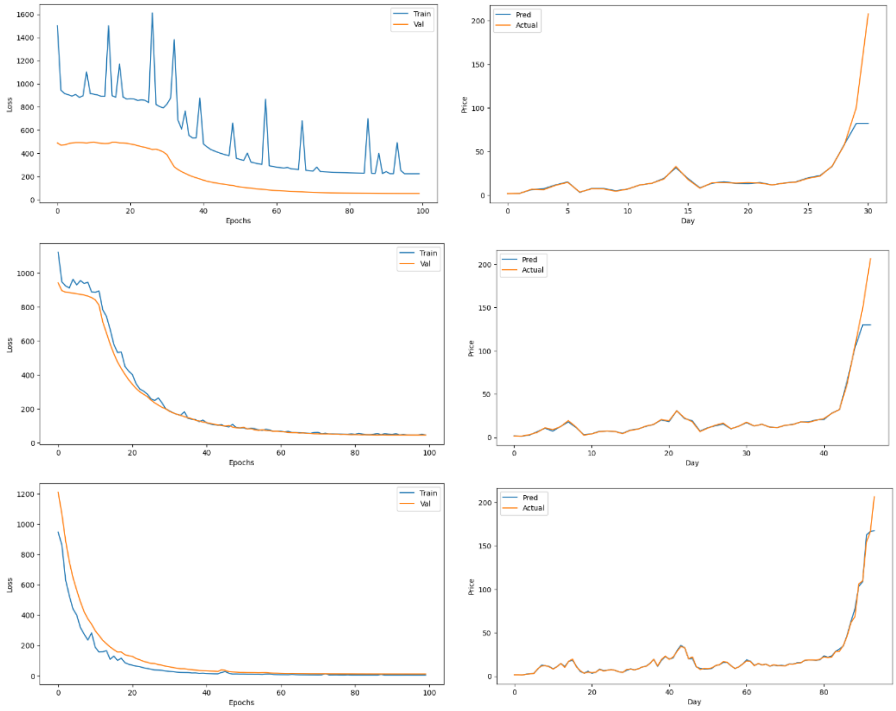
Homework 4

113030511 簡若伃

1. Experiment.

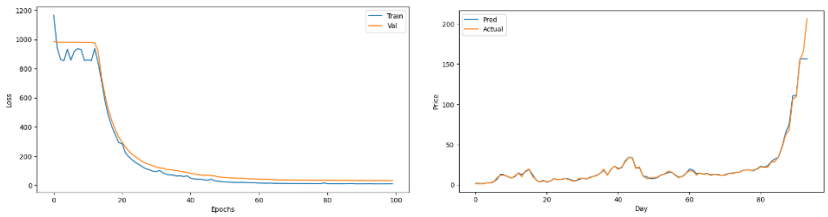
window size	step	MSE
15	15	53.7196
10	10	46.3147
5	5	12.0030



From the results, we observe that smaller window size and step significantly improve model performance. The configuration with a window size of 5 and step size of 5 yields the lowest validation MSE, suggesting that more granular and overlapping sequences help the LSTM model better capture temporal dependencies and further stabilize learning. In contrast, larger windows with larger steps result in higher MSE. Additionally, the learning curves show more stable convergence and better alignment between training and validation loss as the window size and step decreases while the training time may be longer.

2. (i) The validation loss came to 32.9. The learning curves are smooth and convergent, and the prediction curve closely tracks the actual prices, indicating that Volume provides meaningful information for the LSTM to learn market dynamics. Notably, it enhances prediction during high-volatility periods, although minor underestimation remains near price peaks, which is visually observable in final price surge, where the predicted curve lags slightly below the actual peak, indicating conservative prediction behavior near rapid upward shifts. Overall, incorporating Volume helps

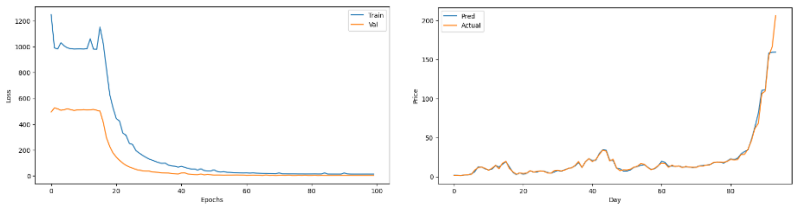
model capture supply-demand behavior, improving trend accuracy and reducing MSE.



