Weak Signals, Broad Attention: A CNN Interpretability Study of Stock Candlestick Charts

Abstract

Historically, traders have read candlestick charts by eye while assuming that past prices may hint at future movement; yet decades of evidence suggest that short-term price changes are difficult to predict. We ask not whether a machine can reliably predict the market from a chart, but where such a machine looks when it tries. We converted daily price data from 30 large technology and growth stocks from 2014–2024 into 30-day grayscale candlestick images, labeled by whether the following ten trading days produced a large volatility-adjusted move. The convolutional neural network performed only slightly above chance on the test set (ROC AUC 0.559, 95% CI 0.537 to 0.581). Averaging occlusion importance across test images showed that the model did not concentrate on a small set of visually salient patterns; instead, importance was distributed across most of the chart, with slightly greater weight on the most recent trading days. This weak but broadly distributed signal suggests that candlestick information may be more useful when combined with complementary inputs than when used in isolation.

1 Introduction

Candlestick charts compress open, high, low, and close prices into a visual sequence that traders use to search for market patterns. Whether those patterns reliably predict future prices is debated. The efficient-market and random-walk literature argues that short-term price changes are difficult to forecast from past prices alone (Fama, 1970; Malkiel, 2003), while statistical studies of technical analysis show that chart patterns can at least be tested systematically rather than treated only as trader intuition (Lo, Mamaysky, and Wang, 2000).

Recent deep-learning work extends this idea by turning price histories into chart images for convolutional neural networks (Sezer and Ozbayoglu, 2019; Kusuma et al., 2019). These studies suggest that CNNs can be applied to financial chart data, but prediction accuracy alone does not explain what visual information the model uses. Our study therefore focuses on interpretability rather than trading performance: we ask where a CNN looks in a candlestick chart when predicting future movement. Using Grad-CAM and occlusion sensitivity, which identify image regions that influence CNN predictions (Zeiler and Fergus, 2014; Selvaraju et al., 2017), we test whether importance is concentrated in specific temporal regions of a 30-day candlestick chart or distributed across the full window.

2 Methods

2.1 Data Construction and Labeling

Our unit of analysis is a single window of 30 trading days: the response is its binary label, and the only predictor is the rendered chart image. We pulled daily open, high, low, and close prices for 30 large U.S. technology and growth stocks from Yahoo Finance for 2014 to 2024 (Appendix Table 1). We focused on technology and growth stocks to keep the sample within a relatively comparable group of companies, while recognizing that this limits generalization beyond that sector. We moved the window forward five trading days at a time, and graphed each one as a 224×224 grayscale candlestick image with axes and gridlines removed, so the model sees only candles rather than chart annotations or scale markings. The five-day step increases sample size but creates overlapping windows.

We labeled each window by its future movement relative to its own recent volatility, not a fixed percentage, since a fixed cutoff treats calm and volatile stocks alike. Recent volatility was estimated as the standard deviation of daily returns within the window and scaled to a ten-day horizon by $\sqrt{10}$. Writing the scaled threshold as τ and the absolute ten-day close-to-close forward return as r, we labeled a window *significant* if $r \geq 1.30\,\tau$, *not_significant* if $r \leq 0.90\,\tau$, and dropped it otherwise. The dropped middle band gives cleaner labels but means the classes are not exhaustive, so our results describe this separated classification problem.

To reduce temporal leakage, splits were assigned by window start date rather than randomly. Windows starting on or before December 31, 2018 were assigned to training (4,073 *not_significant*, 1,318 *significant*), windows starting from January 1, 2019 through February 19, 2020 were assigned to validation (1,129 validation windows), and remaining later windows were assigned to testing (3,606 total test windows). We removed any window whose chart period or future label date overlapped February 20, 2020 through December 31, 2021, because COVID-era market behavior

was unusually volatile and atypical.

2.2 Model Training, Evaluation, and Interpretation

Before training, images were converted to grayscale within the model pipeline so the CNN would focus on candle structure and chart shape rather than red-green color cues. This choice was supported by a saturation analysis showing nearly identical color distributions across classes (Appendix Figure 1). We also applied mild contrast and dilation augmentation during training to reduce dependence on brightness differences or very thin candle lines.

