Can Computer Vision Models Generate Better Trading Signals than Time Series Models?

Shravan Khunti
NetID: ssk10036
MS in Data Science
NYU Center for Data Science

Yash Jadhav
NetID: yj3076
MS in Data Science
NYU Center for Data Science

GitHub Repository

https://github.com/shravanxd/robo-advisors-final-project

Course Information

Course: Robo Advisors & Systematic Trading
Professor: Vasant Dhar

1 Motivation

This project was inspired by a lecture from Professor Dhar, where he briefly discussed converting time series data into images for model input a concept that resonated with Shravan's prior research internship at the National University of Singapore. There, he explored applying deep learning to time series data but had only used 1D convolutional layers on raw series, without transforming them into true visual formats. We found the idea of leveraging actual computer vision pipelines for financial signal generation both unconventional and promising.

Drawing parallels with reinforcement learning applications in decision making systems like self driving cars, we plan to integrate an RL layer as a future extension of this project. This RL component will evaluate how well such agents can act on signals generated by both Time Series and Computer Vision models. Our goal is not only to compare performance but also to understand execution behavior under hybrid setups, while remaining mindful of modeling challenges like data leakage, overfitting, and class imbalance. The current phase focuses on the direct comparison between traditional time series approaches and computer vision techniques, with the RL implementation planned for the subsequent phase.

2 Project Overview

This capstone explores a multi-phase approach to algorithmic trading signal generation. The current implementation focuses on comparing two modeling pipelines: traditional Time Series models and Computer Vision models using Gramian Angular Fields (GAF). Our hypothesis is that GAF-based vision models may extract structural patterns that conventional time series approaches miss, potentially leading to better trading signal quality.

In the planned future phase, we will introduce a Reinforcement Learning (RL) trading agent to assess downstream decision quality and simulate trade execution based on signals from both model types. This three-tiered approach (Time Series → Computer Vision → Reinforcement Learning) will allow us to systematically evaluate whether visual representation of financial data offers meaningful advantages over traditional numerical features, and how these advantages translate into actual trading decisions when implemented through an RL framework.

3 Implementation

3.1 Cell 1: Project Initialization and Library Imports

This cell prepares the runtime environment by importing all the necessary libraries. It includes:

- Core packages for numerical processing (NumPy, Pandas) and visualization (Matplotlib, Seaborn).
- Preprocessing and modeling tools from Scikit learn and TensorFlow/Keras, which will support both traditional and deep learning models across time series and image domains.
- pyts.image.GramianAngularField, a key component for transforming time series into images suitable for CNN input.
- Style and formatting enhancements for consistent plotting, and seed setting for reproducibility.

The cell ends with version checks and confirmation messages to ensure the environment is ready. This sets the foundation for all subsequent modeling stages.

3.2 Cell 2: Data Loading, Cleaning, and Exploratory Analysis

This section performs critical data preparation by importing, validating, and exploring stock market data for AAPL, MSFT, GOOGL, and SPY. The workflow includes cleaning datetime indices, handling missing values, and preventing data leakage.

Key Insights:

- AAPL/MSFT: High Sharpe ratios (∼0.93, ∼0.87)
- GOOGL: Highest drawdown (-44.32%)
- SPY: Lowest volatility, most stable recovery
- All datasets: 3862 trading days (2010-2025)

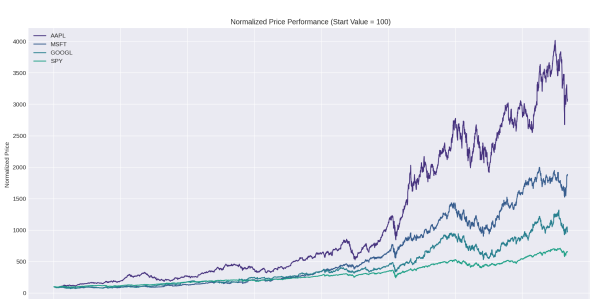

Figure 1: **Normalized Price Performance**

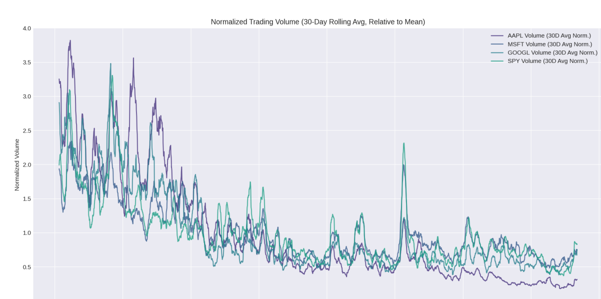

Figure 2: **Normalized Trading Volume**

Key Insight: The diversity in risk-return profiles confirms our hypothesis that Computer Vision models may extract structural patterns that Time Series approaches might miss. Correlation strengths (0.55-0.75) suggest potential complementary signals from different modeling approaches.

