Predicting Intraday Volume and Its Implication on Algorithmic Trading

Hoi Lam Kong, Jekyll Song, Hao Wen, Ziyang Zeng, Yue (Susie) Zhang

2 March 2025

Abstract

Trading volume has many important implications in the financial industry as an indicator of market activities. Accurately forecasting intraday trading volume is also critical to algorithmic trading, as it influences liquidity and market impact. This work develops a novel approach based on the Kalman filter and neural network to predict intraday trading volume efficiently. It also proposes an enhanced version incorporating robust techniques to handle market noise and outliers, leading to superior performance in volume forecasting and Volume Weighted Average Price (VWAP) tracking. We also go beyond predicting the volume alone by leveraging this data to build up a simulated execution algorithm and innovatively separate the order flow direction from the Limit Order Book (LOB) and construct a Directional Volume Weighted Price (DVWAP) method. Our results show that successfully applying these techniques improves the transaction cost and beats the naive price movement assumption.

1 Introduction

With systematic trading becoming more popular in the modern financial trading industry, many firms are putting extra effort into getting a more efficient execution algorithm (Chaboud et al., 2014)[8]. The execution algorithm, also known as algorithmic trading, is a trading algorithm that takes the quantity and direction as input and outputs an optimal trading strategy that could minimize the market impact and achieve the best average executed price. This algorithm does not predict the market direction but uses the information coming from the market to execute the input order (Hendershott et al., 2011)[12]. With a well-designed execution algo, the transaction cost could be greatly optimized (Bertsimas & Lo, 1998)[4]. However, most algorithms assume the process for the midprice process rather than modeling it endogenously. It is shown that under the Brownian Motion assumption, Time-Weighted Average Price (TWAP) is the optimal strategy and if we ignore the risk of the price movement, then Volume-Weighted Average Price is optimal (Almgren & Chriss, 2000)[2]. This leads to our key indicator, the volume, to be critical in designing the optimal strategy.

On the buy side of the financial industry, due to the nature of many institutional financial wealth managers, who put a lot of emphasis on predicting the market movement or exploring the market alpha, they put disproportionally less attention to execution often as it isn't the primary focus of their research, especially those operating with lower-frequency strategies. These funds typically rely on prime brokerage services from banks, executing trades through algorithms provided by the brokers. These algorithms are either following the previous volume-weighted price or just an equal-paced trading strategy, trying to minimize market impact by spreading out trades over time.

However, these algorithms are overly simplistic and inefficient. High-frequency trading firms that anticipate the timing of these brokers' orders could potentially position themselves to profit from these trades. Although pinpointing the active side of each order in the market is challenging, we can approach this from a volume prediction perspective, as many researchers have shown the importance of a VWAP strategy (Obizhaeva & Wang, 2013)[15]. Thus a better execution algorithm with accurate trading volume information could bring some adds-on alphas to the existing strategy.

Our paper leverages order book data to forecast trading volumes over various future time horizons by using a Kalman filter approach, which is more intuitive to understand the nature of the settings, and also implements a nonlinear autoregressive (NAR) framework to capture complex temporal dependencies by using historical trading volumes as inputs to predict future values. The neural network model incorporates multiple time lags as predictive features, effectively learning the sequential patterns and cyclical nature of the market activity. By implementing both one-layer and two-layer configurations with various neuron counts and lag structures, we systematically evaluate model performance across different architectures and select the optimal structure based on prediction accuracy and model stability (Xu & Zhang, 2023)[17].

With the improvement in predicting the trading volume, we explore the prediction on both 1-minute and 15-minute horizons and then develop a VWAP strategy based on the information predicted. In the 1-minute prediction horizon, we apply the VWAP strategy and in the 15-minute, we combine the VWAP with TWAP to smooth the execution volume. Even though the original prediction from a 1-minute interval is quite noisy due to the high volatility nature, both VWAP strategies show some decent improvement compared to the final price.

We then use signed trading volume, e.g., treating bid and ask sides separately, to extend this framework to include the DVWAP strategy to alleviate some issues we observed in VWAP strategies such as too much trade concentration at the open and close. We believe that examining the market microstructure of the trade flow can further improve these results. This not only provides one more dimension of the volume information but also has important implications on LOB flow (Biais et al., 1995)[5]. The LOB order flow will have some indication on the further market price movement, which will help reduce the risk of getting an unfavorable price when the market is trending. Our result shows that even though using the DVWAP strategy yields a higher variance and makes the results mixed, the potential to achieve an overall much better result is higher.

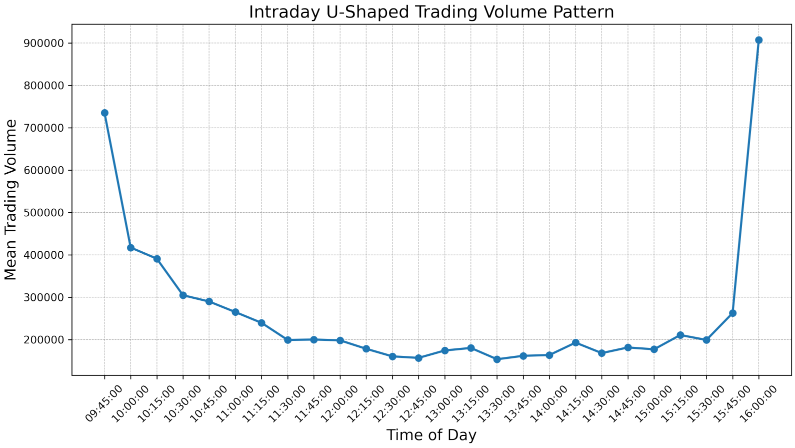


Figure 1: U-shaped Distribution of Daily Trading Volume

2 Data

To investigate whether a Kalman filter and neural network can improve predictions of intraday trading volume and assess their performance through simulated execution, we analyze time series data for the two most traded ETFs, SPY and QQQ. Our data is sourced from Nasdaq TotalView, collected by Databento, and consists of minute-level OCHLV data along with level-2 snapshots of the bid/ask order book. Specifically, we use OCHLV data from 9:30 AM EST to 4:00 PM EST, providing 390 data points per ticker per day, and capturing the 10-level bid/ask order book snapshot at the close of each minute for evaluation.

2.1 Intraday trading volume patterns

Intraday trading volume exhibits a well-documented U-shaped pattern, characterized by heightened activity at the market open and close, with relatively lower trading intensity during the middle of the trading session. This phenomenon is primarily driven by several market microstructure effects. At the opening, traders adjust positions based on overnight information, economic releases, and macroeconomic news, leading to a surge in trading activity. Similarly, toward the close, market participants engage in portfolio rebalancing, hedging, and trade executions driven by institutional investors' order flow, which further elevates trading volume (Admati & Pfleiderer, 1988)[1].

To empirically verify this pattern, we analyze a dataset containing trading volume recorded at 15-minute intervals throughout the trading session. The mean trading volume at each time interval is computed across the sample period, revealing a clear U-shaped distribution. The results from Figure 1 confirm prior literature findings that volume tends to peak during the first and last trading hours, reinforcing the view that trading decisions are influenced by both information arrival and liquidity considerations.

