training Data: Data up to June 30, 2023, used for fitting the LSTM model. Validation Data: Data from July 1, 2023, to December 31, 2023, used for tuning model parameters. Test Data: Data from January 1, 2024, to March 31, 2024, used to evaluate model performance. For visual analysis, we utilised the Matplotlib library to create plots with dates on the horizontal axis and closing prices on the vertical axis. These visuals help illustrate the overall trends and fluctuations in stock prices over the observed period.

* 1. Data Normalisation.

To ensure consistent model training and improve prediction accuracy, data normalisation was conducted using the StandardScaler from the sklearn library. This step standardizes the feature scales, minimising the influence of outlier values. Each data set was reshaped into a two-dimensional array to meet the LSTM model’s input requirements.

* 1. Preparing Predictor and Response Variables.

In this phase, we structured the data to capture temporal relationships crucial for time series forecasting. The data from the current time (t) serves as the predictor, while the data from the next time step (t+1) is used as the response variable. This setup is consistent across training, validation, and test datasets, enabling the LSTM network to learn from past trends to predict future stock prices.

* 1. Data reading and visual display.

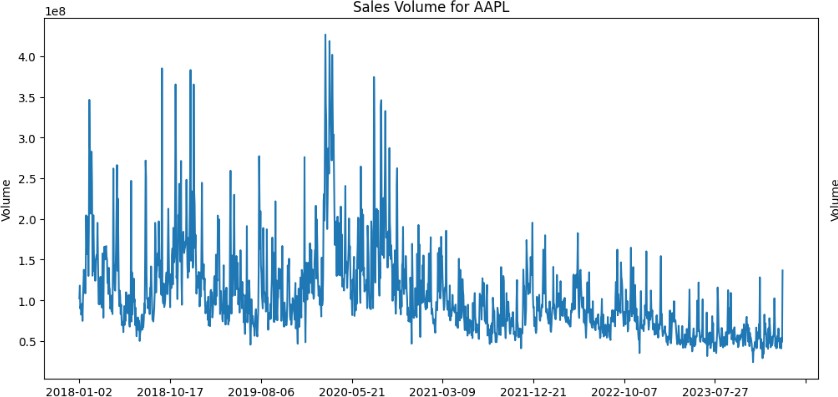
Using Yahoo Finance, we were able to obtain financial information about four different companies, such as stock prices and trading volumes.We divide the data into training, validation, and test sets to ensure the model's performance and dependability in practical applications. This allows us to train, adjust, and assess machine learning models efficiently.

Initially, the U.S. Apple stock index Excel file is divided into time-based sections. These sections include the starting price, maximum price, lowest price, and closing price of the Apple stock index from January 1, 2018, to March 20, 2024. Data for training before June 30, 2023, TrainData.xls， The test data ValidateData.xls from July 1 to December 31, 2023, and the forecast data TestData.xls from January to March 31, 2024.

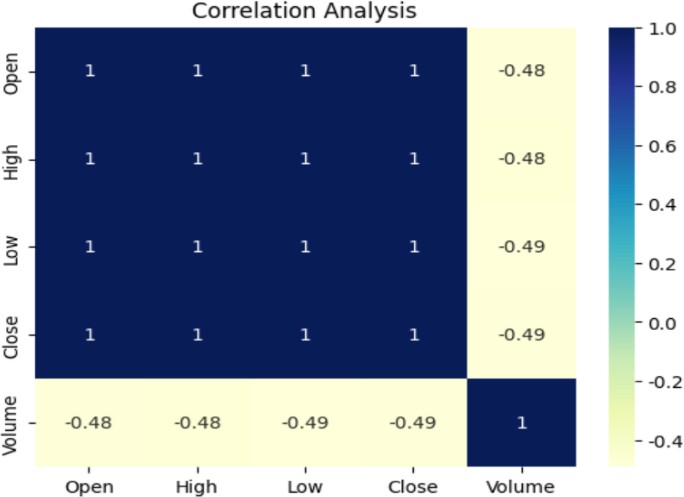
To create a basic visual analysis of the total data of the three files, use the date as the horizontal axis data and the closing price as the vertical axis data. The following are the results of the test set, prediction set and training set:



Based on the operating results, Apple's stock is normally displaying an increasing tendency in its closing price trend, except a few steep declines in the middle. It's easy to see that the closing price of the stock exhibits a varying pattern as you enlarge and examine more closely.