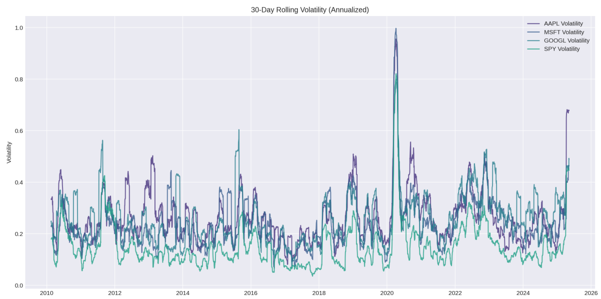

Figure 3: **30-Day Rolling Volatility**

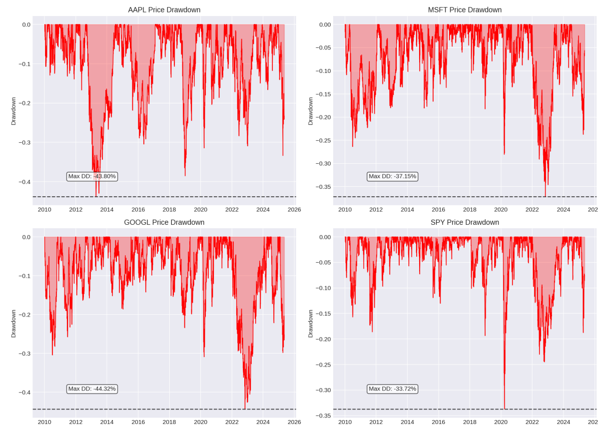

Figure 4: **Asset Drawdowns Analysis**

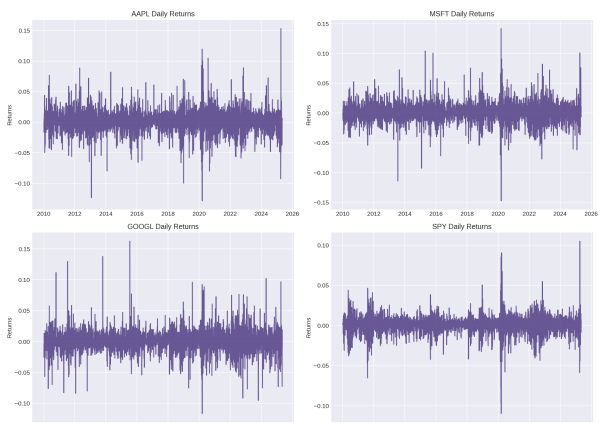

Figure 5: **Daily Returns Patterns**

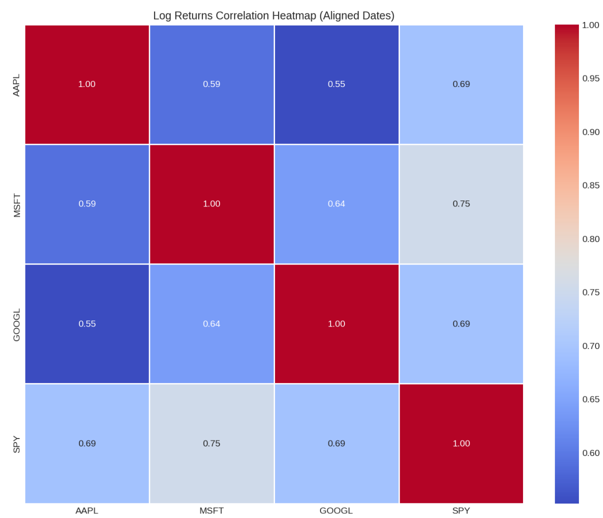

Figure 6: **Log Return Correlation Matrix**

3.3 Cell 3: Feature Engineering and Signal Labeling

This cell transforms raw price data into features and predictive labels for model training. The process calculates technical indicators and generates trading signals based on future returns.