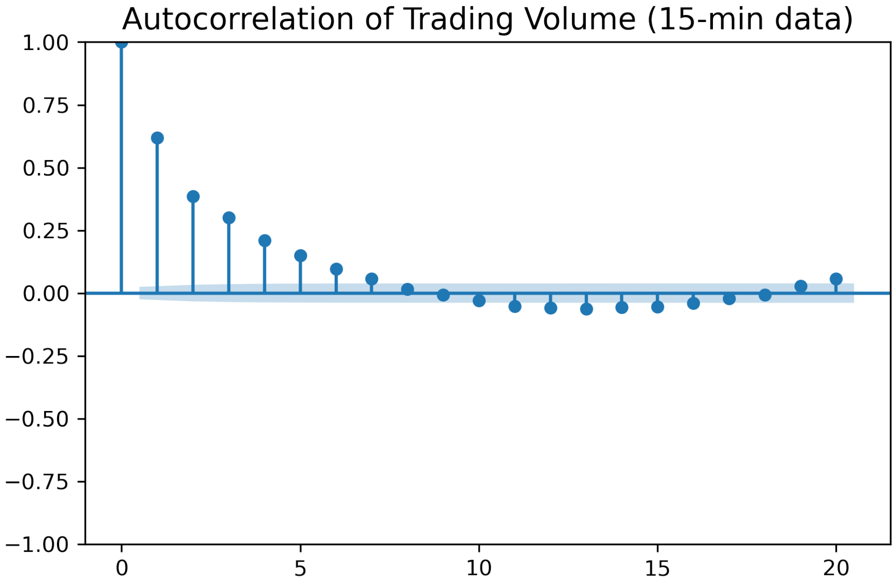


Figure 2: Autocorrelation of Trading Volume (15-minute data)

2.2 Autocorrelation of Trading Volume

Another key characteristic of intraday trading volume is its significant autocorrelation structure. High-frequency trading data typically exhibit strong positive short-term autocorrelation, reflecting order splitting by institutional investors and liquidity providers (Gabaix et al., 2003)[11]. Traders often execute large orders gradually over time to minimize trading costs due to market impact, leading to persistence in trading volume.

To quantify this effect, we compute the autocorrelation function (ACF) of trading volume for different lags. The results indicate that trading volume is highly correlated, especially at shorter lags. This finding aligns with previous studies suggesting that the autocorrelation of volume gradually decays over time but remains significant, implying that past trading intensity influences future activity. Understanding this behavior is crucial for designing volume-based execution algorithms, as well as improving forecasting models used in liquidity prediction and market impact analysis (Bouchaud et al., 2004)[7].

The autocorrelation plot in Figure 2 shows a strong positive autocorrelation at short lags, particularly at lag 1, confirming the persistence in trading activity. As the lag increases, the correlation gradually decreases, with some negative values appearing at longer lags. This suggests that while volume patterns exhibit short-term clustering, there is also a tendency for mean reversion over longer periods. The shaded confidence bands indicate the statistical significance of the autocorrelation values, with most early lags falling outside these bands, reinforcing the reliability of these observations.

3 Kalman Filter for Intraday Volume Prediction

3.1 Methodology

The Kalman filter is a recursive estimation algorithm that provides an optimal solution for tracking latent states in a dynamic system. Originally introduced by Kalman (1960)[13] for linear quadratic estimation problems, it has since been widely applied in various domains, including financial time series modeling, high-frequency trading, and market microstructure analysis (Wood et al., 1985)[16]. The Kalman filter operates within a state-space representation, making it particularly well-suited for modeling time-varying processes such as intraday trading volume. In this study, we model trading volume as a state-space system where the observed volume at time t is composed of three main latent components: a long-term trend component, an intraday periodic component, and a short-term stochastic component. This follows the framework proposed by Andersen and Bollerslev (1998)[3] for modeling intraday seasonality and market microstructure effects.

The standard form of the state-space model consists of a measurement equation and a transition equation:

$$V_t = \eta_t + \phi_t + \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{1}$$

where V_t represents the observed trading volume, η_t denotes the long-term volume trend, ϕ_t captures recurrent intraday periodic fluctuations, and μ_t accounts for dynamic short-term variations. The error term ϵ_t represents measurement noise. The transition equations underline the evolution of these hidden states over time:

$$\eta_t = \eta_{t-1} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \sigma_\nu^2), \tag{2}$$

$$\phi_t = \sum_{j=1}^{J} \alpha_j \cos\left(\frac{2\pi jt}{T}\right) + \sum_{j=1}^{J} \beta_j \sin\left(\frac{2\pi jt}{T}\right) + \omega_t, \quad \omega_t \sim \mathcal{N}(0, \sigma_\omega^2), \tag{3}$$

$$\mu_t = \rho\mu_{t-1} + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_\xi^2), \tag{4}$$

where ν_t, ω_t, and ξ_t are normally distributed and independent of each other innovations. The periodic component ϕ_t is modeled using a Fourier series decomposition, consistent with the approach used in trading volume forecasting literature (Chen et al., 2016)[9]. The short-term component μ_t follows an autoregressive process to capture dynamic fluctuations driven by market conditions.

3.2 Recursive Estimation using the Kalman Filter

Given an initial state estimate $\hat{x}_{t-1|t-1}$ and covariance matrix $P_{t-1|t-1}$, the Kalman filter updates these estimates at each time step through two key steps:

Prediction Step: The state and covariance matrix are projected forward based on the transition equations:

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1} + Bu_t, \tag{5}$$

$$P_{t|t-1} = AP_{t-1|t-1}A' + Q, \tag{6}$$

where A is the state transition matrix, B represents control inputs (if any), and Q is the process noise covariance matrix.

Update Step: When a new observation V_t is available, the state estimate and covariance matrix are updated using the Kalman gain K_t:

$$K_t = P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}, \tag{7}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(V_t - H\hat{x}_{t|t-1}), \tag{8}$$

$$P_{t|t} = (I - K_tH)P_{t|t-1}, \tag{9}$$

where H is the observation matrix and R represents measurement noise. This recursive filtering process allows for real-time estimation of the latent components of trading volume while continuously adapting to new data. The method has been widely used in financial econometrics and market microstructure research due to its adaptability and robustness in tracking non-stationary processes (Chen et al., 2016)[9].

3.3 Performance Evaluation Metrics

To assess the effectiveness of the Kalman filter in trading volume prediction, we employ two common error metrics: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |V_i - \hat{V}_i|, \tag{10}$$

where V_i is the actual observed trading volume, $\hat{V}_i$ is the predicted volume, and N is the total number of observations. The MAPE is given by:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{V_i - \hat{V}_i}{V_i} \right|. \tag{11}$$

MAPE provides a relative measure of prediction accuracy, making it useful for comparing models across different instruments and timeframes.