The classifier is a convolutional neural network with four convolutional blocks, batch normalization, L_2 regularization, dropout, global average pooling, and a final sigmoid output. Because only about 24% of images were labeled *significant*, a naive classifier could achieve about 76% accuracy by always predicting *not_significant*. To address this imbalance, we trained with balanced sampling and binary focal cross-entropy loss (Lin et al., 2017), which gives more weight to difficult or minority-class examples. Model selection used the validation set only: early stopping was based on validation PR AUC, and the final decision threshold was chosen on validation F1 score before being applied unchanged to the test set.

We evaluated performance using ROC AUC and PR AUC. PR AUC was emphasized because it is more informative under class imbalance: it measures how well the model identifies the relatively rare *significant* windows without being dominated by the majority class. Finally, we used Grad-CAM (Selvaraju et al., 2017) and occlusion sensitivity (Zeiler and Fergus, 2014) to examine which image regions influenced the fitted model. Grad-CAM gives local heatmaps for individual predictions, while occlusion sensitivity directly measures the change in predicted probability after masking parts of the chart. We averaged occlusion importance across all test images to estimate whether the model's reliance was concentrated in particular temporal regions or spread across the 30-day window.

3 Results

3.1 Predictive Performance

On the test set, the model reached 55% accuracy, ROC AUC 0.559, and PR AUC 0.276, compared with a positive-class baseline of 0.24 for PR AUC (Appendix Table 2). For the *significant* class, recall was 0.52 and precision was 0.27, meaning the model identified about half of the true large-move windows but produced many false positives. Accuracy is therefore less informative than PR AUC in this setting, since a classifier that always predicted *not_significant* would reach about 76% accuracy.

To quantify sampling variability, we used a nonparametric bootstrap over the test windows with 2,000 resamples (Appendix Figure 2). The 95% interval was $[0.537, 0.581]$ for ROC AUC and $[0.256, 0.300]$ for PR AUC. Both intervals lie above their corresponding chance baselines, 0.50 for ROC AUC and 0.24 for PR AUC, suggesting weak above-baseline performance. However, because adjacent windows overlap and stocks are correlated, these intervals likely understate uncertainty. We therefore interpret the result as evidence of a small predictive signal rather than useful standalone predictive skill.

3.2 Temporal Importance Across the Chart

To summarize which temporal regions influenced the model, we split each 30-day chart into three ten-day vertical strips along the time axis, with the oldest days on the left and newest days on the right. We occluded one strip at a time and recorded the average drop in predicted probability across all test windows (Appendix Figure 3). Importance was broadly distributed: the three average occlusion scores fell in a narrow range from 0.0731 to 0.0848. The most recent third of the chart, days 21 to 30, had the highest average importance score (0.0848), while the first two-thirds were slightly lower and similar to each other: 0.0750 for days 1 to 10 and 0.0731 for days 11 to 20.

Individual Grad-CAM and occlusion examples showed the same broad pattern rather than isolated hot spots (Appendix Figures 4, 5, 6, and 7). Thus, the model did not appear to rely on a single visually salient chart region. Instead, its limited predictive signal was spread across most of the 30-day window, with only a mild tilt toward the most recent trading days.

4 Discussion and Conclusion

The results support a modest conclusion: the CNN detected a weak signal in candlestick-chart images, but not one strong enough to suggest useful standalone prediction. Test performance was only slightly above baseline, with ROC AUC 0.559 and PR AUC 0.276 compared with a 0.24 positive-class baseline. Because precision for the *significant* class was only 0.27, many predicted large-move windows were false positives. These results are consistent with the efficient-market and random-walk literature, which suggests that short-term price movements are difficult to predict from past prices alone.

The interpretability results answer our main research question more directly. Rather than relying on a small visually salient region, the CNN's occlusion importance was distributed across the full 30-day chart, with only a mild tilt toward the most recent ten trading days. This suggests that the model did not learn a sharp candlestick pattern in one part of the image. Instead, whatever signal it found appears faint and broadly distributed. We cannot claim that this matches human chart reading, because we did not collect eye-tracking or human decision data. The safer conclusion is that the CNN relied on a broad chart structure rather than a localized pattern.