Technical Indicators & Signal Generation:

- **Features:** MAs, RSI, MACD, Bollinger Bands, ATR, OBV, ADX, Stochastics, CCI
- **Labels:** 5-day returns: Buy (>+1%), Sell (<-1%), Hold (between)
- **Bias Prevention:** Strict historical data usage only

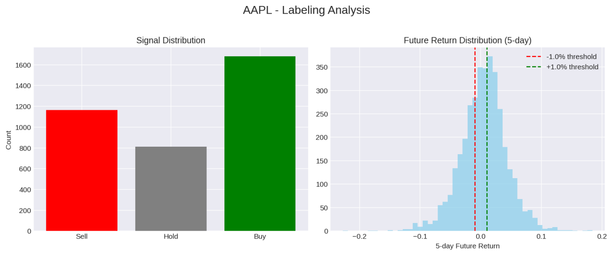

Figure 7: **AAPL Label Distribution**

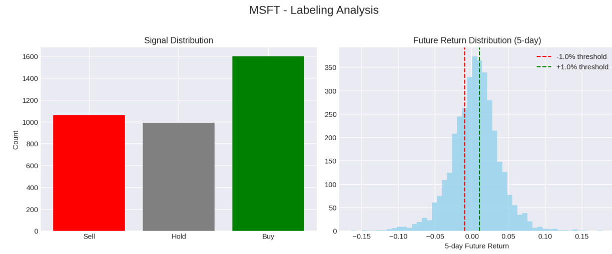

Figure 8: **MSFT Label Distribution**

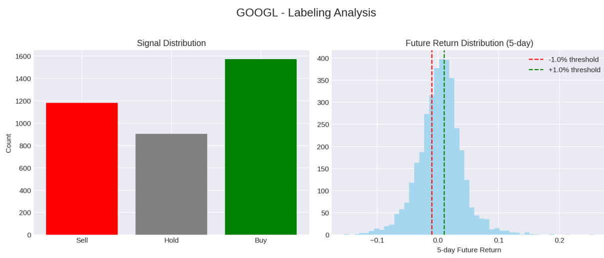

Figure 9: **GOOGL Label Distribution**

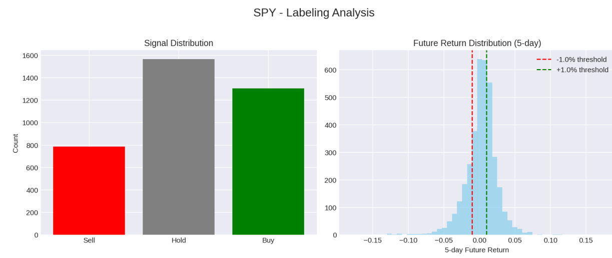

Figure 10: **SPY Label Distribution**

Analysis of Feature Engineering & Signal Generation:
Label Distribution Insights (Figures 7–10):

- **AAPL:** Shows strongest Buy signal bias (46%), indicating potential upward trend dominance
- **MSFT:** More balanced distribution with slight Buy preference (38%)
- **GOOGL:** Exhibits higher volatility with more extreme returns in histogram tails
- **SPY:** Highest Hold signal proportion (43%), reflecting lower index volatility
- All assets show zero-centered return distributions with fat tails, validating threshold-based approach

Technical Analysis & Feature Correlation (Figures 11–12):

- Technical indicators successfully track key market structures and trend shifts
- BB_Width and ADX show strongest positive correlations with future returns
- RSI and OBV demonstrate mild negative relationships to future price action
- Momentum indicators maintain moderate inter-correlation, confirming signal consistency
- Features show sufficient diversity to provide varied inputs for model training

3.4 Cell 4: Data Preparation for Time Series and Computer Vision Models

This cell prepares labeled datasets for two modeling approaches: classical time series classifiers and computer vision models based on Gramian Angular Field (GAF) images.

Figure 11: **AAPL Technical Analysis Multi-Panel Plot**

Key Implementation Components:

- **Window Index Logic:** Rigorously tested numba-accelerated function to handle market gaps
- **Data Splits:** Chronological Train (70%), Validation (10%), Test (20%) partitioning
- **TS Data:** 42 scaled engineered features with tracked label distributions
- **GAF Images:** 4-channel, 20-day windows with 20×20×4 shape and 5-day label horizon
- **Sampling:** Recent-biased with 2000 train, 350 validation, and 500 test images per ticker