3.4 Empirical Implementation and Results

We apply our Kalman filter and its robust variant to intraday trading data for SPY and QQQ from 2021 onwards, covering a period of heightened market volatility. The estimation is performed at both 1-minute and 15-minute intervals, allowing for comparison with traditional rolling mean and CMEM benchmarks. Table 1 reports mean absolute error

Table 1: Prediction Errors by Method and Symbol, 1-min

	MAE		MAPE	
Symbol	Rolling Mean	Kalman Filter	Rolling Mean	Kalman Filter
Panel A: Estimation using 1-minute interval				
SPY (Buy)	3583.59	3378.68	0.5811	0.5479
SPY (Sell)	3358.47	3165.14	0.5851	0.5514
QQQ (Buy)	3556.78	3338.50	0.5692	0.5343
QQQ (Sell)	3373.84	3165.43	0.6025	0.5653
Panel B: Estimation using 15-minute interval				
SPY (Buy)	38802.23	34911.62	0.4216	0.3793
SPY (Sell)	35212.78	31570.80	0.4110	0.3686
QQQ (Buy)	35673.39	31572.56	0.3821	0.3382
QQQ (Sell)	34492.90	30517.71	0.4123	0.3648

(MAE) and mean absolute percentage error (MAPE) results across different methods. The Kalman filter consistently outperforms the rolling mean model. Also, the 15-minute-level prediction outperforms the 1-minute-level prediction, which contains more noise because of the inherited characteristics of the trading data. This is best illustrated in Figure 3 reporting intraday volume predictions for SPY (Panel A) and QQQ (Panel B) using 15-minute intervals.

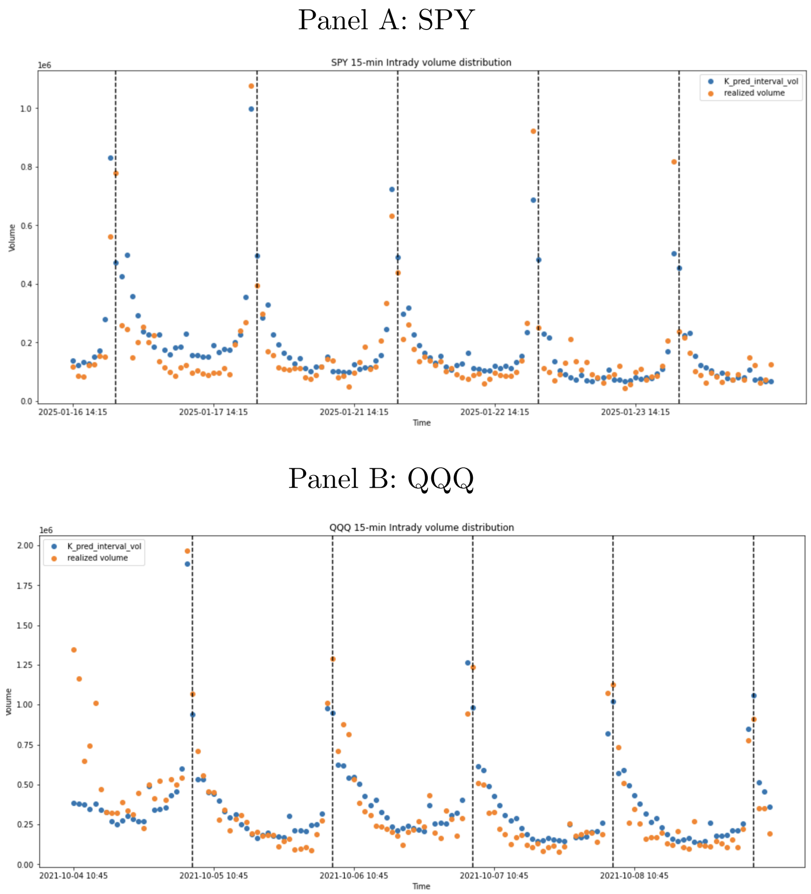
These findings highlight the practical advantages of state-space modeling and robust filtering techniques in algorithmic trading and execution strategy optimization. Future work could extend this framework to incorporate deep learning-based state estimation methods, as explored in recent studies on financial time series forecasting (Ma & Li, 2021)[14].

4 Neural Network for Intraday Volume Prediction

With the advancement of machine learning, neural networks have gained much more popularity. Among them, forward-feeding neural networks are the most commonly used techniques for a variety of applications in finance (Coakley & Brown, 2000)[10]. In this study, we employ a nonlinear autoregressive (NAR) neural network model to predict one-minute trading volumes of major equity index ETFs (SPY, QQQ). The NAR model is particularly suitable for this application due to trading volumes' strong temporal dependencies and nonlinear patterns, which are difficult to capture using traditional linear models.

4.1 Train/Test Split

Pre-processing data of the train-validation-test split is crucial to the performance of the neural network model (Bichri et al., 2023)[6]. We partition our data using a 60%/20%/20%


Panel A: SPY

Panel B: QQQ

Figure 3: Intraday volume prediction (15-minute)

split for training, validation, and testing, respectively. The training set builds the model, the validation set prevents overfitting through early stopping, and the test set provides a robust out-of-sample performance evaluation. The training sample covers September 2024 through November 2024 and the validation sample spans over December 2024, while the testing sample encompasses January 2025, providing a clear temporal separation for out-of-sample prediction evaluation.

Beyond aggregate volume, the directional flow of trading activity offers critical insights for microstructure strategy design. Therefore, we apply the same NAR model to analyze buy-only and sell-only datasets in addition to total volume. Table 2 presents a summary of the volume datasets along with their respective observation counts for training, validation, and testing periods.

4.2 Model

To predict the highly irregular trading volumes of SPY and QQQ, we implemented an artificial neural network (ANN) model. ANNs excel at learning complex patterns within datasets and making predictions by iteratively adjusting connection weights between neu-

Table 2: Summary of Datasets Used for NN Training

Index ETF	Freq	No of observation (Train)	No of observation (Validation)	No of observation (Test)
SPY	1min	24,439	8,146	7,800
	15min	1,638	546	520
QQQ	1min	24,467	8,115	7,800
	15min	1,638	546	520

rons based on training data. Specifically, we employed a nonlinear autoregressive (NAR) neural network model where the sole predictive inputs are historical lags of trading volume for the same index ETF. This relationship can be expressed as

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-d}) + \epsilon_t, \tag{12}$$

where y represents trading volume, t denotes time on a 1-minute or 15-minute basis, d is the number of lags (or delays in neural network terminology), and $f(\cdot)$ represents the nonlinear functional form of the NAR model.

To accommodate the varying sample sizes in our training datasets while maintaining an optimal sample-to-parameter ratio for effective generalization and fit, we implemented both one-layer and two-layer feedforward architectures. The one-layer neural network employs a sigmoid activation function in the hidden layer, while the two-layer neural networks utilize ReLU (Rectified Linear Unit) activation functions in both hidden layers, with a linear transfer function in the output layer. We configured the model in an open-loop form for efficient training, where actual historical values are used as inputs rather than feeding back predicted values.

Our model implementation includes the following specifications. For hidden neuron configurations, we tested single-layer networks with 5, 10, 20, and 40 neurons, and two-layer networks with configurations of [2,2], [5,2], [5,5], [16,8], and [32,16] neurons. We selected the sigmoid activation function for the single-layer model due to its smooth gradient and effective performance with smaller networks. For multi-layer models, we chose ReLU activations because they help mitigate the vanishing gradient problem during the training of deeper networks and generally enable faster convergence.

For time lags, we evaluated periods of 2, 3, 5, 10, 20, 30, and 50. This range was selected to capture both immediate short-term dependencies (2-5 lags) that might reflect immediate market reactions, medium-term patterns (10-20 lags) that could capture intraday trading patterns and longer-term dependencies (30-50 lags) that might reflect broader market rhythms within trading sessions.