(ii) window size=5, step=5

Feature Combination	**MSE**
Open, High, Low, Close	12.0030
Open, High, Low, Close, Volume	32.9267
Open, Close, Volume	4.0484

Picture below is for features: Open, Close, Volume. This combination achieved the lowest validation MSE suggests that including all price features like High and Low may introduce redundancy or noise, while features like Open, Close, and Volume provide a stronger signal-to-noise ratio. In the plots, the model trained with this configuration converges stably, and the predicted prices closely follow actual prices, even during rapid price surges. Volume appears to help the model recognize upcoming price changes, while Open and Close encapsulate essential price dynamics effectively.



3. With and without normalized.

Feature Combination	**without normalized**	**with normalized**
Open, High, Low, Close, Volume	1099.7740	32.9267
Open, Close, Volume	765.9852	4.0484

By applying Min-Max normalization to input features reduced the MSE significantly. Normalization's ability to scale features to a uniform range, mitigating issues like exploding gradients and facilitating faster convergence. Without normalization, the model failed to converge and produced nearly constant predictions regardless of actual price trends. Reflected by the flat prediction line and static validation loss. The result confirms that unscaled features dominate the gradients. Normalization is critical for time series models like LSTM, especially when dealing with multiscale financial features.

4. Window size should be **greater than or equal to** step size. If the step size ≥ window

size, some parts of the time series are skipped, reducing the number of training samples and missing potentially useful patterns. Using step size ≤ window size ensures overlapping windows, which increases training data and captures more sequential dynamics.
Reference: Jason Brownlee - Deep Learning for Time Series Forecasting - Predict the Future with MLPs, CNNs and LSTMs in Python
https://www.inf.u-szeged.hu/~korosig/teach/books/Jason%20Brownlee%20-%20Deep%20Learning%20for%20Time%20Series%20Forecasting%20-%20Predict%20the%20Future%20with%20MLPs,%20CNNs%20and%20LSTMs%20in%20Python%20(2018).pdf

5. Window Warping (WW) randomly selects a window of a time series and speeds it up or slows it down to create new synthetic samples. This technique increases data diversity while preserving essential temporal patterns. The length of the slice will change after deformation, window slicing (WS) will be performed to cut the deformed data into uniform length. Time series are prone to changes in the time axis, such as how fast a person walks. Warping allows the model to ignore this natural variation and focus only on the essential pattern of the data, thereby improving the model's robustness to time distortions.
Reference: Arthur Le Guennec, Simon Malinowski, Romain Tavenard. Data Augmentation for Time Series Classification using Convolutional Neural Networks
https://shs.hal.science/halshs-01357973
6. (i) For convolution-based models, using fixed-size sliding window when the sequence is long or unbounded, allowing scalable and memory-efficient inference. Also we can use global inference, helping the model capture overall trends or long-term dependencies better, but it also requires more memory and computation.
(ii) For recurrent-based models, step by step prediction, using its internal hidden states to remember past information, allows it to handle sequences of arbitrary length without needing a fixed window. And if the model was trained using fixed-size windows inference should also feed inputs with the same window size.
(iii) For transformer-based models, if computational resources allow, the full self-attention mechanism can be fed into the model at once. For very long sequences, inference is done in overlapping chunks or sliding windows to manage memory usage. Overlap is necessary to preserve context across chunks. Transformers are highly flexible with input length but are memory intensive. Thus, balancing window size and memory constraints is crucial during inference.