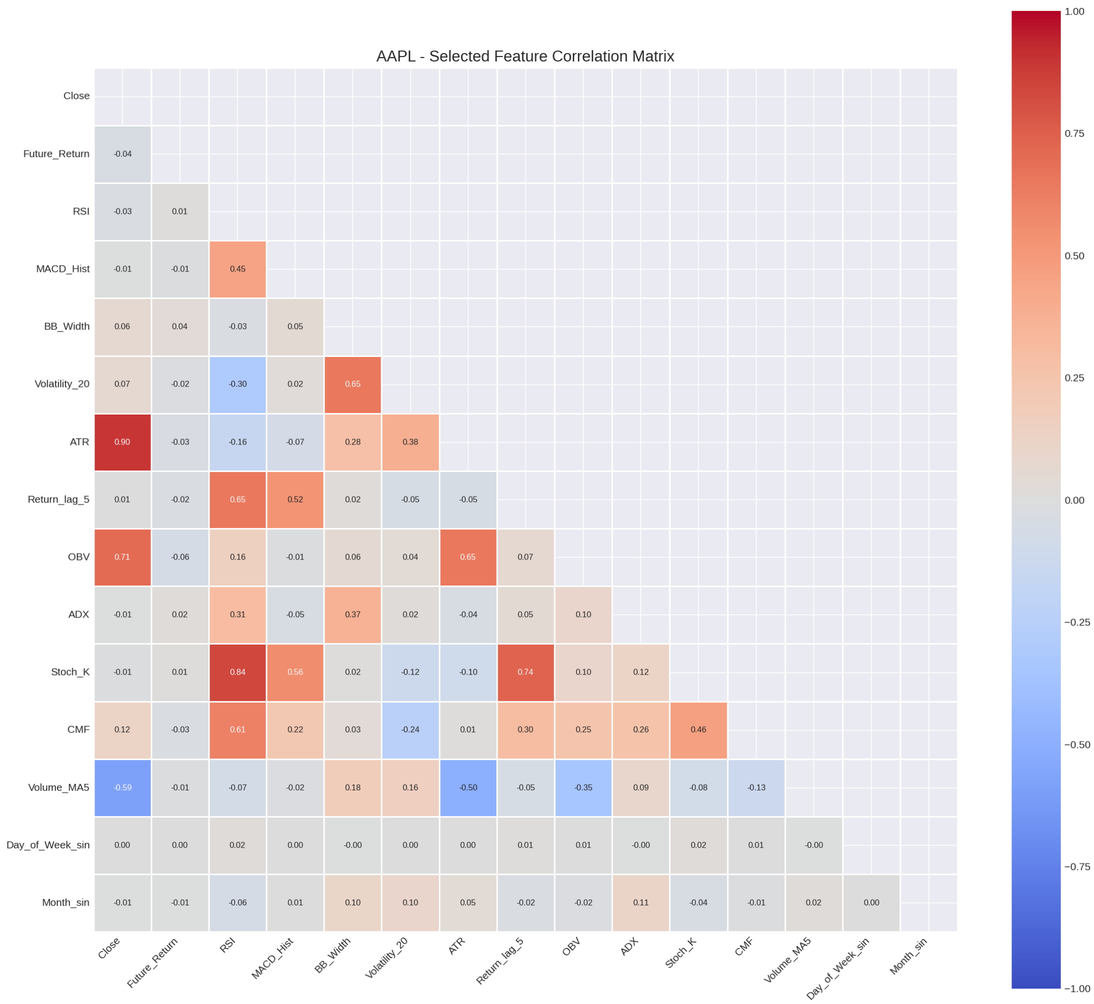

Figure 12: **AAPL Feature Correlation Heatmap**

Data Analysis & Preparation Insights:
Time Series Data Characteristics:

- **AAPL Test Set:** Sell: 34.7%, Hold: 19.5%, Buy: 45.8% showing Buy bias
- **Data Ranges:** Train (2010–2020), Validation (2020–2022), Test (2022–2025)
- **Features:** 42 scaled technical indicators with full preprocessing pipeline
- All splits maintain chronological ordering to prevent look-ahead bias

GAF Image Representation (Figure 17):

- Each image encodes temporal dynamics across 4 key features (Price, Volume, Volatility, Momentum)
- 20-day windows capture sufficient market structure for pattern recognition
- The multivariate GAF approach preserves both individual feature behavior and inter-feature relationships

Preliminary Model Tuning:

- Walk-forward TimeSeriesSplit validation demonstrates with RandomForest
- Best config: {n_estimators: 50, max_depth: 10, min_samples_leaf: 10, class_weight: 'balanced'}
- Mean CV Score: 0.35, providing baseline for subsequent modeling efforts