For optimization, we employed the Adam optimizer with a mean squared error (MSE) loss function to iteratively update weights and biases. Adam was selected for its adaptive learning rate capabilities and effective performance with noisy gradients common in financial time series data.

Table 3: Model Settings of NN Tested

Hyperparameter	Setting
Delay	2 / 3 / 5 / 10 / 20 / 50
1-layer Hidden Neuron	[5] / [10] / [20] / [40]
2-layer Hidden Neuron	[2,2] / [5,2] / [5,5] / [16,8] / [32,16]

4.3 Results

We estimated all models based on the settings listed in Table 3 for predictions of the trading volume of SPY and QQQ. To evaluate the model's performance, we used the relative root mean square error (RRMSE) and computed it across training, validation, and testing phases for each model setting. When determining the optimal model setting for the prediction task, we need to consider both prediction accuracy and model stability. We will first examine the fitness of 1-layer and 2-layer neural network models and select the optimal delay and architecture based on average RRMSE and the trend of the loss function.

4.3.1 Choice between 1-layer and 2-layers NN

Our comprehensive evaluation of single-layer and two-layer neural network architectures for trading volume prediction revealed consistent performance advantages for the two-layer configuration across all test scenarios. The two-layer architecture demonstrated the following notable strengths:

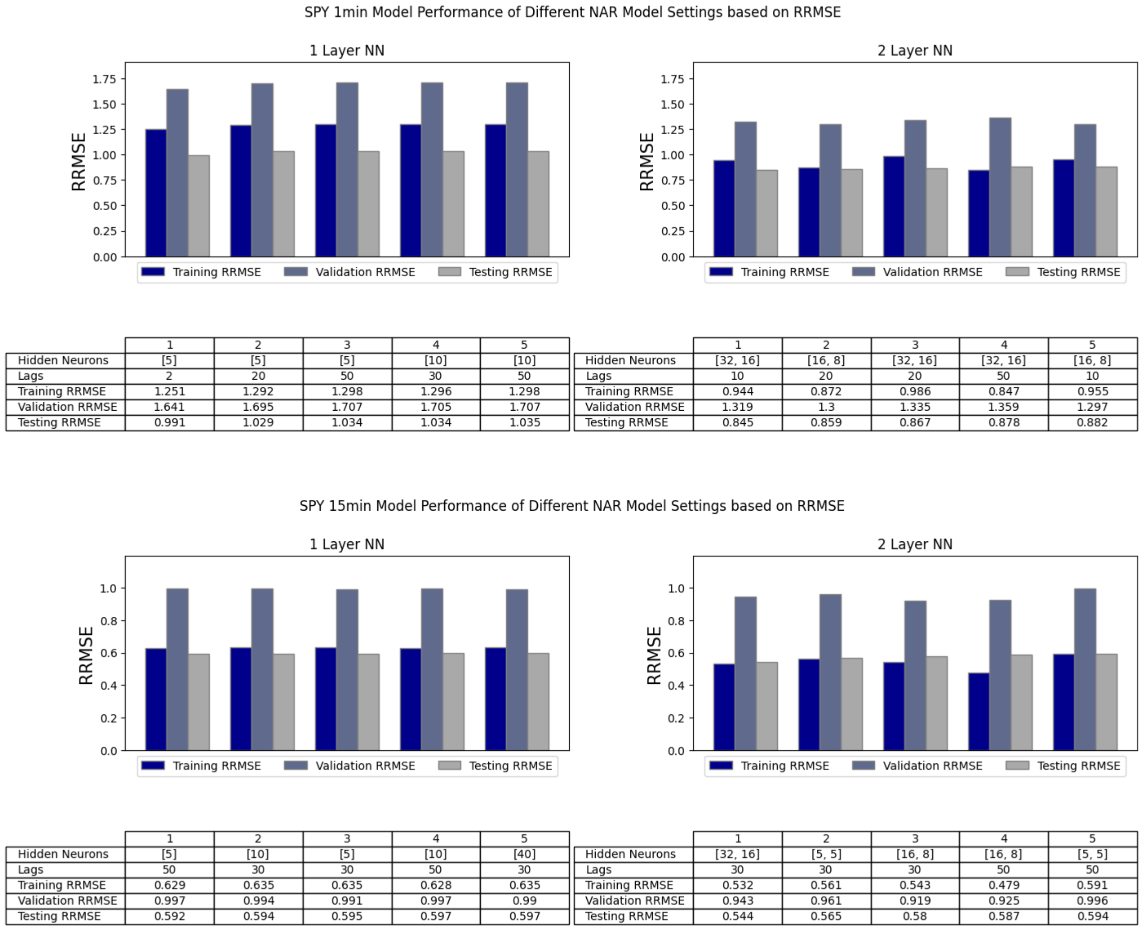
Superior Prediction Accuracy: Two-layer models consistently produced lower Relative Root Mean Square Error (RRMSE) values during testing phases. This improvement was particularly pronounced for 1-minute frequency data, where the temporal patterns exhibit greater complexity and volatility.

Enhanced Generalization Capability: The two-layer architecture exhibited a smaller disparity between training and testing errors, indicating superior generalization performance and reduced overfitting tendencies. This is critical for maintaining prediction reliability under varying market conditions.

Advanced Pattern Recognition: The additional layer provides increased model capacity to capture complex nonlinear relationships inherent in high-frequency trading volume data. This architectural advantage proved especially beneficial for processing the intricate temporal patterns present in 1-minute interval data.

Greater Temporal Adaptability: Two-layer models maintained strong performance across various lag configurations, demonstrating enhanced adaptability to different temporal dependencies within the data.

While both architectures delivered acceptable results for 15-minute frequency data, the more nuanced patterns in 1-minute data benefited substantially from the additional complexity of the two-layer network. This performance differential was most evident for QQQ



SPY 1min Model Performance of Different NAR Model Settings based on RRMSE

1 Layer NN

2 Layer NN

	1	2	3	4	5
Hidden Neurons	[5]	[5]	[5]	[10]	[10]
Lags	2	20	50	30	50
Training RRMSE	1.251	1.292	1.298	1.296	1.298
Validation RRMSE	1.641	1.695	1.707	1.705	1.707
Testing RRMSE	0.991	1.029	1.034	1.034	1.035

	1	2	3	4	5
Hidden Neurons	[32, 16]	[16, 8]	[32, 16]	[32, 16]	[16, 8]
Lags	10	20	20	50	10
Training RRMSE	0.944	0.872	0.986	0.847	0.955
Validation RRMSE	1.319	1.3	1.335	1.359	1.297
Testing RRMSE	0.845	0.859	0.867	0.878	0.882

SPY 15min Model Performance of Different NAR Model Settings based on RRMSE

1 Layer NN

2 Layer NN

	1	2	3	4	5
Hidden Neurons	[5]	[10]	[5]	[10]	[40]
Lags	50	30	30	50	30
Training RRMSE	0.629	0.635	0.635	0.628	0.635
Validation RRMSE	0.997	0.994	0.991	0.997	0.99
Testing RRMSE	0.592	0.594	0.595	0.597	0.597

	1	2	3	4	5
Hidden Neurons	[32, 16]	[5, 5]	[16, 8]	[16, 8]	[5, 5]
Lags	30	30	30	50	50
Training RRMSE	0.532	0.561	0.543	0.479	0.591
Validation RRMSE	0.943	0.961	0.919	0.925	0.996
Testing RRMSE	0.544	0.565	0.58	0.587	0.594

