STOCK MARKET DIRECTION PREDICTION

Direction of USD/CAD in stock market for next minute

NAME: SALMA AHMED RADWAN
ID: 2333663

1. Introduction:

Artificial Intelligence (AI) and Machine Learning (ML) are integral to modern life, playing a crucial role in various fields. In medicine, they assist in predicting diseases and detecting tumors. Similarly, in finance, AI is widely used to forecast stock market trends and price fluctuations.

Predicting price direction in the Foreign Exchange (Forex) market is a significant challenge, as traders rely on accurate predictions to make informed decisions. Correctly identifying price movement is essential for successful trading. This project focuses on predicting the price direction of USD/CAD (U.S. Dollar/Canadian Dollar) for the next minute.

The data used in this project was sourced from the Dukascopy demo Forex account (Dukascopy Bank SA, 2024), covering minute-by-minute data from September 1, 2024, to November 30, 2024. It includes key information such as Open, High, Low, Close (OHLC) prices, and volume for both ask and bid prices. The objective is to predict the price direction of USD/CAD using machine learning models, including Long Short-Term Memory (LSTM), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

2. Literature Review:

The prediction of Forex direction has gained significant attention in recent years. This section highlights several projects that have focused on predicting Forex direction, particularly for currencies like USD/CAD and others, using similar techniques.

A study by Ayitey Junior et al. (2022) proposed a method for dataset collection utilizing the Hurst component alongside a Two-Layered Stacked Long Short-Term Memory (TLS-LSTM) model. This approach enhanced the accuracy of predicting the Forex direction for AUD/USD. The study compared the performance of TLS-LSTM with Single-Layer LSTM and Multi-Layer Perceptron (MLP) models. Evaluation metrics such as Mean Squared Error

(MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) demonstrated that the TLS-LSTM model with the Hurst component outperformed other models in accuracy.

Another study utilized LSTM models to predict Forex direction using combined datasets of macroeconomic and technical indicator data. The technical indicators included OHLC (Open, High, Low, Close) prices and volume, processed using lagging techniques. Moving Averages (MA) were applied to smooth price data, while additional indicators like Moving Average Convergence Divergence (MACD), Rate of Change (ROC), and Relative Strength Index (RSI) were incorporated. Covering data from January 2013 to 2018 with 1,234 data points, the model predicted directions for 1-day, 3-day, and 5-day intervals. Results showed that for 1-day predictions, individual LSTM models slightly outperformed ME-LSTM, with less than a 1% difference. For 3-day predictions, individual models achieved 5.81% higher accuracy, while both models performed similarly for 5-day predictions (Yıldırım, Toroslu & Fiore, 2021).

In a more recent paper, Guyard and Deriaz (2024) predicted the direction of EUR/USD for 1-day ahead using various machine learning models. The analysis pipeline incorporated both decorrelated and non-decorrelated feature sets via Principal Component Analysis (PCA). Data from April 30, 2013, to December 31, 2022, with the final year used for evaluation, included economic indicators, market data, and Forex-specific data. Their findings revealed that Gradient Boosting achieved the highest accuracy for 1-day predictions at 58.52%, outperforming other models such as Logistic Regression (LR), K-Nearest Neighbors (KNN), SVM, Decision Trees, and Random Forest (RF).

3. Methodology

The approach to solving the problem is presented in the following pipeline, inspired by the review conducted by Jiang (2021). It outlines the solution in four main steps:

1. Data Acquisition

2. Data Preprocessing
3. Model Prediction
4. Model Evaluation

3.1. Dataset acquisition:

The dataset used in this project spans a period of three months, from September 1, 2024, to November 30, 2024, with each row representing one minute of data. It combines bid and ask price datasets, collected from a Dukascopy demo account, into a single dataset. This merged dataset includes OHLC (Open, High, Low, Close) prices and volumes for both the bid and ask sides, commonly referred to as market data according to the review by Jiang (2021). The final dataset contains 93,600 rows and 9 columns.

3.2. Data Preprocessing:

The data used in previous studies for predicting market direction generally includes OHLC prices and volume, but it is often unclear whether bid or ask prices were utilized. In this project, the midpoint of the bid and ask prices is used, a standard approach to measure the bid-ask spread, chosen for its simplicity (Ardia, Guidotti & Kroencke, 2024).

Since market direction is influenced by the relationship between the previous and current close prices, a lagging feature was created by shifting the midpoint close column by one row. In addition, two Simple Moving Averages (SMA) were applied: one with a 10-minute period and another with a 50-minute period. SMA, a trend-following indicator, smooths prices by averaging them over a set time frame (Likhith et al., 2023). This method helps reduce noise and provides insights into support and resistance levels (Yıldırım, Toroslu & Fiore, 2021). Furthermore, the Relative Strength Indicator (RSI) was calculated. The RSI is a valuable tool for identifying whether a market is overbought (above 70) or oversold (below 30), and it has been shown to offer useful insights for market direction, as demonstrated in the study by Indah and Mahyuni (2022).