This finding helps clarify the role of chart images in prediction. Prior CNN-based chart studies show that price histories can be represented as images for deep learning, but predictive performance alone does not reveal what visual information the model uses. In our setting, candlestick images alone appear to carry little predictive information, though a weak signal spread across the chart could still be useful alongside stronger inputs such as numerical price, volume, or text-based market information. Several limitations affect this conclusion. First, neighboring windows share up to 25 of their 30 days, and many technology stocks move together, so the observations are not independent and the bootstrap likely understates uncertainty; future work could use block bootstrap methods or stock-level/time-blocked resampling. Second, the same 30 tickers appear in every split, so the test set measures new time periods for familiar stocks rather than generalization to new companies; future studies should test on a broader market sample. Third, the sample is concentrated in technology and growth stocks, the window length is fixed, and removing the middle band creates a cleaner but less realistic classification problem. Within these limits, the most useful next step is to test whether this weak image signal adds value after stronger numerical or text-based features are included.

Statement on Generative AI Use

We used a generative AI tool (ChatGPT) in a support role only: to fix grammar and wording, to help debug code, to write simple starter scripts that we then rewrote for our own data, and to explain statistics ideas and possible methods that we tested ourselves. We made all of the modeling, analysis, and writing decisions, and we checked them. None of the results, figures, or numbers in this report were produced by AI.

References

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417.

Kusuma, R. M. I., Ho, T.-T., Kao, W.-C., Ou, Y.-Y., & Hua, K.-L. (2019). Using deep learning neural networks and candlestick chart representation to predict stock market. *arXiv preprint* arXiv:1903.12258.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance, 55*(4), 1705–1765.

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives, 17*(1), 59–82. https://doi.org/10.1257/089533003321164958

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Sezer, O. B., & Ozbayoglu, A. M. (2019). Financial trading model with stock bar chart image time series with deep convolutional neural networks. *arXiv preprint* arXiv:1903.04610. https://arxiv.org/abs/1903.04610

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, LNCS vol. 8689, 818–833. Springer. https://doi.org/10.1007/978-3-319-10590-1_53

Appendix

Ticker	Company	Ticker	Company
AAPL	Apple	MSFT	Microsoft
GOOGL	Alphabet	AMZN	Amazon
META	Meta Platforms	NVDA	NVIDIA
TSLA	Tesla	ADBE	Adobe
CRM	Salesforce	INTC	Intel
AMD	Advanced Micro Devices	ORCL	Oracle
CSCO	Cisco Systems	QCOM	Qualcomm
IBM	IBM	NFLX	Netflix
AVGO	Broadcom	TXN	Texas Instruments
UBER	Uber Technologies	PYPL	PayPal
NOW	ServiceNow	SHOP	Shopify
SNOW	Snowflake	PLTR	Palantir Technologies
MU	Micron Technology	AMAT	Applied Materials
LRCX	Lam Research	ADI	Analog Devices
PANW	Palo Alto Networks	CRWD	CrowdStrike

Table 1: Stocks used in the analysis.

Confusion matrix

	Actual sig.	Actual not
Pred. sig.	450	1,200
Pred. not	419	1,537

Summary metrics

Metric	Value
Accuracy	0.55
ROC AUC	0.56
PR AUC	0.28
PR baseline	0.24
Recall	0.52
Precision	0.27

Table 2: Test-set performance of the CNN.

Note. Results are based on 3,606 test windows. The classification threshold was selected on the validation set and then held fixed for testing.