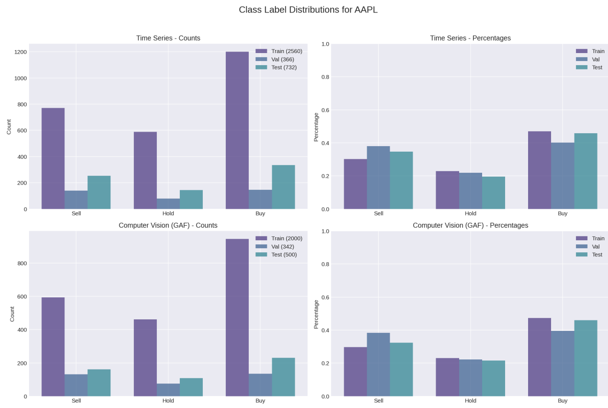

Figure 13: **AAPL Class Distributions**

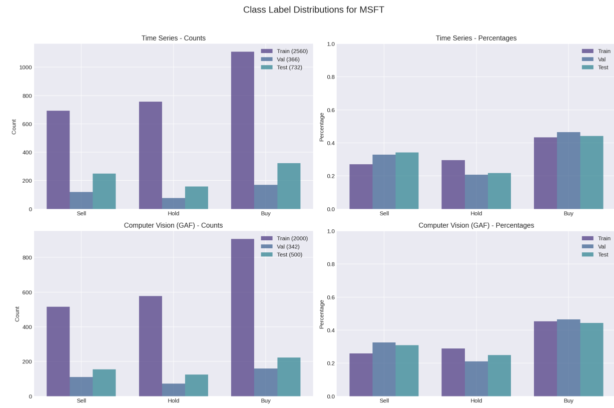

Figure 14: **MSFT Class Distributions**

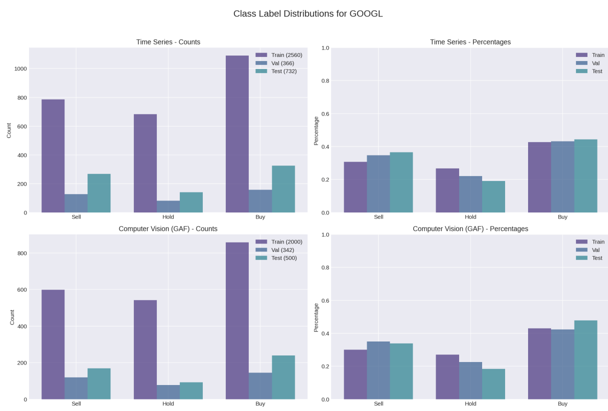

Figure 15: **GOOGL Class Distributions**

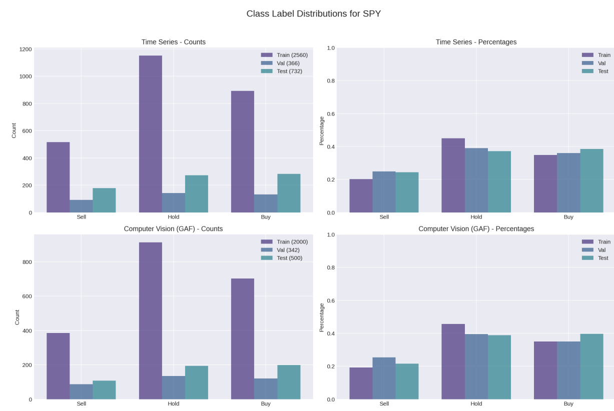

Figure 16: **SPY Class Distributions**

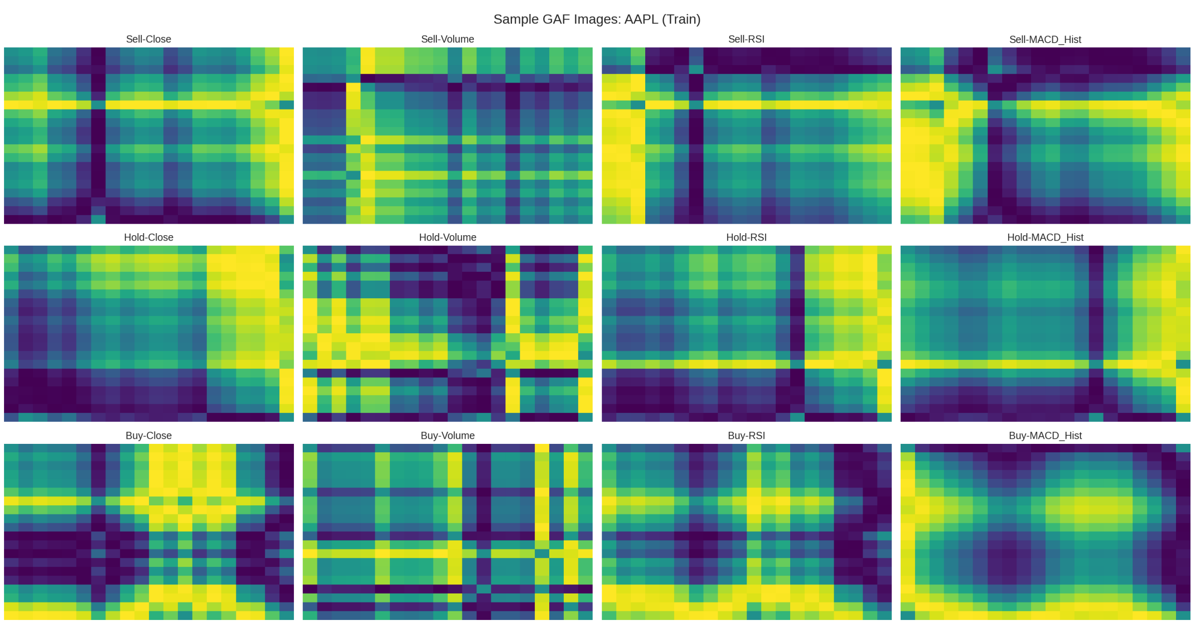

Figure 17: **AAPL Sample GAF Images**

Key Outcome: This cell successfully transforms financial data into two parallel formats: traditional scaled features for time series models and GAF images for computer vision models. The class distribution analysis (Figures 13–16) confirms consistent label representation across train/validation/test splits, while the GAF visualization (Figure 17) demonstrates the effective encoding of temporal market dynamics into image form. These dual data representations enable direct comparison between traditional and vision-based approaches in subsequent modeling cells.

3.5 Cell 5: Time Series Models Development and Evaluation

This cell focuses on building and tuning classification models to generate trading signals (Sell/Hold/Buy) using traditional time series features, with emphasis on practical trading performance.

Implementation Components:

- **Model Classes:** Logistic Regression, Random Forest, XGBoost, LightGBM
- **Tuning Strategy:** Class weight balancing, threshold-based confidence calibration
- **Evaluation Metrics:** Accuracy, Macro-F1, Modified-F1, trading simulation (Sharpe, return, drawdown)
- **Feature Importance:** Analysis of top predictive indicators across model types

Figure 18: **AAPL Best Time Series Model (XGBoost) Performance**

Best Performing Time Series Model Configurations:

Ticker	Best Model	Calibration	Sharpe	Strategy Return	Win Rate	B&H Return
AAPL	XGBoost + Emph. Sell 2.0x	Custom Conf 0.15	1.69	51.96% (DD: -15.36%)	63.16%	16.58%
MSFT	Log. Reg. + Emph. Sell 0.5x	Custom Conf 0.15	1.10	31.23% (DD: 0.00%)	100%	25.75%