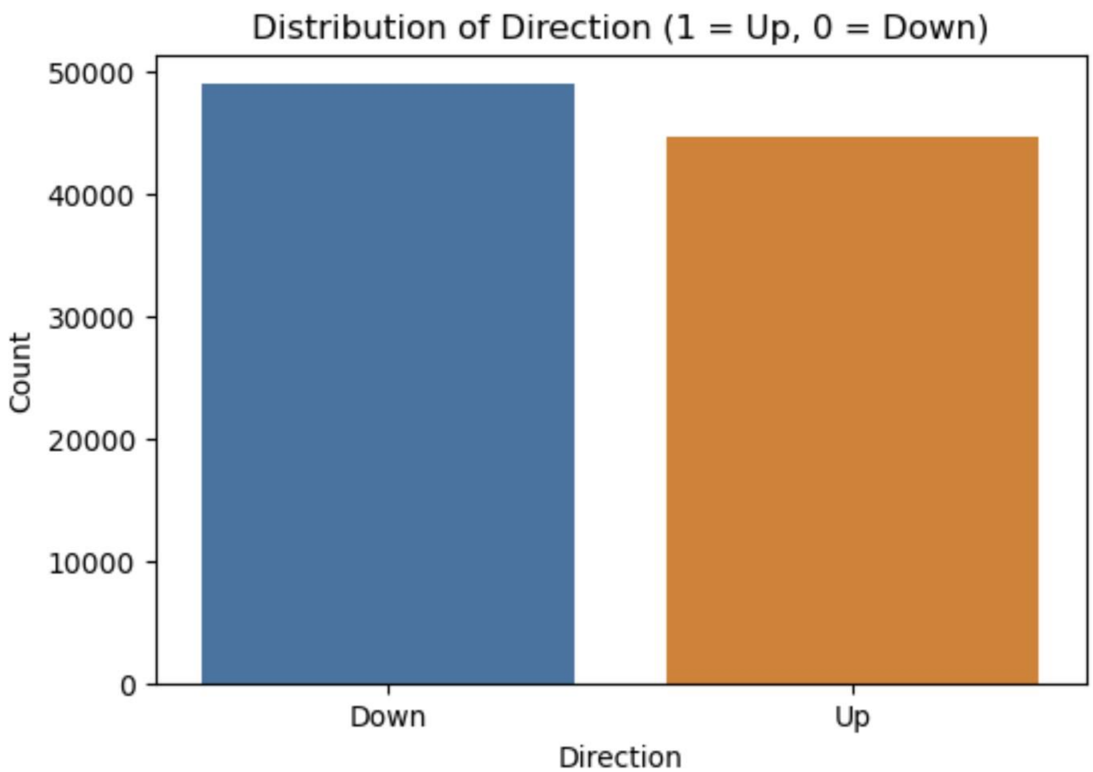
To define the target column, **Direction**, the following rule was applied:

- If close_mdp+1 > close_mdp, then direction = 1; otherwise, direction = 0.

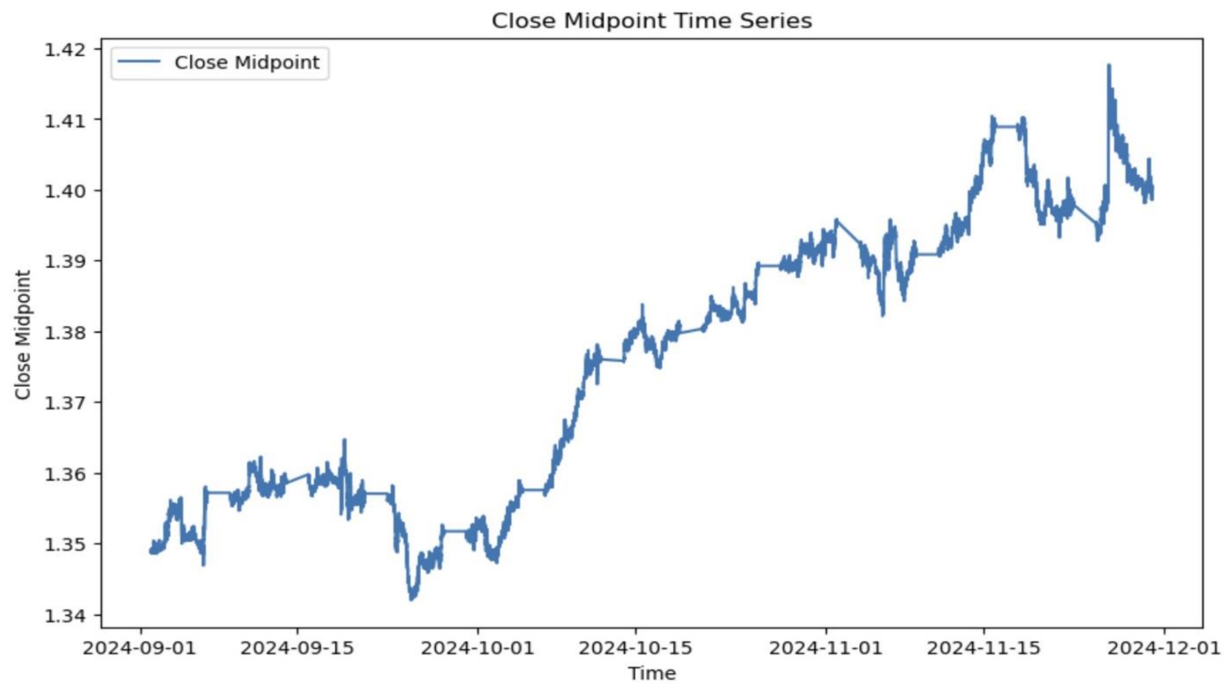
Additionally, recognizing that time is a crucial factor in market prediction, a feature was created to indicate the start of a new hour. After applying these preprocessing techniques, the dataset consisted of 93,551 rows, 9 features, and 1 target column.