Figure 4: SPY 1&15 min Model Performance of Different NAR Settings based on RRMSE

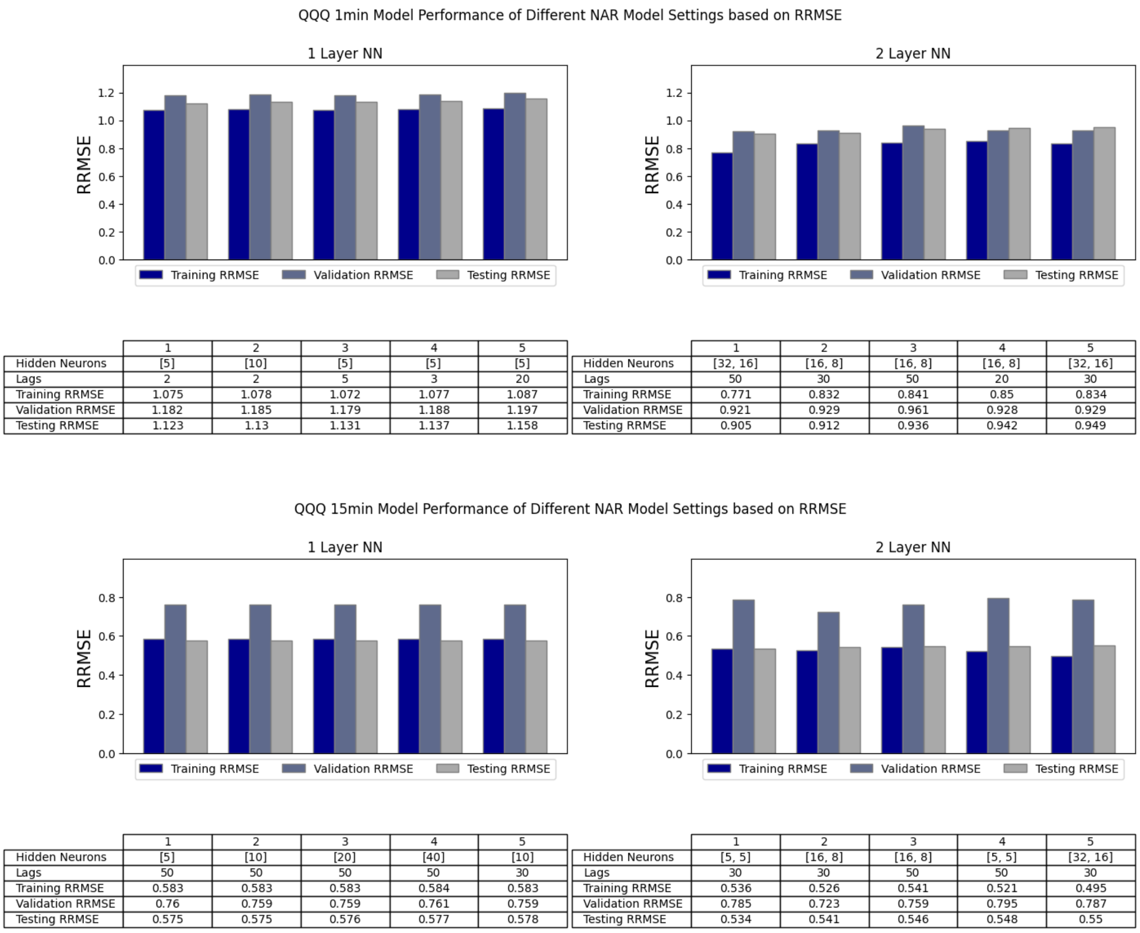
predictions, where the two-layer architecture reduced testing RRMSE by approximately 20% compared to single-layer implementations.

Based on these findings, we selected the two-layer neural network architecture as the optimal foundation for our trading volume prediction models.

4.3.2 Optimal Model Setting

Looking at the 2-layer neural network results, we picked the optimal configuration that balances low testing RRMSE, good degradation rate from training to testing loss, and appropriate data-to-parameter ratio.

SPY Analysis (1-minute Frequency): From the top subplot of Figure 4, the two-layer architecture with [32, 16] neurons and 20 lags shows the optimal result for SPY 1-minute trading volume prediction as it shows a good balance between prediction power, model generalization, and stability. The model consistently demonstrated superior predictive performance with a low testing RRMSE (0.867) and 10% loss reduction from training to testing. With 24,439 training observations and 1217 trainable parameters, the model main-



	1	2	3	4	5
Hidden Neurons	[5]	[10]	[5]	[5]	[5]
Lags	2	2	5	3	20
Training RRMSE	1.075	1.078	1.072	1.077	1.087
Validation RRMSE	1.182	1.185	1.179	1.188	1.197
Testing RRMSE	1.123	1.13	1.131	1.137	1.158

	1	2	3	4	5
Hidden Neurons	[32, 16]	[16, 8]	[16, 8]	[16, 8]	[32, 16]
Lags	50	30	50	20	30
Training RRMSE	0.771	0.832	0.841	0.85	0.834
Validation RRMSE	0.921	0.929	0.961	0.928	0.929
Testing RRMSE	0.905	0.912	0.936	0.942	0.949

	1	2	3	4	5
Hidden Neurons	[5]	[10]	[20]	[40]	[10]
Lags	50	50	50	50	30
Training RRMSE	0.583	0.583	0.583	0.584	0.583
Validation RRMSE	0.76	0.759	0.759	0.761	0.759
Testing RRMSE	0.575	0.575	0.576	0.577	0.578

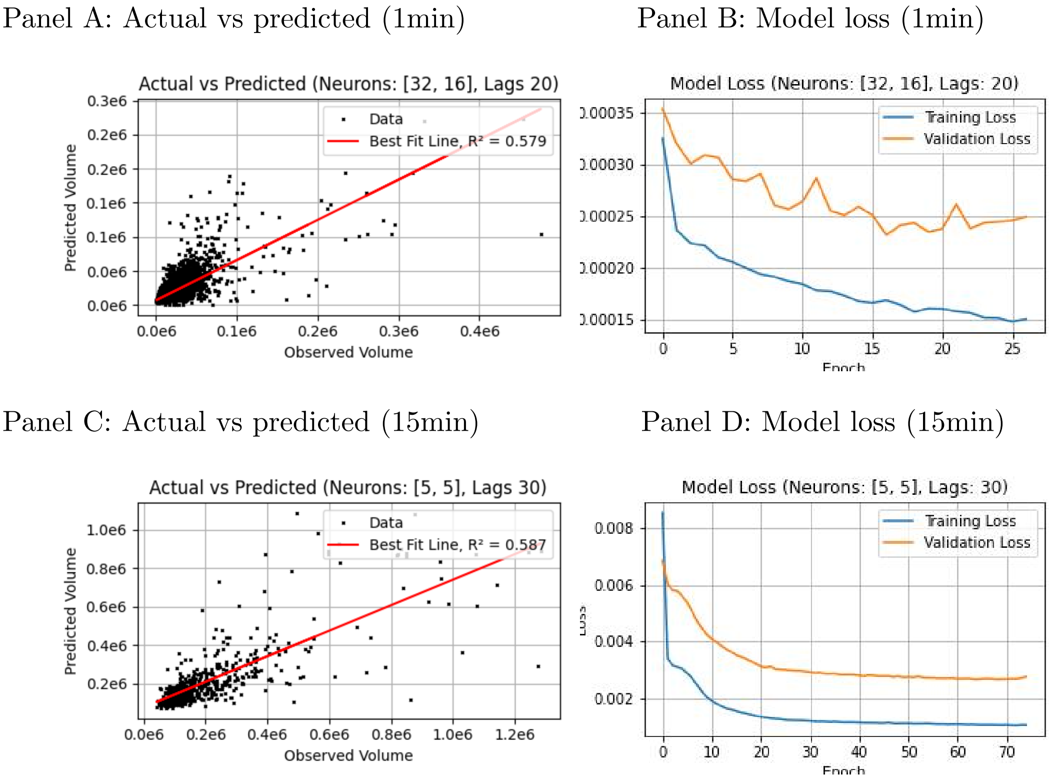
	1	2	3	4	5
Hidden Neurons	[5, 5]	[16, 8]	[16, 8]	[5, 5]	[32, 16]
Lags	30	30	50	50	30
Training RRMSE	0.536	0.526	0.541	0.521	0.495
Validation RRMSE	0.785	0.723	0.759	0.795	0.787
Testing RRMSE	0.534	0.541	0.546	0.548	0.55