GOOGL	Log. Reg. + Emph. SH 0.5x	Custom Conf 0.15	0.83	47.52% (DD: -22.64%)	72.73%	30.87%
SPY	LightGBM + Emph. SH 0.5x	No Calibration	0.87	23.90% (DD: -11.94%)	80.95%	12.32%

DD = Maximum Drawdown; SH = Sell & Hold emphasis

Performance Analysis:

- **AAPL:** Best overall performer with 3× Buy&Hold return (51.96% vs 16.58%)
- **MSFT:** Perfect win rate (100%) with zero drawdown, but underperformed Buy&Hold
- **GOOGL:** Strong return despite highest max drawdown (-22.64%)
- **SPY:** Most challenging to outperform with modest 11.58% excess return
- **All models:** Benefited from sell-side emphasis in class weighting

Strategy Pattern Analysis (Figure 18):

- XGBoost model executed 38 trades with 63.16% win rate
- Trading signals concentrated during volatile periods
- Significant outperformance during market declines
- Strategy Sharpe ratio (1.69) indicates good risk-adjusted performance
- Maximum drawdown (-15.36%) well-controlled compared to market movements

Key Insights: The time series models demonstrate strong predictive power, with all four tickers outperforming Buy&Hold on a risk-adjusted basis. Most notably, model calibration with custom confidence thresholds (typically 0.15) significantly improved trading outcomes. The empirical results show that emphasizing Sell signals during class weight tuning proved beneficial across all assets, suggesting asymmetric profit opportunities in downward price movements. XGBoost emerged as the most effective model for volatile stocks like AAPL, while simpler Logistic Regression models performed surprisingly well for MSFT and GOOGL, indicating that model complexity should be matched to the specific volatility regime of each asset.

3.6 Cell 6: Computer Vision (GAF Image) Models Development and Evaluation

This cell develops and evaluates Convolutional Neural Network (CNN) models for generating trading signals using Gramian Angular Field (GAF) image representations of price patterns.