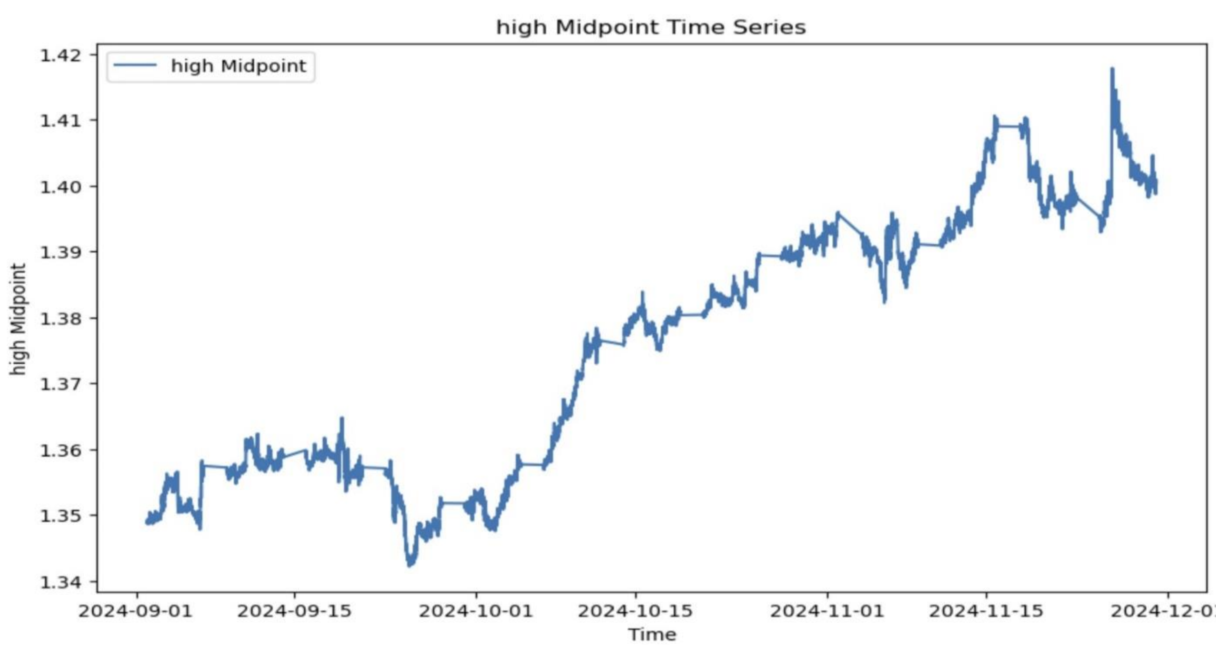
3.2.1. Data Analysis:

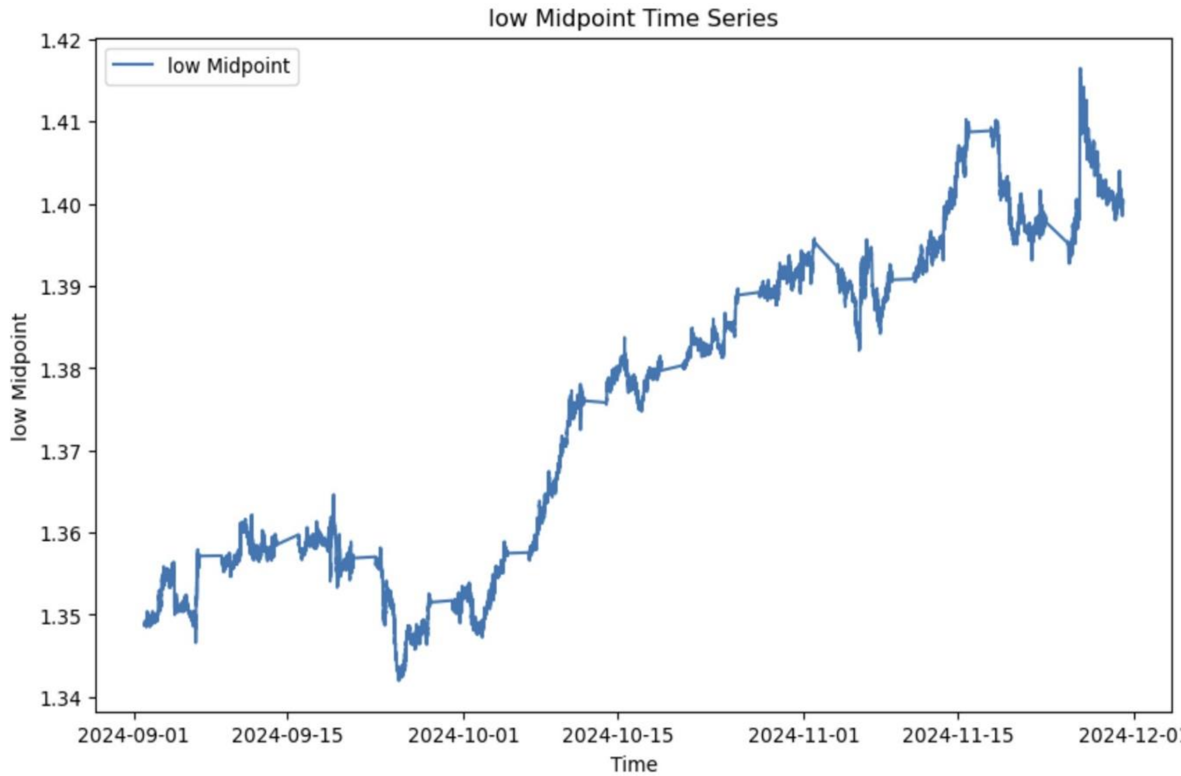
The model is designed to predict price movements as either upward (1) or downward (0). Over the three-month period, the data reveals a higher frequency of downward movements compared to upward ones. The imbalance in the dataset, with more downward movements, is also common in real-world financial data, where market corrections and volatility spikes often lead to more frequent price drops (Yıldırım, Toroslu & Fiore, 2021).