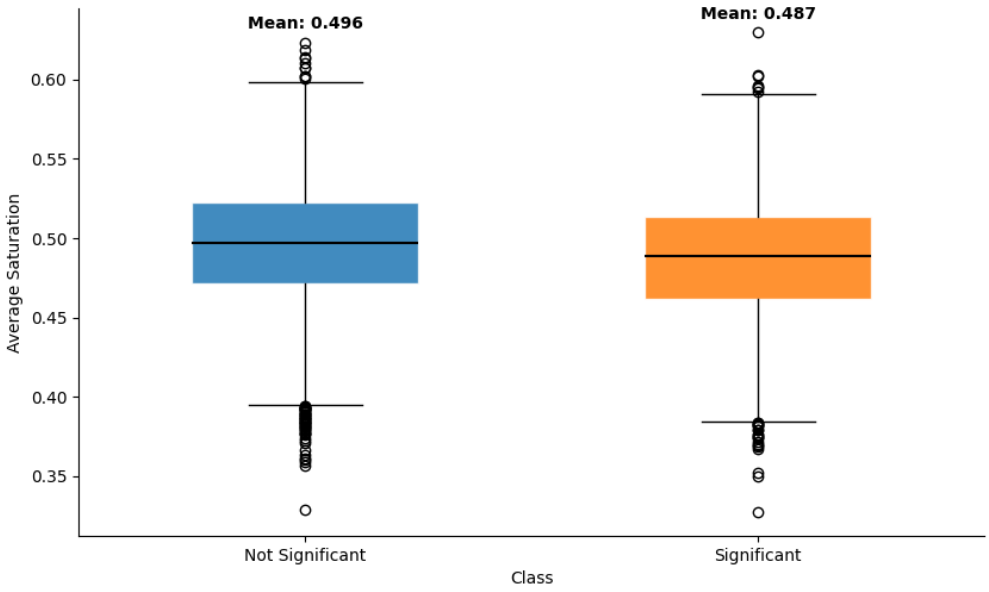

Figure 1: Color saturation by class

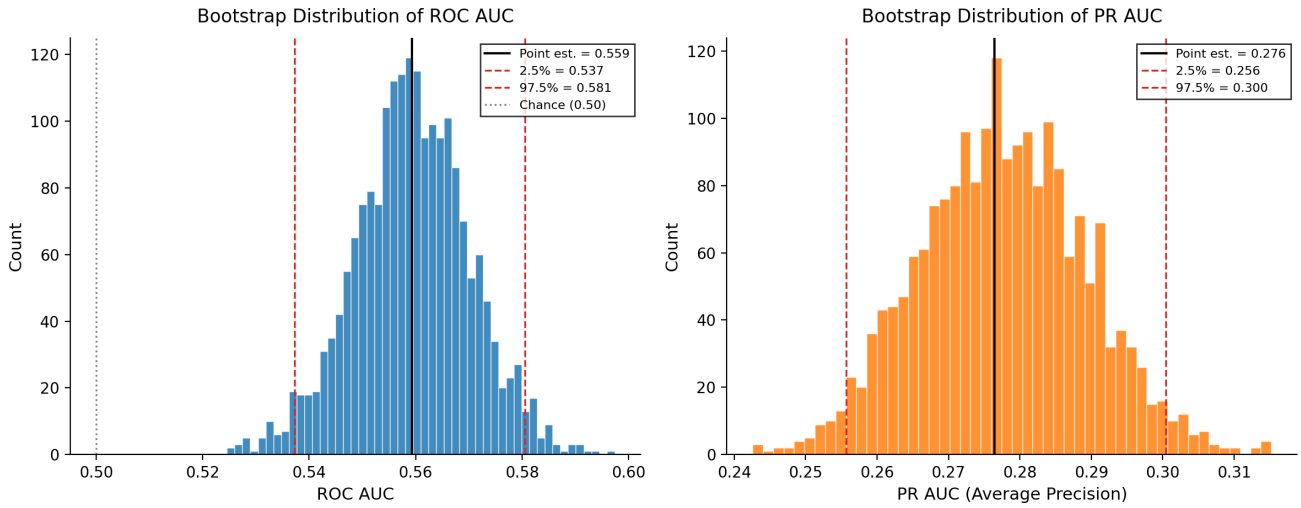

Figure 2: Bootstrap Distributions of ROC AUC and PR AUC

Note. Distributions are based on 2,000 test-set resamples. Dashed lines mark the 95% intervals, and dotted lines mark chance-level performance.