Figure 5: QQQ 1&15 min Model Performance of Different NAR Settings based on RRMSE

tains an excellent data-to-parameter ratio of 20:1, ensuring robust generalization without overfitting.

The 20-lag structure strikes an ideal balance between capturing relevant temporal dependencies and maintaining model responsiveness to recent market dynamics. This architecture effectively captures the complex nonlinear relationships in high-frequency trading volume data while maintaining computational efficiency for practical deployment. The learning curve (the top right subplot of Figure 6) shows that both training and validation losses decrease in parallel throughout the training process. The consistent downward trajectory of both metrics without divergence confirms the model is learning effectively without overfitting, suggesting reliable generalization capabilities for this complex time series prediction task.

SPY Analysis (15-minute Frequency): Among the SPY 15-minute trading volume prediction models (the bottom subplot of Figure 4), we selected the [5, 5] neuron configuration with 30 lags as optimal based on our comprehensive evaluation. With 1,638 training observations, this architecture maintains an appropriate data-to-parameter ratio


Panel A: Actual vs predicted (1min)

Panel B: Model loss (1min)

Panel C: Actual vs predicted (15min)

Panel D: Model loss (15min)

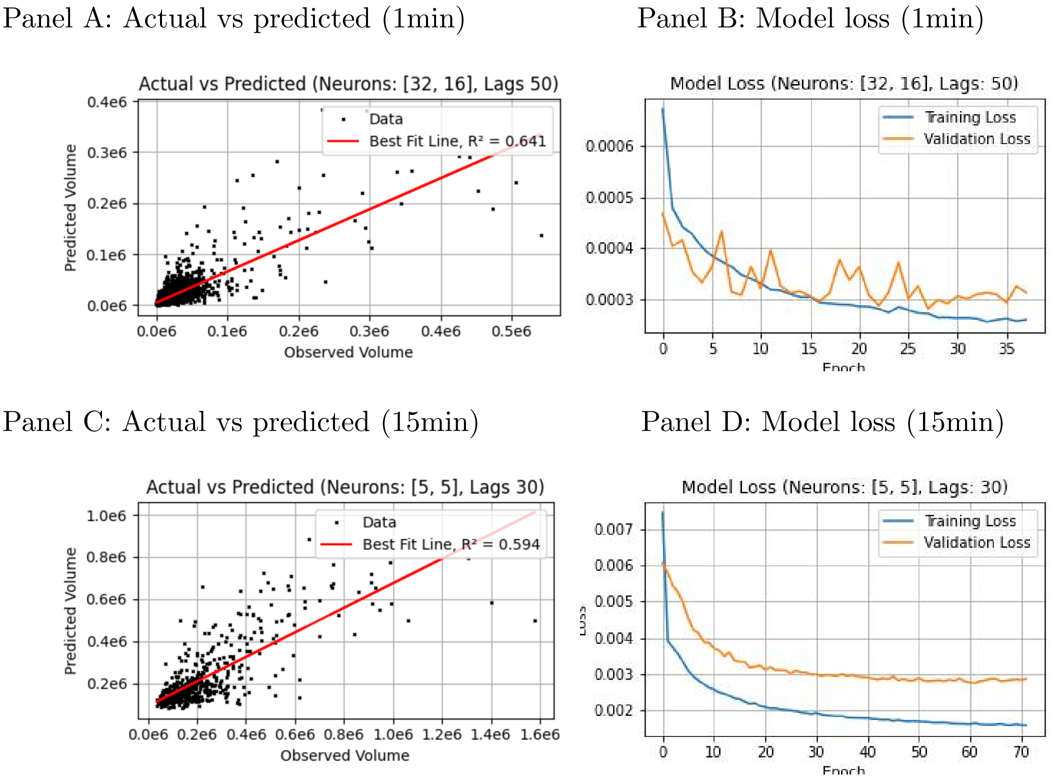
Figure 6: SPY Visualizations

of approximately 8:1, ensuring reliable generalization. The model demonstrates exceptional accuracy with the second lowest testing error across all settings.

The 30-lag structure is particularly well-suited for 15-minute data as it captures a 7.5-hour historical window, effectively spanning one trading day. This allows the model to incorporate critical cyclical patterns, as trading volumes at the same time across consecutive trading days typically exhibit strong correlations. This temporal structure enables the model to recognize and leverage daily seasonality patterns that are fundamental to market microstructure, particularly the characteristic U-shaped volume patterns observed during trading sessions.

Looking at the learning curve (the bottom right subplot of Figure 6), both training and validation losses decrease steadily until approximately epoch 40, after which they stabilize, with training loss converging at approximately 0.001 and validation loss at 0.0025. This pattern confirms the model has reached an optimal balance between fitting the training data while maintaining strong generalization capability, with the consistent gap between training and validation losses indicating appropriate model complexity for the available data.

QQQ Analysis (1-minute Frequency): For the QQQ 1-minute trading volume prediction (the top subplot of Figure 5), we identified the [32, 16] neuron architecture with 20 lags as the optimal configuration. Similar to SPY, this model demonstrates a relatively


Panel A: Actual vs predicted (1min)

Panel B: Model loss (1min)

Panel C: Actual vs predicted (15min)

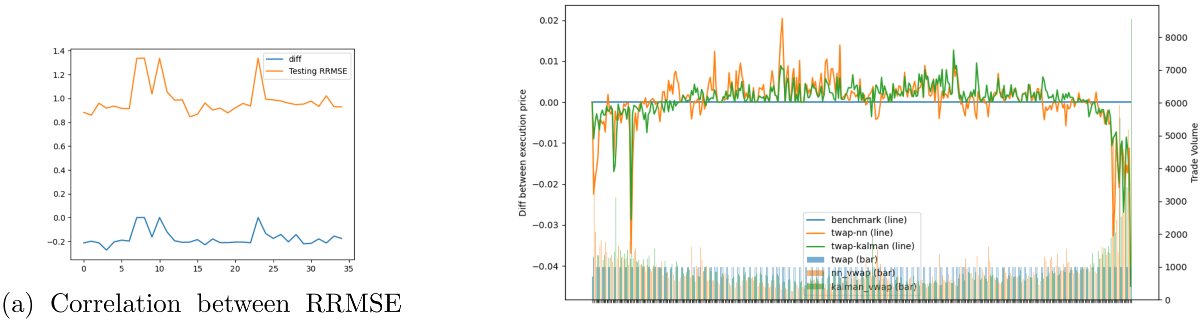
Panel D: Model loss (15min)