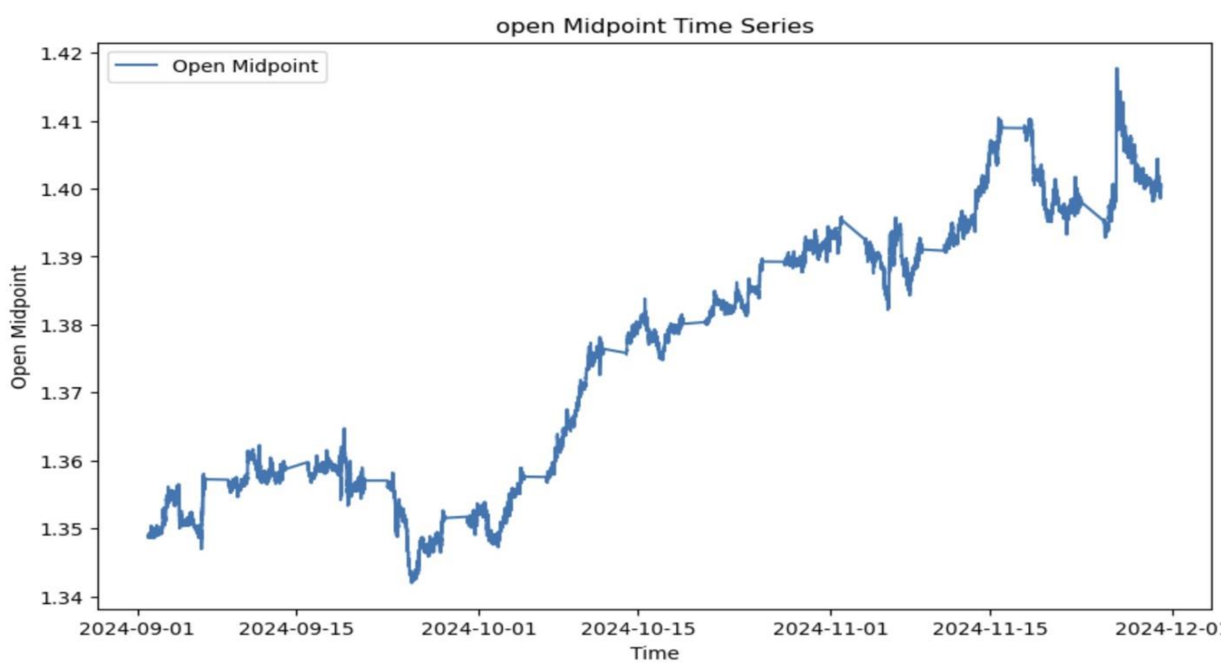
The following figures display the changes in the close, open, high, and low midpoints over time.





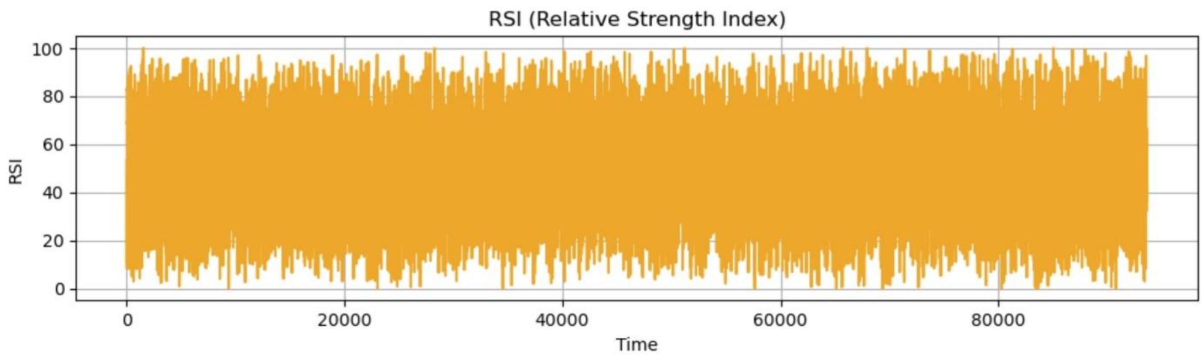


low Midpoint Time Series
low Midpoint
1.42
1.41
1.40
1.39
1.38
1.37
1.36
1.35
1.34
2024-09-01
2024-09-15
2024-10-01
2024-10-15
2024-11-01
2024-11-15
2024-12-01
Time



open Midpoint Time Series
Open Midpoint
1.42
1.41
1.40
1.39
1.38
1.37
1.36
1.35
1.34
2024-09-01
2024-09-15
2024-10-01
2024-10-15
2024-11-01
2024-11-15
2024-12-01
Time