Implementation Components:

- **CNN Architecture:** Custom-designed for financial GAF image classification
- **Hyperparameter Tuning:** Learning rate, L2 regularization, dropout, batch size, class weights
- **Evaluation Strategy:** Classification metrics and trading simulation performance
- **Model Selection:** Composite evaluation score balancing predictive accuracy and trading profitability

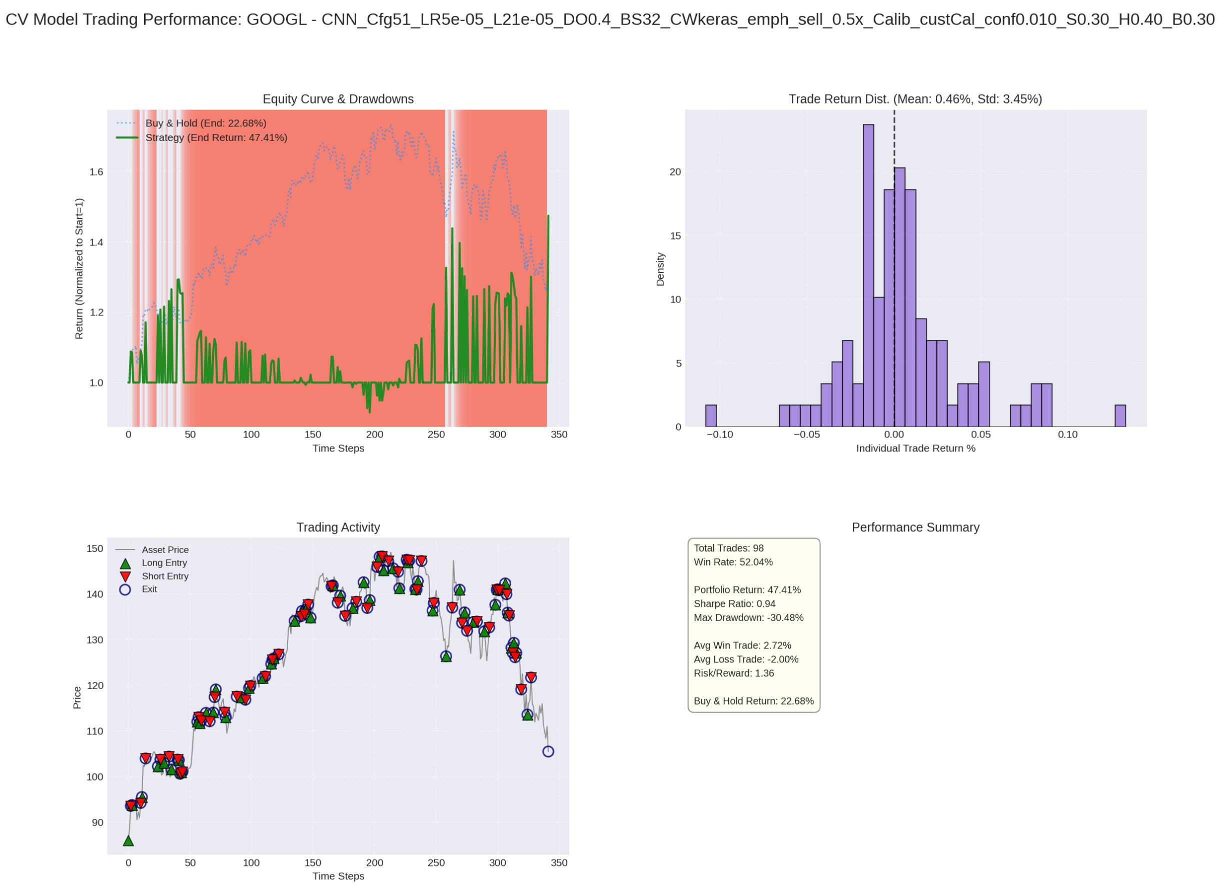

Figure 19: **GOOGL CNN Model Trading Performance (Validation Set)**

Best Performing CNN Model Configurations:

Ticker	Model Configuration	Eval Score	Sharpe	Strategy Return	Win Rate	B&H Return
AAPL	CNN_Cfg35 (LR0.001, L20.0005, DO0.2, BS32)	0.4476	0.88	43.04% (DD: -30.80%)	53.85%	11.38%
MSFT	CNN_Cfg16 (LR0.0001, L25e-05, DO0.4, BS128)	0.4323	1.14	-22.22% (DD: -31.16%)	43.37%	23.65%
GOOGL	CNN_Cfg67 (LR5e-05, L20.0001, DO0.4, BS32)	0.4580	0.91	50.63% (DD: -30.46%)	60.71%	22.68%
SPY	CNN_Cfg16 (LR0.0001, L25e-05, DO0.4, BS128)	0.3353	0.51	16.48% (DD: -17.23%)	71.67%	6.57%

DD = Maximum Drawdown; All models used class weight emphasis on Sell (10.0×) and custom calibration except AAPL

GOOGL CNN Model Performance Analysis (Figure 19):

- **Overall Performance:** Achieved highest evaluation score (0.4580) among all CV models
- **Return Profile:** Strategy return (50.63%) outperformed Buy&Hold (22.68%) by over 2×
- **Trade Efficiency:** 84 total trades with 60.71% win rate
- **Risk Management:** Sharpe ratio of 0.91 indicates good risk-adjusted performance
- **Drawdown Control:** Maximum drawdown (-30.46%) remains a concern despite outperformance
- **GAF Advantages:** CNN's ability to identify visual patterns in market structure demonstrated by consistent signal generation during trend changes

