Learning Predictive Candlestick Patterns: Vision Transformers for Technical Analysis

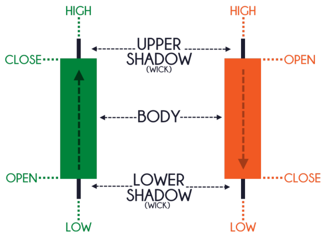

Figure 1. A OHLC Candle

Abstract

Stock market traders have long used technical analysis of candlestick charts to predict price movements, but the predictive value of these charts is contested. In this paper, I examine different frameworks of training ViTs (classification for price prediction and self-supervised learning) to see if transformers are able to learn semantically meaningful patterns. This work demonstrates the efficacy of transformers in price movement forecasting and the value of candlestick charts over numerical OHCL data.

1. Introduction

The candlestick chart, first developed by a Japanese rice trader in the 18th century, have long been used to visually represent price data for financial assets. Each candlestick encodes four critical price points—open, high, low, and close—within a specific time period, creating visual patterns that proponents believe reveal market psychology and predict future price movements.

Despite widespread adoption, the predictive validity of candlestick patterns remains highly debated. Efficient market hypothesis suggests that all available information is already reflected in current prices, making technical analysis futile. However, behavioral finance argues that recurring psychological patterns create exploitable inefficiencies. Previous empirical studies have yielded mixed results, often limited by subjective pattern definitions and small sample sizes.

Recent advances in computer vision, particularly Vision Transformers (Dosovitskiy et al., 2020), offer an unprecedented opportunity to objectively test these claims. Unlike traditional approaches that rely on hand-crafted rules, deep learning can discover patterns directly from data, unblemished by human psychology. Furthermore, self-supervised learning techniques like Masked Autoencoders (He et al., 2022) enable models to learn meaningful representations without extensive labeled data.

In this paper,I investigate whether ViTs can learn meaningful patterns from candlestick charts that predict future price movements. I use two different fine-tuning approaches for ImageNet-pretrained models: (1) multi-class classification to directly predict price movements and (2) self-supervised finetuning used masked autoencoders (MAE), where learned patterns are then clustered and backtested. I find that transformers are able to learn effective representations of the market through candlestick charts, with my models showing impressive results given their compute limitations.

2. Literature Review

Existing literature on the application of computer vision techniques for candlestick charts is limited, partly because of the general consensus that visual data is noisier and inherently inferior to numerical OHLC data in regards to time-series forecasting.

1. Chen, 2025: This paper utilizes vision transformers as a baseline for the task of price forecasting from candlestick data. It's useful as it shows the relatively high accuracy of ViT on this task, especially compared to CNNs. While the multi-modal LLM used outperformed the ViT, this doesn't account for the massive disparity in training costs between the two – my paper compares two vision transformers trained with the same dataset and same number of epochs.

2. Kusuma et al., 2020: This paper used a number of CNN architectures for the task of price forecasting from candlestick data. It indicated promising results and served as a useful point of comparison for my models.

3. Huang, 2024: This former 231N project explores training an object-detection model to detect and classify candlestick patterns on a chart. While the task strongly differs from my project, the pipeline for the creation of candlestick charts from numerical data served as inspiration for my own.

As you can see, the existing literature is extremely sparse – my work is unique in it's use of vision transformers for interpreting candlestick charts and particularly the idea of interpreting learned chart representations via SSL.

3. Dataset

I constructed a dataset of 50,000 224 x 224 candlestick chart snapshots across 10 different U.S. equities, including index-funds, individual stocks, and leveraged ETFs.

I first collected 5-minute OHLC data using the Polygon.io API. The data spans January 2020 to June 2025 and includes only regular market hours (9:30am–4:00pm). Each snapshot covers 30 5-minute candles (2.5 hours of trading history). Importantly, each snapshot is a continguous trading period to allow the models to focus on learning intraday patterns. For the classification task, I additionally included a 3-class label based on the closing price 25 minutes after the chart snapshot. If the price was more than 0.5% it was an ”Up” label, if it was more than -0.5% it was ”down”, and otherwise it was labeled ”flat”. The class distribution was roughly 80/10/10 for flat/up/down, but I left this untouched as it most accurately represented market conditions.

The images were then generated via mplfinance and cleaned of axes, grids, and text to focus purely on patterns. All images use green up candles and red down candles. I used a 70/15/15 Train/Val/Test split.

4. Methods

I formulate candlestick pattern recognition as a multi-class image classification problem, where given an input candlestick chart image $I \in \mathbb{R}^{H\times W\times 3}$,I aim to predict the probability distribution over three classes representing future price movements: $\{\text{Down}, \text{Flat}, \text{Up}\}$. My approach leverages Vision Transformers (ViTs), which have demonstrated superior performance in image classification tasks by treating images as sequences of patches and applying self-attention mechanisms to capture global dependencies.