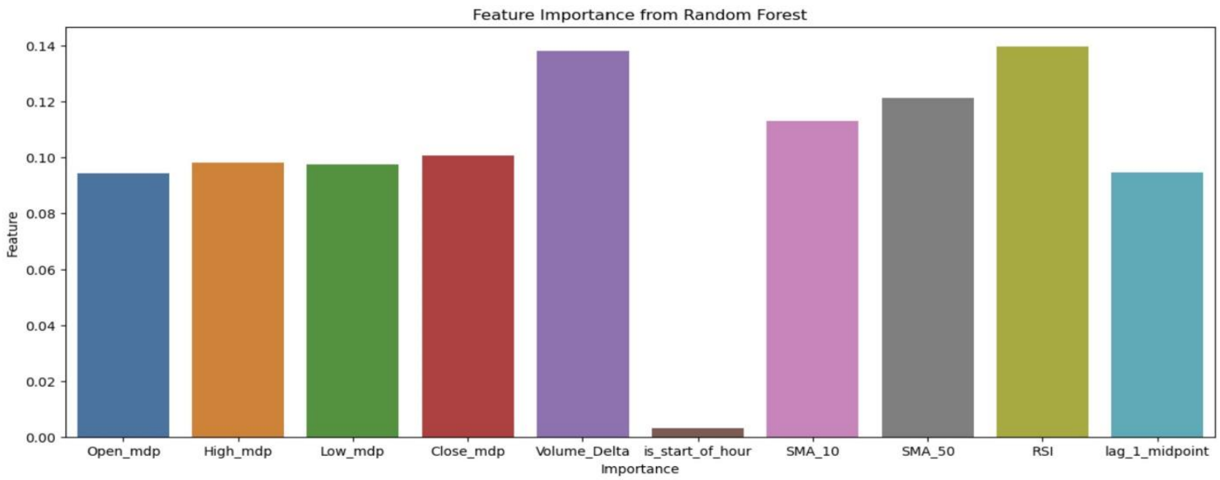
This graph illustrates the Relative Strength Index (RSI) over time. It appears that the market has remained mostly stable.

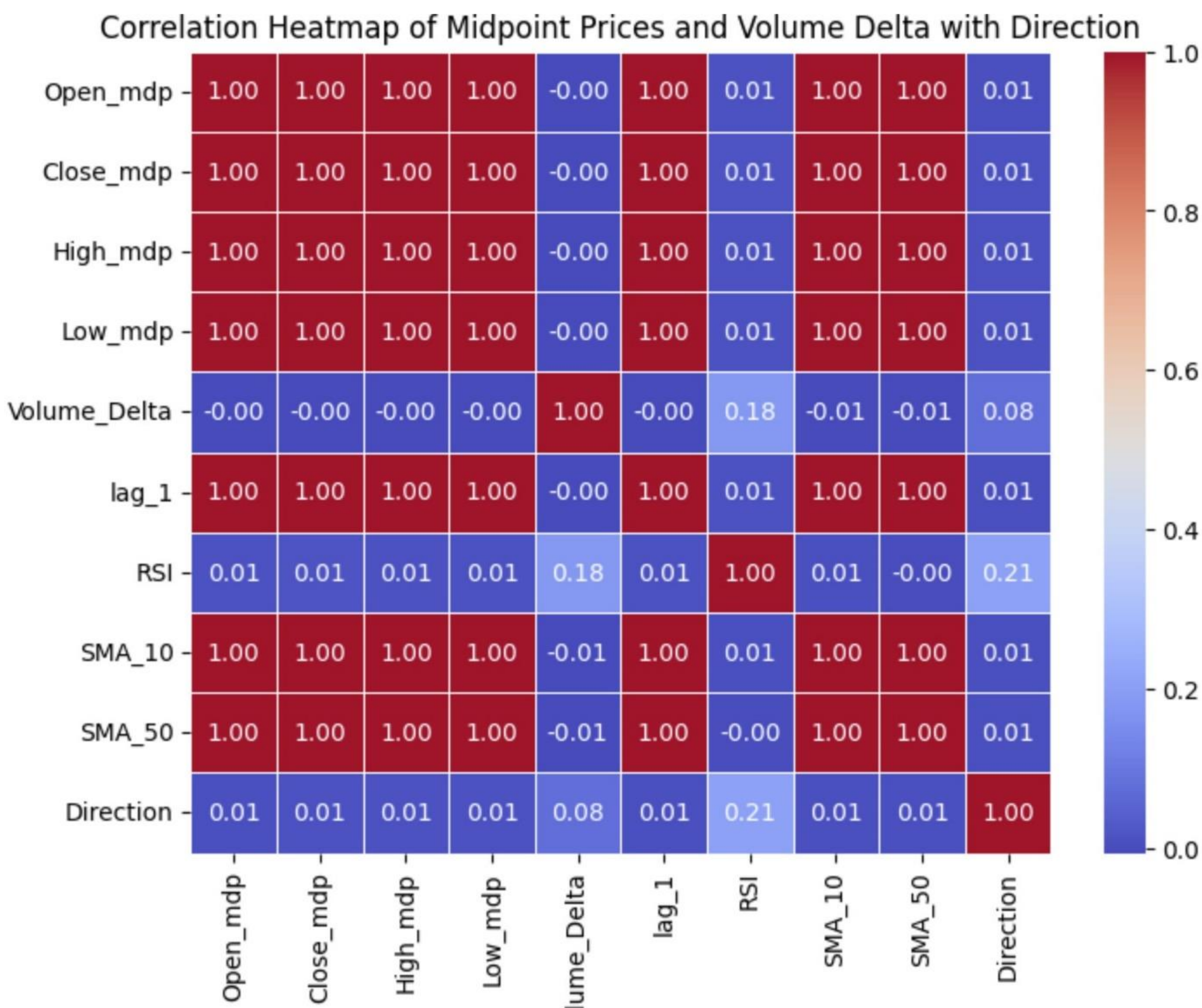


3.3. Model training:

The dataset is divided into training and testing sets in an 80:20 ratio. The features and target are separated, with the features structured in 2D and the target in 1D. These 2D feature datasets are then used for training with models like Random Forest Classifier, SVM, and MLP. Before applying the models, the data is normalized using MinMaxScaler. Feature importance is assessed, revealing that **Volume Delta** and **RSI** are the most influential features, while the start of a new hour has the least impact. Subsequently, the dataset is reshaped into a 3D format to train and test with the LSTM model.