In order to illustrate the trading behaviour of AAPL from January 2018 to about July 2023, a time series plot of stock sales was made. The abscissa shows temporal trends, and the ordinate measures sales, generally represented in stocks. Plotted sales volume varies from 0 to 100 million shares; chart peaks reflect trading activity peaks on particular days. When everything is said and done, the chart's varying pattern offers a visual representation of sales volume over time.



The correlation between several factors in stock market data is represented by this correlation matrix. A perfect positive correlation is represented by a value of 1, no correlation by a value of 0, and a perfect negative correlation by a value of -1.

“The four variables, "Open", "High", "Low", and "Close", have a perfect positive correlation of 1, indicating a perfect relationship. This is to be expected as these Prices will frequently change concurrently when the market moves during the same trading day.

On the other hand, the correlation between "Volume" and "Open", "High", "Low" and "Close" shows negative numbers (-0.48, -0.48, -0.49 and -0.49), which indicates that the volume There is a moderate negative correlation with these prices. In other words, when price rises, volume may fall and vice versa. This could indicate lower trading volume during larger price changes, or less price movement on days with higher trading volumes.

* 1. Modelling.

This model uses the Keras library, a widely used deep learning library that allows researchers and developers to define, train, and deploy a variety of complex neural network models in a relatively simple way. In the process of building the model, first build a Sequential model, which is a linear stacked layer model for processing a series of input data. The first layer of the model is an LSTM layer, which is a kind of recurrent neural network specifically intended for analysing time series data. The first LSTM layer is set to return the output of the entire sequence, which is necessary to stack another LSTM layer. The input shape is (1,1), and the first LSTM layer is set to 128 cells, which means that each time step will have a 128- dimensional output vector. Immediately after, a second LSTM layer is added to the model, set to 64 cells, and it does not return the output of the entire sequence, meaning that the output is only returned at the last time step of the sequence. After that, the model adds a Dropout layer, a regularization technique that reduces the risk of overfitting the model by randomly discarding (i.e., setting to zero) the output of neurons during training. In this model, the Dropout layer is set to randomly drop 20% of the cells each update cycle.

Next, a Dense layer of 25 cells is added to the model, which is a fully connected layer that compresses the information received from the previous LSTM layer into a 25-dimensional representation. Finally, the model ends with a fully connected layer of single cells, Dense (1). This indicates that the final output is a single continuous value.

* 1. Compiling and training model.

After the model framework is built, the command to compile the model is carried out, which uses the 'adam' optimizer and 'mse' (mean square error) as the loss function. This is because the 'adam' optimizer is a highly efficient variant of gradient descent that is particularly suitable for problems with large-scale data or parameters. The mean square error loss function is a common loss function in regression tasks,

which measures the average square value of the difference between the predicted value and the real value of the model and is one of the standards to evaluate the prediction accuracy of the model.

Finally, the model is trained using historical data by model.fit function. This function trains the model for a specified number of iterations on the dataset, called epochs. To avoid overfitting, the model is set to train 100 epochs, and the batch size is 32. train\_X and train\_Y are training data inputs and labels, while validate\_X and validate\_Y are used to validate the data to evaluate the model's performance against data that has not yet been trained.