Figure 7: QQQ Visualizations

low testing RRMSE (0.867) among all evaluated configurations. With 24,467 training observations, this configuration maintains a healthy data-to-parameter ratio despite its larger parameter count, allowing it to effectively capture complex patterns without overfitting. Interestingly, QQQ benefits from a longer lag structure (50 periods) compared to SPY (10 periods), suggesting that NASDAQ-100 trading volumes exhibit longer-term dependencies. This may reflect the tech-heavy composition of QQQ, where trading patterns could be influenced by more persistent factors than those affecting the broader market represented by SPY.

The learning curve of this model (the top right subplot of Figure 7) exhibits a more volatile training pattern compared to the SPY models. The training loss (blue line) demonstrates steady improvement, declining from 0.0007 to approximately 0.00026 over 35 epochs while the validation loss (orange line) displays considerable fluctuation throughout training, suggesting the model is sensitive to the specific data in each validation batch.

QQQ Analysis (15-minute Frequency): For the QQQ 15-minute trading volume prediction (the bottom subplot of Figure 5), we selected the [5, 5] neuron configuration with 30 lags as optimal, achieving the lowest testing RRMSE (0.534) among all evaluated models. With 1,638 training observations, this compact architecture ensures an appropriate data-to-parameter ratio similar to what we observed with SPY 15-minute data. As with SPY, the 30-lag structure proves effective for QQQ's 15-minute data, capturing es-



(a) Correlation between RRMSE and simulation

(b) Volume and VWAP by three strategies

Figure 8: 1 minute prediction

sential daily cyclical patterns in NASDAQ-100 trading volumes. This consistent finding across both ETFs reinforces that 15-minute trading volume data benefits from incorporating approximately one trading day's worth of historical information, allowing the model to recognize and leverage the characteristic intraday volume patterns common to equity markets.

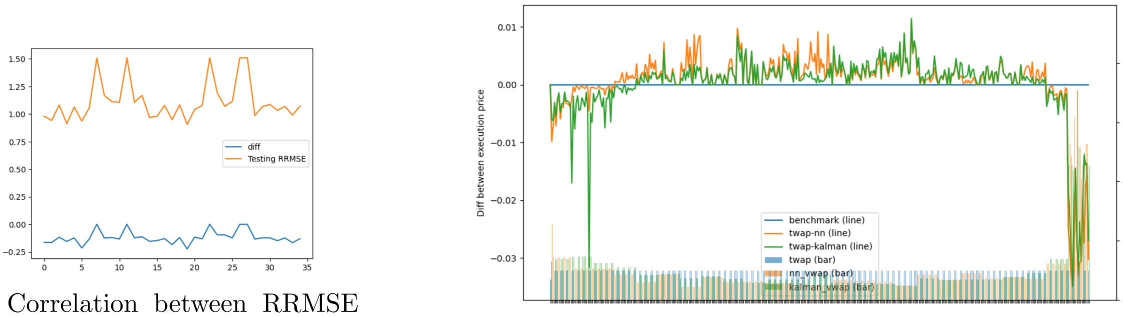
Like the SPY 15 mins model, the learning curve of the QQQ 15 mins model (the top right subplot of Figure 7) shows a similar convergence pattern across 70 epochs. Both training and validation losses decline steadily until epoch 40, after which they stabilize at approximately 0.0016 and 0.0028 respectively. The consistent gap between curves without divergence confirms the model achieves an optimal balance between fitting training data and maintaining generalization capability.

5 Trading Volume and Algorithmic Trading

We leverage the predicted trading volumes to develop an execution algorithm that is benchmarked against a naive TWAP assumption. Specifically, we design a VWAP strategy based on predictions computed at both 1-minute and 15-minute intervals and extend this framework to a Directional VWAP (DVWAP) execution algorithm. The fundamental premise of the VWAP strategy can be explained in the equation below. Historical data are used to estimate the β factor, which calibrates our predicted volumes to the actual executed volumes. Consequently, at the beginning of each trading interval, the target volume to be executed is allocated and subsequently traded via market orders

$$\begin{aligned} V_{actual}^{t-1} &= \hat{\beta} \times V_{predicted}^{t-1}, \\ V_{target}^{t} &= \hat{\beta} \times V_{predicted}^{t}. \end{aligned} \tag{13}$$

The VWAP strategy incorporating 15-minute predictions integrates both VWAP and TWAP methodologies. Within each 15-minute interval, the algorithm follows a similar logic to determine the volume to be executed in that period. To enhance execution stability, the allocated volume is then distributed evenly across each minute using a TWAP approach,



(a) Correlation between RRMSE and simulation

(b) Volume and VWAP by three strategies

Figure 9: 15 minute prediction

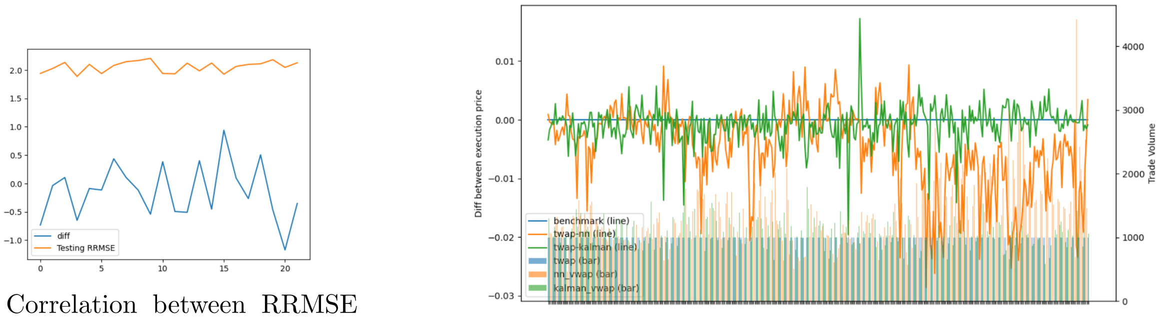
ensuring a smoother execution profile. By adjusting the hyperparameters of the neural network, we generate multiple models with varying statistical metrics. Encouragingly, the performance of these models aligns well with our execution algorithm. Notably, we observe that as the testing error increases, our edge of using VWAP is also diminishing and models with lower errors tend to achieve more favorable execution prices, reinforcing the effectiveness of our approach.

A closer examination of Figure 8 reveals that the executed trade volume exhibits a smile-shaped pattern, consistent with our previous discussion. Similarly, the price difference follows a comparable distribution. At the market open and close, where trading volumes are higher, the execution price tends to be less favorable. Conversely, during mid-session periods, lower executed volumes result in more advantageous pricing, highlighting the trade-off between execution quantity and price efficiency.