A correlation matrix is also generated to show the relationship between the direction and other columns, indicating that **RSI** and **volume** significantly impact the direction prediction.


Feature Importance from Random Forest
Feature
Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Open_mdp
High_mdp
Low_mdp
Close_mdp
Volume_Delta
is_start_of_hour
SMA_10
SMA_50
RSI
lag_1_midpoint


Correlation Heatmap of Midpoint Prices and Volume Delta with Direction
Open_mdp
Close_mdp
High_mdp
Low_mdp
Volume_Delta
lag_1
RSI
SMA_10
SMA_50
Direction
lume_Delta
0.0
0.2
0.4
0.6
0.8
1.0

The MLP model consists of two hidden layers and one output layer, while the LSTM model is structured with one input layer and one output layer. Both models were trained for 20 epochs, with early stopping employed to monitor changes in accuracy. If no significant changes were observed for five consecutive epochs, training was halted to prevent overfitting.

Model: "sequential_4"

Layer (type)	Output Shape	Param #
dense_12 (Dense)	(None, 128)	1,408
dense_13 (Dense)	(None, 64)	8,256
dense_14 (Dense)	(None, 1)	65

Total params: 9,729 (38.00 KB)
Trainable params: 9,729 (38.00 KB)
Non-trainable params: 0 (0.00 B)
None

Model: "sequential_6"

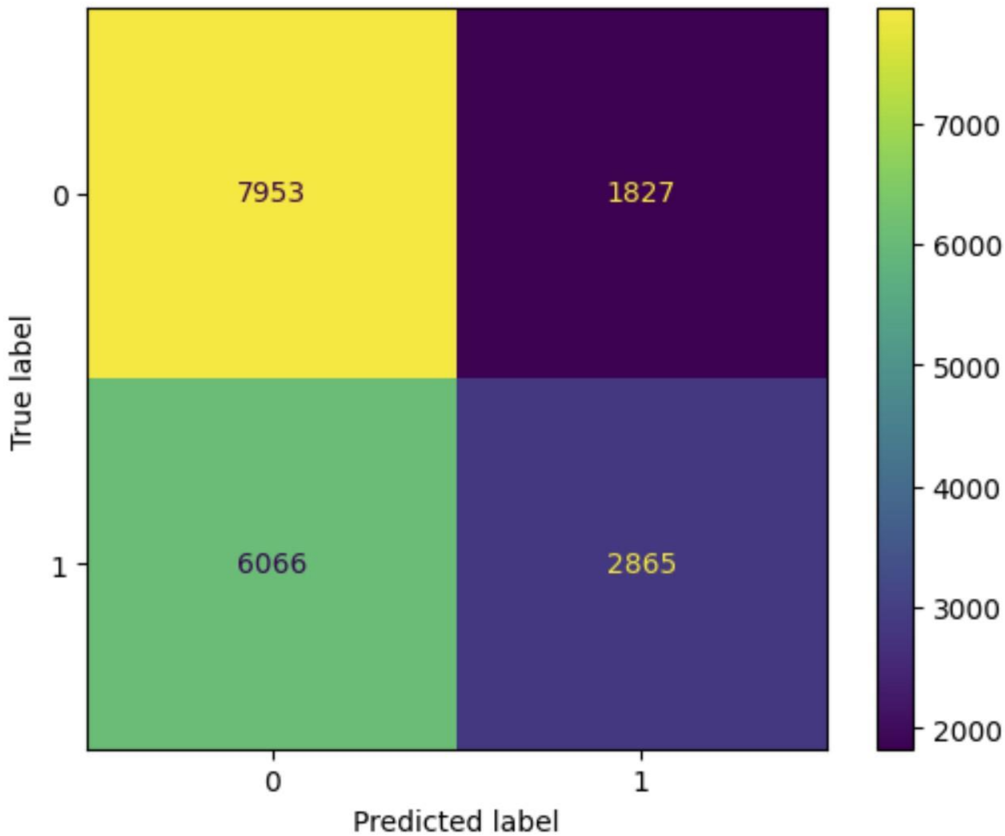
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 50)	12,200
dense_17 (Dense)	(None, 1)	51

Total params: 12,251 (47.86 KB)
Trainable params: 12,251 (47.86 KB)
Non-trainable params: 0 (0.00 B)
None