# Chapter 4. Findings and Analysis.

4.1. Results and Analysis.

The model was assessed using three metrics: mean square error (MSE), mean absolute error (MAE), and R- squared (R2).To prevent overfitting, we compared the model's performance on both training and validation datasets. After model training, we plotted the two loss functions, revealing a consistent loss trend with only slight fluctuations across iterations. This indicates a set of stable parameters that generalize well beyond the training data.

Both the training and validation losses converged to similar values, confirming that the model did not overfit. This stability and low sensitivity to the training dataset specifics suggest good model generalizability.

Subsequently, we made predictions using the model and compared these to the validation targets, calculating the root mean square error (RMSE) to quantify prediction accuracy. Finally, we compared the actual stock data with our predictions. The resulting graphs (refer to Plot 3) show substantial alignment between the predicted and actual values, underscoring the model’s predictive reliability.

# (P1)

A graph of loss and value

Description automatically generated

**(P1)**

A graph of training and validation loss

Description automatically generated

(P2)

A graph showing the difference between the price and the price

Description automatically generated

# (P3)

**Chapter 5. Conclusion, Recommendations, Suggestions for Future Research’ and limitations.**

* 1. Conclusion.

In the selection of stock prediction methodologies, multiple techniques were considered, including moving average, exponential smoothing, Autoregressive Integrated Moving Average (ARIMA), and Random Forests. Following an extensive review of literature and model evaluation, Long Short-Term Memory networks (LSTM) were chosen as the preferred predictive model.

Due to the superior performance of LSTM in handling time series data, particularly in capturing long-term dependencies, it was identified as an ideal choice for analysing stock market data.

The findings suggest that machine learning effectively captures complex data patterns, providing insights into future market trends. Performance metrics like mean square error and R-squared values indicate the model's ability to generalize without overfitting, a common challenge in predictive modelling.

* 1. Recommendations.
     1. Model Enhancement.

Future work should incorporate diverse datasets like macroeconomic indicators and sentiment analysis from news and social media to better understand the factors influencing stock prices.

* + 1. Risk Management Integration.

The integration of risk management strategies with predictive models can enhance trading strategies' resilience to market volatility and unexpected economic events.

* 1. Suggestions for Future Research.
     1. Hybrid Models.

Future research should explore hybrid models that combine machine learning with traditional financial forecasting methods to enhance their robustness and reliability, especially during market anomalies.

* + 1. Real-Time Data Processing.

The potential for real-time data processing capabilities to enhance the trading system's adaptability to real market conditions could be a significant advancement.

* + 1. Cross-Market Analysis.

The study of global market correlations and influences could potentially lead to the creation of more globally aware trading strategies.

* 1. Limitations of the Research.

The primary challenge faced by the researcher was the limited time allocated for completing the research. As a result, primary data could not be collected, nor could multiple models be combined to enhance the study. To mitigate these limitations, secondary data that was relevant and from trusted sources was employed. Additionally, in terms of combining multiple models, efforts were made to address the weaknesses found in models previously utilized in other studies.

# References

1. A. B. Omar, S. Huang, A. A. Salameh, H. Khurram, and M. Fareed, "Stock Market Forecasting Using the Random Forest and Deep Neural Network Models Before and During the COVID-19 Period," Front. Environ. Sci., vol. 25, Jul. 2022 [Online]. Available: https://[www.frontiersin.org/journals/environmental-science.](http://www.frontiersin.org/journals/environmental-science) [Accessed: Apr. 17, 2024].
2. J. Wang, T. Sun, B. Liu, Y. Cao, and D. Wang, "Financial Markets Prediction with Deep Learning," Apr. 2021. [Online]. Available: https://paperswithcode.com/paper/financial-markets-prediction-with-deep. [Accessed: Apr. 17, 2024].
3. N. Rouf, M. B. Malik, T. Arif, S. Sharma, S. Singh, S. Aich, and H.-C. Kim, "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions," Electronics, vol. 10, no. 2717, Nov. 2021. [Online]. Available: https://doi.org/10.3390/electronics10212717. [Accessed: Apr. 17, 2024].