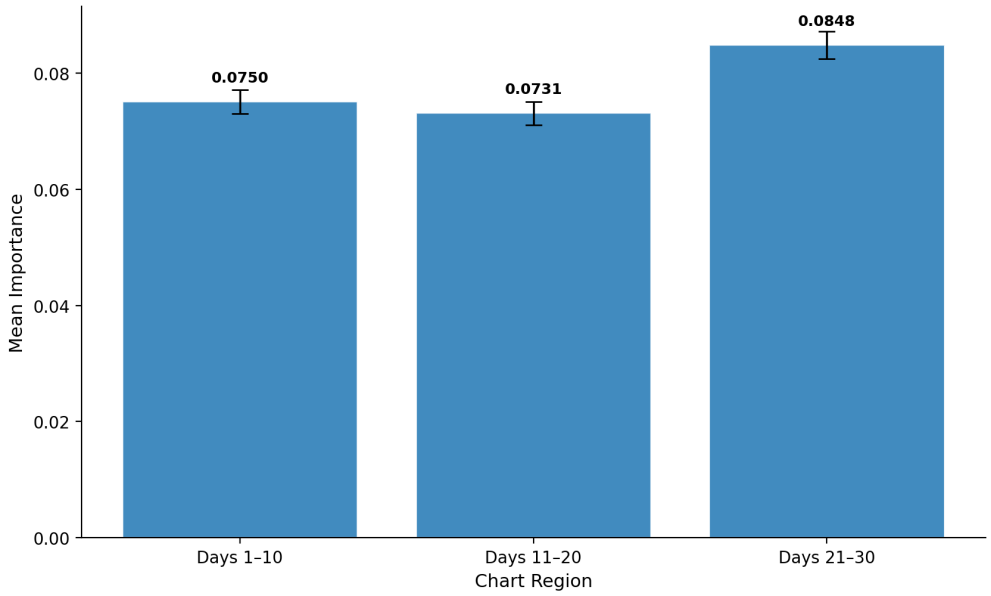

Figure 3: Average Occlusion Importance by Time Strip

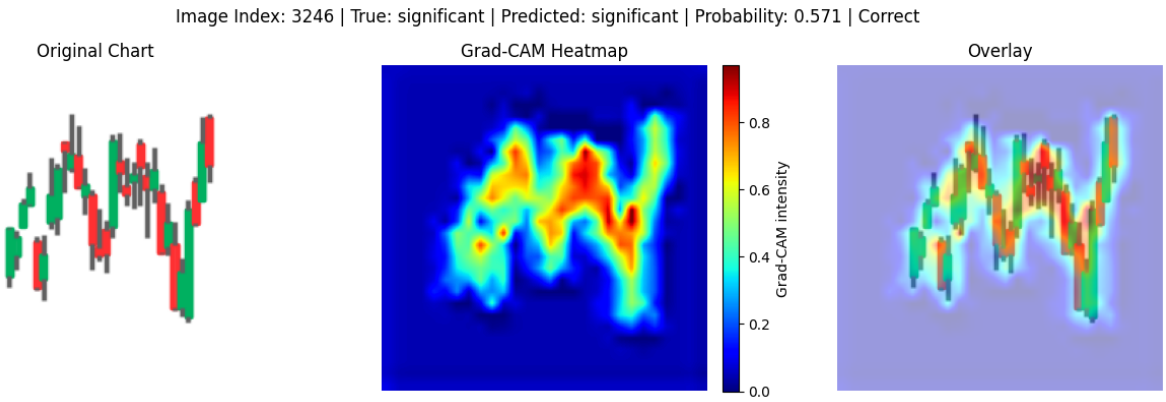

Figure 4: Grad-CAM attention map for a correctly classified test window

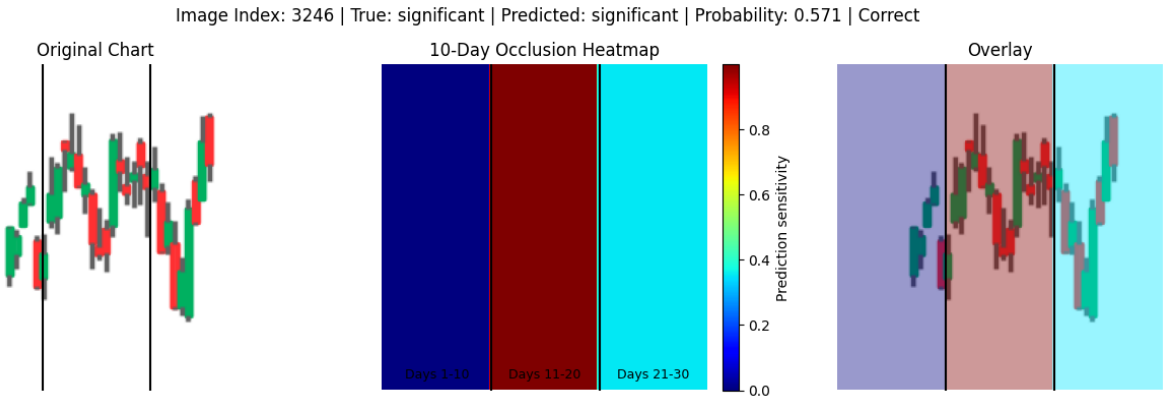