A similar pattern is observed with the 15-minute interval, as evident from Figure 9, where we also note a slight performance improvement. This enhancement aligns with the more accurate predictions at the 15-minute frequency. We attribute this improvement to the aggregating effect, which smooths volume volatility by mitigating noise. By offsetting white noise, the model can better capture the underlying systematic patterns, leading to more effective execution.

Besides the slight improvement in overall performance from the 15-minute VWAP strategy, we find it closely resembles the 1-minute strategy, with trading volumes heavily concentrated at the market open and close. However, a key distinction observed in the graph is that, apart from these peak periods, the VWAP strategy tends to consistently outperform the benchmark in the mid-session. This improvement is likely attributed to the smoothing effect, where the 15-minute execution is distributed evenly across 1-minute intervals, reducing short-term fluctuations and contributing to the overall enhancement in performance.

We further investigate whether the DVWAP strategy can enhance overall execution performance. Under the DVWAP framework, predictions for both buy and sell quantities are generated at the beginning of each time interval. The strategy then utilizes the following formula to compute the historical buy-to-sell ratio, serving as a key factor in determining



(a) Correlation between RRMSE and simulation

(b) Volume and VWAP by three strategies

Figure 10: DVWAP prediction

execution allocation

$$
\begin{aligned}
\hat{\alpha}_{imbalance} &= \frac{V_{sell}^{t-1}}{V_{buy}^{t-1}}, \\
V_{target}^{t} &= \hat{\alpha}_{imbalance} \times \frac{V_{pred\_buy}^{t}}{V_{pred\_sell}^{t}} \times b, \\
b &= \frac{V_{total}}{T}.
\end{aligned}
\tag{14}
$$

This implies that if the algorithm predicts higher buy volume, it will execute more buy orders. This decision is driven by two factors: first, a higher predicted buy VWAP suggests increased buying pressure, and second, an anticipation that the market may move upward, leading to less favorable execution prices in the future. If the prediction is accurate and correctly captures the market direction, the strategy has the potential to optimize execution costs and achieve more efficient trade execution.

As evident from Figure 10, we find that the performance of DVWAP is more mixed compared to the straightforward improvements observed in our earlier strategies. Additionally, the relationship between statistical metrics and simulation performance becomes less clear. However, when DVWAP does yield an improvement, the magnitude of this enhancement is significantly larger than that of other strategies. The graph also reveals a fundamental shift in the execution behavior: instead of volume being heavily concentrated at the market open and close, it is now more evenly distributed throughout the trading session.

To better understand the source of this improvement, we analyze the execution patterns in greater details. Figure 10 does not reveal specific instances where the executed price outperforms the benchmark, missing the details where the overall results indicate a significant enhancement. To investigate further, we examine the price chart of the instrument in question, allowing us to gain deeper insights into the underlying factors contributing to this improvement.

Figure 11 illustrates how the DVWAP strategy outperforms the benchmark by strategically executing volume at optimal price levels. In this example, when the market opens and price fluctuations are high, the algorithm refrains from immediate execution and avoids

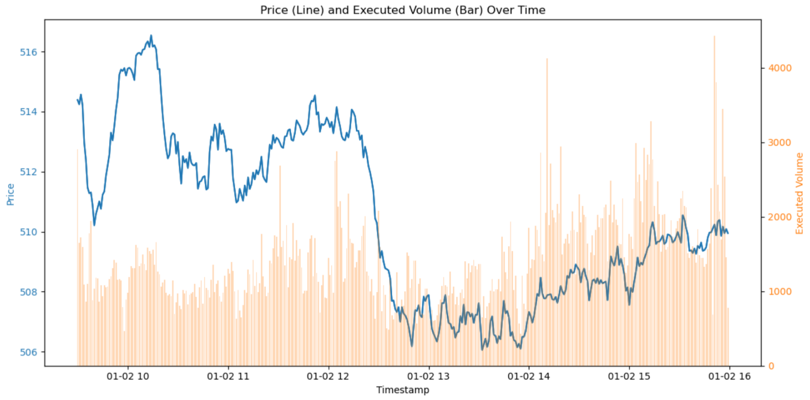


Figure 11: Price movement and executed volume with DVWAP

premature realization, particularly when selling volume surpasses buying, as this signals a potential downward price movement. By delaying execution, the algorithm anticipates a more favorable price later in the session. As the market declines later in the afternoon session, we observe a significant increase in executed volume compared to the morning session. This aligns with a more favorable price environment, allowing the strategy to capitalize on lower prices and convert this price advantage into improved execution performance.

6 Conclusion

In this paper, we analyzed the properties of trade volume and explored various modeling approaches, including the Kalman filter and neural networks, to predict trading volume and achieved robust statistical metrics. Leveraging these predictions, we developed a VWAP-based execution strategy aimed at reducing transaction costs. We further extended this approach to DVWAP and examined its effectiveness in execution strategies.

Our findings indicate that within the VWAP-based framework, utilizing a 15-minute interval improves transaction costs by concentrating execution around the market open and close while achieving better volume smoothing. Additionally, in the DVWAP setup, it proves more effective in capturing market microstructure, providing insights into future price movements, and potentially delivering greater execution improvements. However, the DVWAP-based strategy also introduces significant uncertainty due to increased time risk. To fully capitalize on its advantages, we suggest incorporating additional market microstructure factors to enhance execution robustness and mitigate potential risks.

References

[1] Admati, A. R., & Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, 1(1), 3–40.

[2] Almgren, R., & Chriss, N. (2000). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2), 5–39.

[3] Andersen, T. G., & Bollerslev, T. (1998). Deutsche Mark–Dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance*, 53(1), 219–265.

[4] Bertsimas, D., & Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, 1(1), 1–50.

[5] Biais, B., Hillion, P., & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse. *Journal of Finance*, 50(5), 1655–1689.

[6] Bichri, H., Chergui, A., & Hain, M. (2023). *Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets.* AICSE Lab, Department of Computer Science, ENSAM Casablanca, Morocco.

[7] Bouchaud, J. P., Gefen, Y., Potters, M., & Wyart, M. (2004). Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quantitative Finance*, 4(2), 176–190.

[8] Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., & Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *Journal of Finance*, 69(5), 2045–2084.

[9] Chen, R., Feng, Y., & Palomar, D. (2016). Forecasting intraday trading volume: A Kalman filter approach. *SSRN Electronic Journal.*

[10] Coakley, J. R., & Brown, C. E. (2000). *Artificial neural networks in accounting and finance: Modeling issues.* Intelligent Systems in Accounting, Finance & Management, 9(2), 119–144.

[11] Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937), 267–270.

[12] Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *Journal of Finance*, 66(1), 1–33.

[13] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering – Transactions of the ASME*, 82(1), 35–45.

[14] Ma, S., & Li, P. (2021). Predicting daily trading volume via various hidden states. *arXiv Preprint arXiv:2101.12345*.

[15] Obizhaeva, A., & Wang, J. (2013). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1), 1–32. (Originally circulated as a working paper in 2006)

[16] Wood, R. A., McInish, T. H., & Ord, J. K. (1985). An investigation of transactions data for NYSE stocks. *Journal of Finance*, 40(3), 723–739.

[17] Xu, X., & Zhang, Y. (2023). A high-frequency trading volume prediction model using neural networks. *Decision Analytics Journal*, 7, 100235.