3.4. Model Evaluation and Results

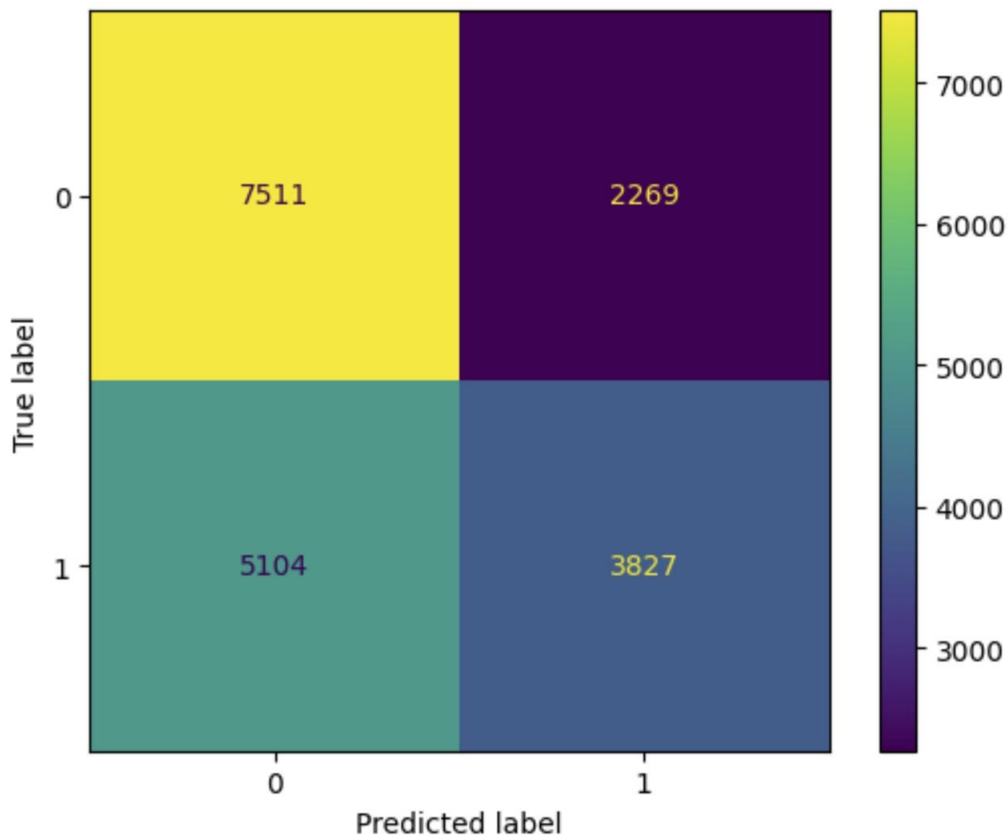
The four models were evaluated using unseen test data, yielding varying results. The RF model achieved an accuracy of 58%, correctly labeling 10,818 out of 18,711 rows.

Random Forest Confusion Matrix:

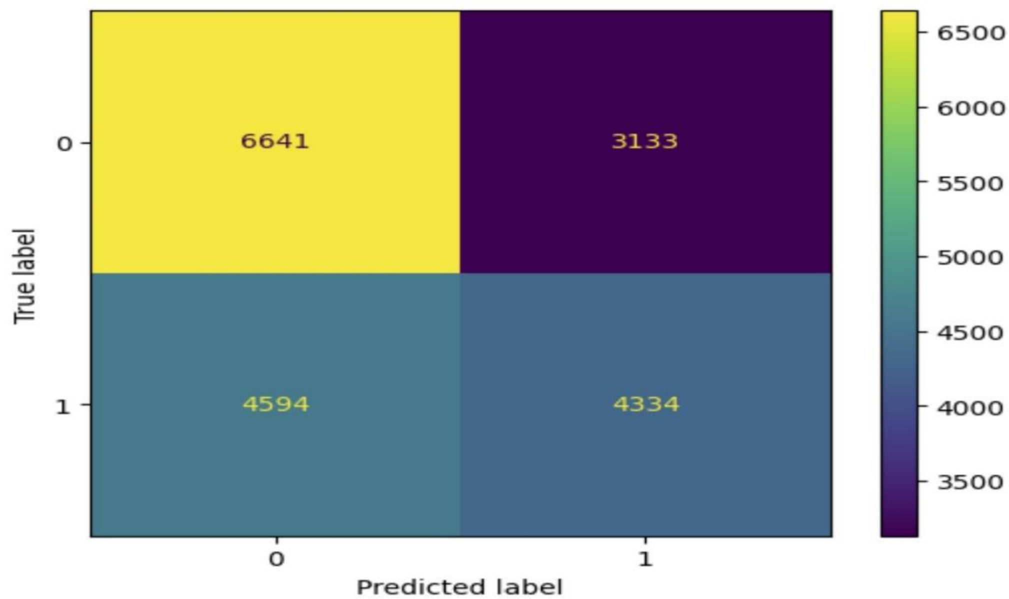


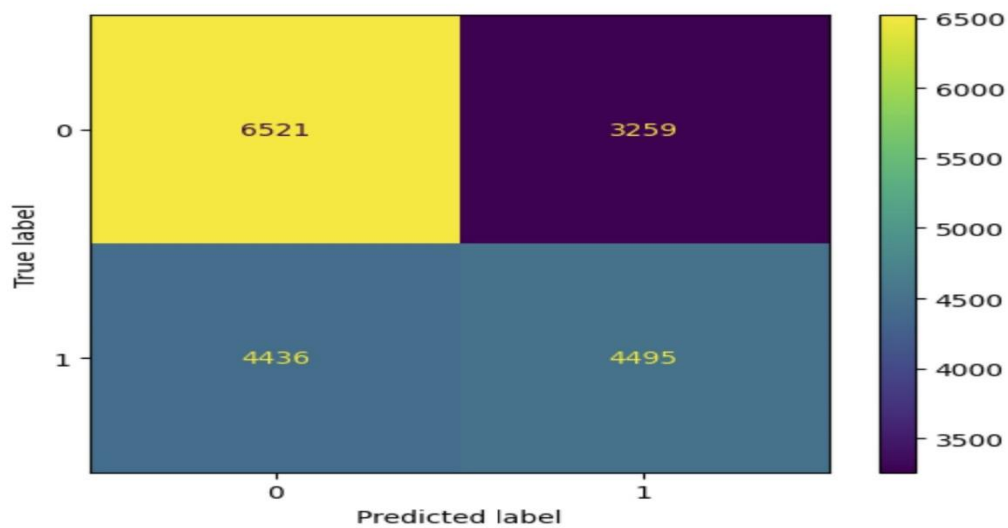
The SVM model achieved a higher accuracy of 61%, correctly predicting 11,338 out of 18,711 rows.