Figure 5: Occlusion sensitivity map for a correctly classified test window

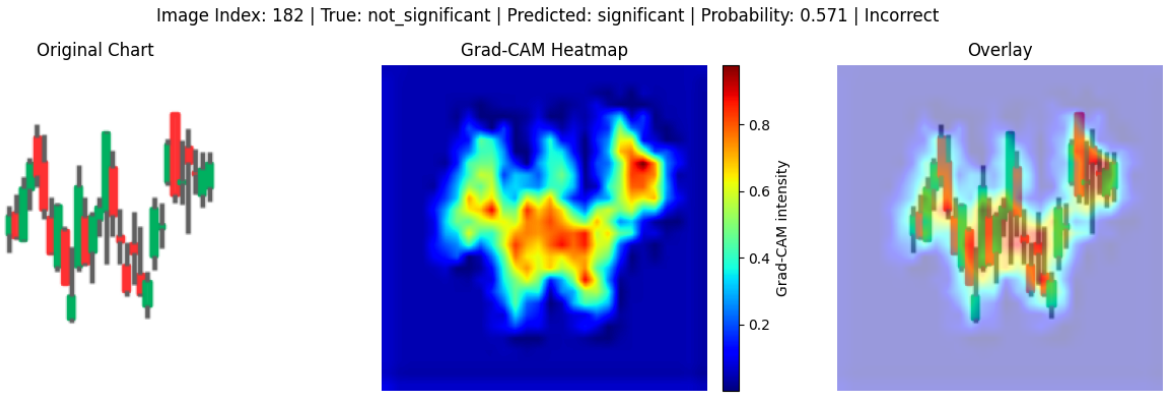

Figure 6: Grad-CAM attention map for an incorrectly classified test window

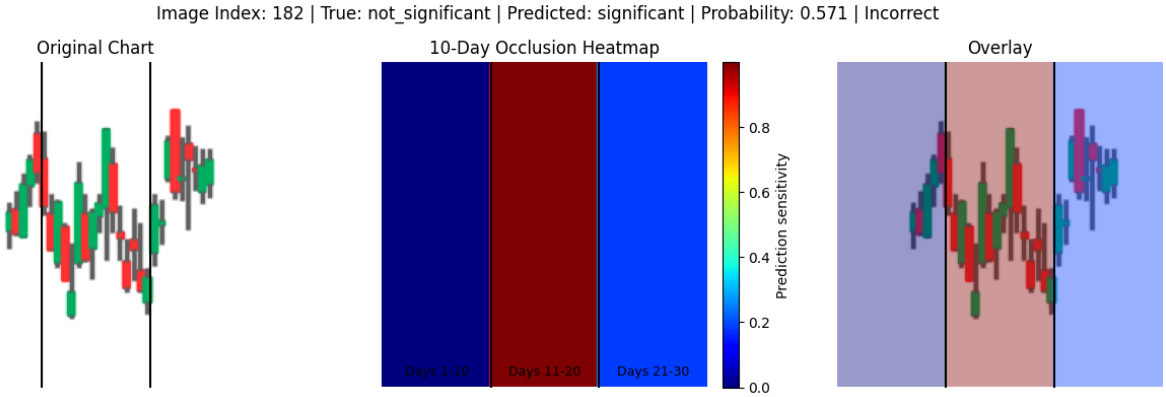

Figure 7: Occlusion sensitivity map Example for an incorrectly classified test window