4.1 Vision Transformer Architecture

I employ ViT-Tiny (Dosovitskiy et al., 2020) as my base architecture, which divides each input image into a grid of non-overlapping patches. Specifically, for an input image of size 224×224 pixels,I use patches of size $P = 16$, resulting in $N = (H/P) \times (W/P) = 196$ patches. Each patch $x_p \in \mathbb{R}^{P^2\cdot 3}$ is flattened and linearly projected to an embedding dimension $d = 192$ through a learnable linear projection $E \in \mathbb{R}^{(P^2\cdot 3)\times d}$:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{\text{pos}}$$

where x_{class} is a learnable classification token prepended to the sequence, and $E_{\text{pos}} \in \mathbb{R}^{(N+1)\times d}$ represents learnable position embeddings added to retain positional information. The resulting sequence is then processed through $L = 12$ transformer encoder blocks, each consisting of multi-head self-attention (MSA) and MLP blocks with residual connections and layer normalization:

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}$$
$$z_\ell = \text{MLP}(\text{LN}(z'_\ell)) + z'_\ell$$

The multi-head self-attention mechanism with $h = 3$ heads enables the model to attend to different positions simultaneously, learning complex relationships between different regions of the candlestick chart.

4.2 Training Strategies

Both training approaches leverage transfer learning by initializing the model with weights pretrained on ImageNet-1k, which contains 1.2 million natural images across 1,000 categories. While candlestick charts differ substantially from natural images, the pretrained model has already learned fundamental visual features such as edge detection, shape recognition, and hierarchical feature composition.

For my first approach, I simply took the described transformer architecture and attached a classification layer. The classification is performed using the transformed classification token z_L^0 through a linear classifier head:

$$y = \text{LN}(z_L^0) W_c$$

where $W_c \in \mathbb{R}^{d\times 3}$.

This approach was motivated by the idea that the best way to learn semantically meaningful representations is to penalize when those representations are not predictive.

4.3 Self-Supervised Pretraining with Masked Autoencoders

My second approach employs MAE, a form of self-supervised learning, on unlabeled candlestick chart data. Unlike the other model, there is no direct classification objective. Instead, the MAE framework learns meaningful representations by reconstructing randomly masked portions of input images, forcing the model to develop a rich understanding of candlestick patterns. I randomly mask $m = 75\%$ of image patches and train the model to reconstruct the original image. The MAE architecture consists of an encoder f_θ that processes only visible patches and a lightweight decoder g_ϕ that reconstructs the full image from the encoded representation.

The masking process follows a random sampling strategy where, for each image,I sample a random permutation π of patch indices and keep only the first $\lceil N(1-m)\rceil$ patches. The encoder processes these visible patches through the standard ViT architecture, while masked patches are replaced with learnable mask tokens before being processed by the decoder. The reconstruction loss is computed only on masked patches:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|M|}\sum_{i\in M}\|x_i - \hat{x}_i\|^2$$

where M denotes the set of masked patch indices, x_i is the original patch, and $\hat{x}_i$ is the reconstruction. This objective encourages the model to learn rich representations that capture the essential structure of candlestick patterns, as accurate reconstruction requires understanding both local patterns (individual candle shapes) and global context (trend formations).

4.4 Implementation Details

All models are trained using the AdamW optimizer (Loshchilov & Hutter, 2019) with weight decay $\lambda = 0.05$ to prevent overfitting. For both models,I finetuned over 30 epochs with batch size set to 128 for optimal GPU utilization on a NVIDIA L4 VM. Training took approximately 2 hours per model on the VMs.

Data augmentation is kept minimal to preserve the integrity of candlestick patterns. I apply only slight horizontal shifts ($\pm 10\%$ of image width) to simulate different time window selections, avoiding rotations or color distortions that would destroy the semantic meaning of price movements. All images are normalized using ImageNet statistics (mean = $[0.485, 0.456, 0.406]$, std = $[0.229, 0.224, 0.225]$) to maintain compatibility with pretrained weights.