SVM Confusion Matrix:

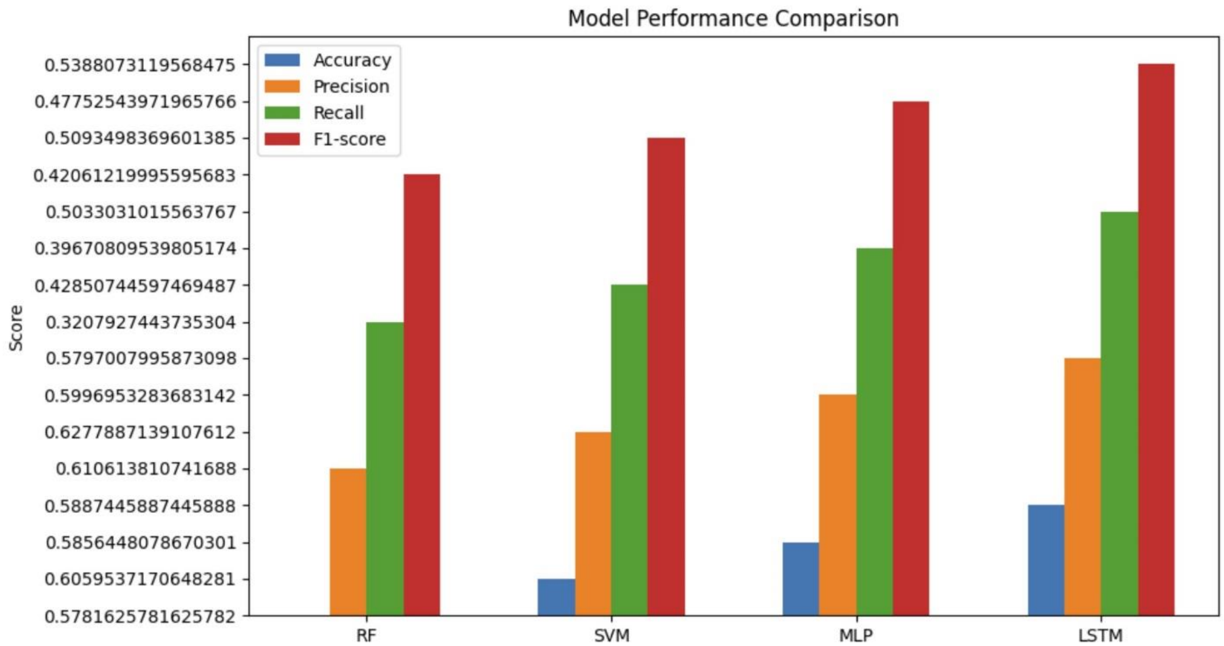


The two neural network models, MLP and LSTM, achieved similar accuracy values of 58.22% and 58.76%, respectively. Both models showed better performance in predicting upward price movements compared to the SVM and RF models. However, overall, the SVM model delivered the best accuracy.





The models were also evaluated based on other metrics such as precision, recall, and F1-score. The comparison revealed that the SVM model performed best across all four metrics, outperforming the other three models.



4. Conclusion

The Forex direction for the USD/CAD currency pair was predicted for the next minute using four different machine learning models: RF, SVM, MLP, and LSTM. Among these models, SVM demonstrated the best performance, with an accuracy exceeding 61%. This result can potentially be improved further by exploring different Moving Average (MA) techniques and capturing the relationship between the close price and the market direction.

5. Reference:

1- Dukascopy Bank SA. (2024) *Dukascopy Forex Demo Account*. Available at: https://www.dukascopy.com (Accessed: 10 December 2024).

2- Ayitey Junior, M., Appiahene, P. & Appiah, O., 2022. Forex market forecasting with two-layer stacked Long Short-Term Memory neural network (LSTM) and correlation analysis. *Journal of Electrical Systems and Information Technology*, 9(14). Available at: https://doi.org/10.1186/s43067-022-00054-1.

3- Yıldırım, D.C., Toroslu, I.H. & Fiore, U., 2021. Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. *Financial Innovation*, 7(1). Available at: https://doi.org/10.1186/s40854-020-00220-2 [Accessed 10 December 2024].

4- Guyard, K.C. & Deriaz, M., 2024. Predicting Foreign Exchange EUR/USD Direction Using Machine Learning. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*. New York, NY, USA: Association for Computing Machinery, pp. 1–9. Available at: https://doi.org/10.1145/3696271.3696272.

5- Ardia, D., Guidotti, E., & Kroencke, T.A., 2024. Efficient estimation of bid–ask spreads from open, high, low, and close prices. *Journal of Financial Economics*, 161, p.103916. Available at: https://doi.org/10.1016/j.jfineco.2024.103916.

6- Jiang, W. (2021) 'Applications of deep learning in stock market prediction: Recent progress', *Expert Systems with Applications*, 184, p. 115537. Available at: https://doi.org/10.1016/j.eswa.2021.115537.

7- Likhith, S.R., Nausheen, N., Abbas, A., Sarayu, V.S., Hemanth, B. and Varvatti, R.K. (2023) 'Predicting future stock market trends using Python and machine learning', *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, Bangalore, India, pp. 1-9. Available at: https://doi.org/10.1109/ICCAMS60113.2023.10526019.

8- Indah, Y.R. and Mahyuni, L.P., 2022. The accuracy of relative strength index (rsi) indicator in forecasting foreign exchange price movement. *Inovbiz: Jurnal Inovasi Bisnis*, *10*(1), pp.96-101.