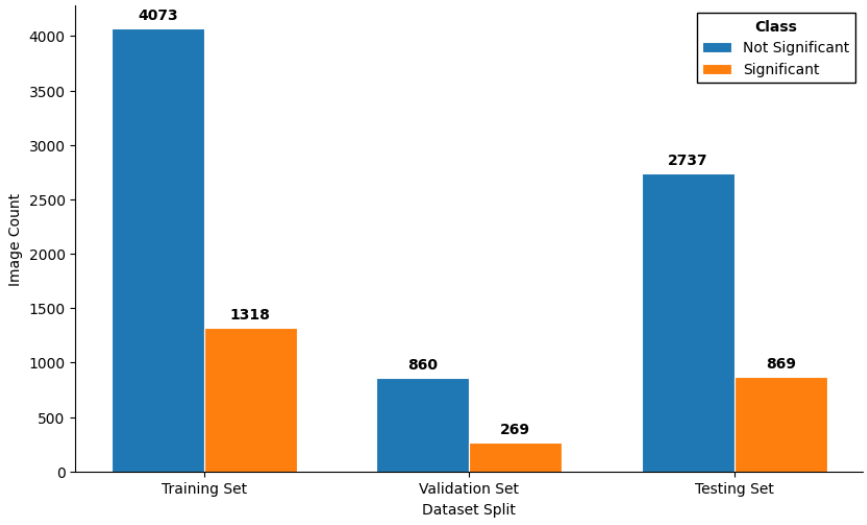

Figure 8: Class Distribution by Split

Note. The *significant* class makes up about 24% of the data, motivating the use of balanced sampling during training.

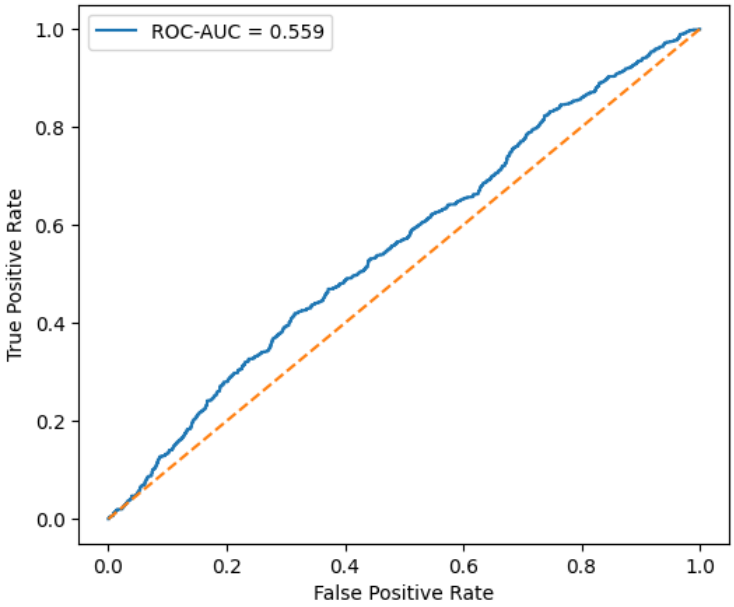

Figure 9: Test ROC curve.

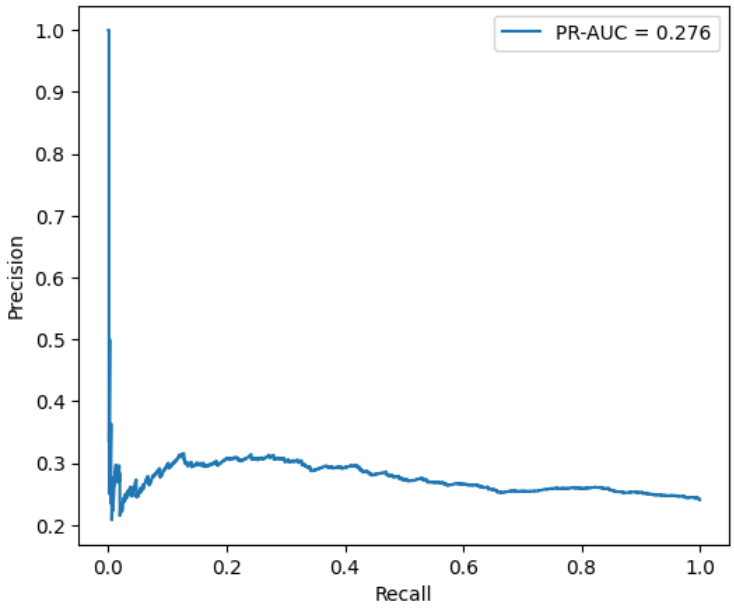

Figure 10: Test Precision-Recall curve.