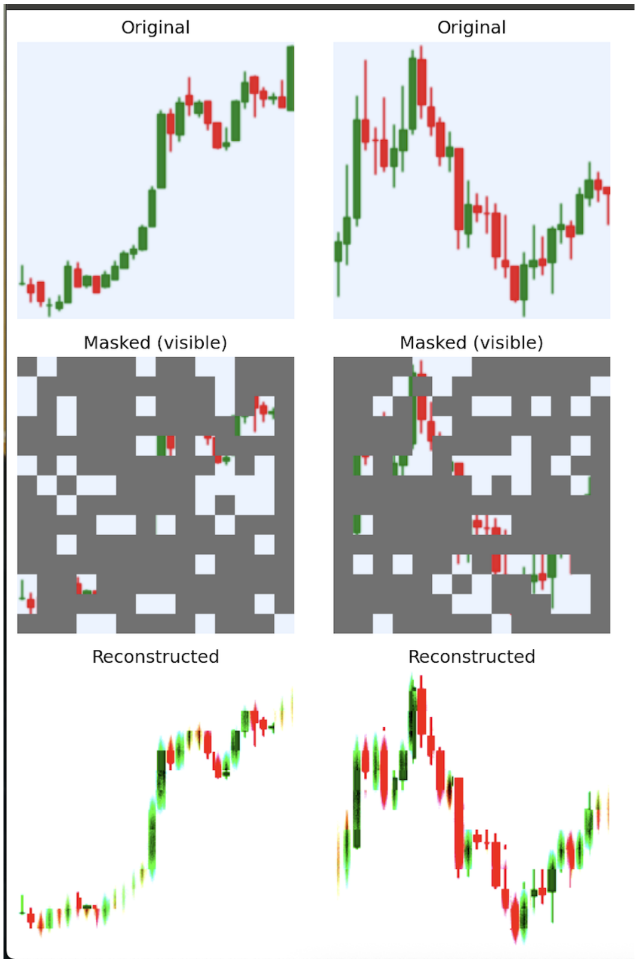

Figure 2. Example of MAE Reconstruction

4.5 Interpretability Analysis

To understand what patterns the MAE model learned, I perform clustering analysis on the learned representations to discover groups of similar patterns.I extract feature vectors from the encoder's output (the transformed classification token) for all test samples and apply K-means clustering with $k = 20$ clusters. For each cluster,I analyze the distribution of true labels to assess predictiveness, calculating the cluster purity as:

$$p_c = \max(p_{\text{down}}, p_{\text{flat}}, p_{\text{up}})$$

where p_y represents the proportion of samples with label y in cluster c. Clusters with high purity indicate that the model has learned to group patterns with similar predictive outcomes, validating the existence of meaningful visual patterns in candlestick charts.

5. Experiments

5.1. Baselines

We used two non-vision baselines, to represent strategies that are more commonly used in the financial world for price forecasting.

(a) RSI Statistical Strategy: Uses common statistical indicators to identify potential trend reversals and confirm overbought/oversold conditions. Specifically, RSI (Relative Strength Index) is the magnitude of recent price changes and is widely used in fundamental analysis in the finance world. In this study, I considered oversold conditions to predict an upward movement and overbought to predict downwards movement (neutral conditions were considered to be a ”flat” prediction).

(b) A RNN with LSTM architecture trained for time-series prediction. It employs three hidden layers with 128, 64, and 32 units respectively, dropout rates of 0.2 between layers, and tanh activation functions.

5.2. Results

Model	Flat	Up	Down
RSI Strategy	0.32	0.13	0.16
RNN	0.38	0.05	0.03
ViT Classifier	0.81	0.32	0.39
ViT MAE Clustered	0.86	0.47	0.42

Table 1. Class-wise accuracy for different models predicting asset price movement

N Samples	Dominant Class	Purity %	Accuracy
2271	Flat	90.84	90.51
2782	Flat	89.22	88.89
2782	Flat	88.17	87.84
3153	Flat	88.07	87.74
1783	Flat	88.05	87.72
2569	Flat	87.70	87.37
2949	Flat	87.66	87.33
3206	Flat	87.52	87.19
2764	Flat	87.05	86.72
3590	Flat	87.02	86.69

Table 2. Top-10 Performing MAE Clusters

6. Conclusion

This work demonstrates that Vision Transformers can successfully learn predictive patterns from candlestick charts, achieving high accuracy on short-term price movement prediction—far exceeding numerical RNNs and traditional technical indicators. Our key findings:

1. Candlestick patterns may contain genuine predictive signal, validating their use in technical analysis
2. Self-supervision leads to highly effective clustering with substantial increase in predictive power over the multi-class classifier. 3. MAE gives transformers a strong ability to reconstruct candlestick charts.

6.1. Limitations

:

1. Limited to liquid stocks in normal market conditions
2. Does not incorporate volume or fundamental data
3. Restricted to intraday movement using a 2.5 hour window and 25 minute prediction window. 4. Most importantly: our models did much better at predicting no movement than they did up/down.

6.2. Future Work

:

1.Extend to multi-modal models incorporating news sentiment 2. Test on other asset classes (forex, cryptocurrencies, commodities) 3. Develop real-time trading system with proper risk management 4. Investigate longer-term patterns with daily/weekly candles

7. Works Cited

(a) Huang, (2022): ”Fast Candlestick Patterns Detection... Using RGB Granular Angular Field and YOLO-Lite- V1”. CS 231N Project, Spring 2022.

(b) Chen, Q. (2025). Image-Driven Stock Price Prediction with LLaMA: A Prompt-Based Approach. International Journal of Modeling and Optimization, 15(1), 17–24. https://www.ijmo.org/vol15/IJMO-V15N1-867.pdf

(c) Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

(d) He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000-16009).

(e) Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650-9660).