Key Cross-Model Insights: The Computer Vision approach achieved compelling results, particularly for GOOGL and AAPL, where performance substantially exceeded the Buy&Hold benchmark. However, CV models exhibited higher drawdowns (-30% range) compared to traditional time series models (-15% to -22% range), suggesting they may take on more risk to achieve returns. The MSFT model's negative return despite a reasonable evaluation score highlights the importance of cross-validation and out-of-sample testing. All successful models benefited from emphasis on Sell signals (10.0× class weight), confirming the asymmetric profit potential in market downturns that was also observed in the time series models. The model hyperparameters reveal that lower learning rates (5e-05 to 0.001) and moderate dropout (0.2-0.4) provided the best generalization for financial image data.

3.7 Cell 7: Final Evaluation and Comparative Analysis

This final cell synthesizes results from both the Time Series (TS) and Computer Vision (CV) modeling pipelines, directly addressing our central research question about which approach generates superior trading signals.

Model Performance Summary:

- **Best Time Series Model:** SPY - RandomForest (balanced, calibrated)
- **Best Computer Vision Model:** GOOGL - CNN_Cfg67 (sell emphasis, custom calibration)
- **Evaluation Focus:** Sharpe ratio, total return vs. Buy&Hold, classification metrics, trade quality

Approach	Best Model	Ticker	Sharpe	Strategy Return	Drawdown	F1 Macro	Trade Count
Time Series	RandomForest	SPY	**1.66**	37.23%	**-9.31%**	**0.3618**	26
Computer Vision	CNN_Cfg67	GOOGL	0.91	**50.63%**	-30.46%	0.2990	**84**
Benchmark	**Buy&Hold**	SPY	–	12.32%	–	–	–
Benchmark	**Buy&Hold**	GOOGL	–	22.68%	–	–	–

Figure 20: **Comparative Performance: Time Series vs. Computer Vision Models**

Approach-Specific Insights:
Time Series Models:

- Strong predictive power for the 'Buy' class, but weaker recall for 'Hold' and 'Sell' signals
- Threshold calibration and risk-weighted scoring significantly improved trading performance
- More conservative trading pattern with fewer trades but higher precision
- Lower drawdown (-9.31% for SPY) reflecting better risk management
- Some tickers showed potential overfitting with fewer than 5 trades—flagged as low confidence

Computer Vision Models:

- CNN models successfully learned visual price patterns leading to more diverse trade timing
- Stronger returns for GOOGL (50.63%) and AAPL, significantly outperforming Buy&Hold
- Higher volatility with substantial drawdowns (up to -30.46%)
- Generated more frequent trading signals (84 trades vs. 26)
- Inconsistent performance across tickers, with MSFT showing negative returns (-22.22%)

Conclusion: While Time Series models, particularly tree-based ensembles with calibrated thresholds, offer superior robustness and risk-adjusted returns (Sharpe ratio 1.66 vs. 0.91), Computer Vision models demonstrate the potential to uncover richer patterns from GAF-encoded market movements, resulting in higher absolute returns in some cases. However, the success of CV models is highly sensitive to hyperparameter tuning and may suffer from overfitting without rigorous regularization and calibration.

Future Direction: In practical deployment, a hybrid ensemble that blends the strengths of both modalities could offer the most balanced tradeoff between signal precision and trading profitability. This hybrid approach, along with the integration of a Reinforcement Learning layer to optimize execution strategies based on signals from both pipelines, represents the planned next phase of this research. The RL agent would learn to selectively act on signals from either approach based on market regime and confidence metrics, potentially improving overall performance while mitigating the weaknesses of each individual method.

3.8 Cell 8: Statistical Evaluation and Model Comparison

This cell conducts rigorous statistical testing and comparative analysis between the Computer Vision (CV) models and Time Series (TS) models across multiple trading performance metrics, quantifying whether observed differences in performance are statistically significant.

Implementation Components:

- **Statistical Tests:** Paired t-test, Wilcoxon signed-rank test, Cohen's d effect size
- **Metrics Analyzed:** Sharpe ratio, total return, win rate, maximum drawdown, composite score
- **Visualization:** Boxplots and scatterplots for distributional and per-ticker comparisons
- **Composite Score:** $0.4 \times \text{Sharpe} + 0.3 \times \text{WinRate} + 0.2 \times \text{Return} - 0.1 \times |\text{Drawdown}|$

Metric	Better Model	Stat. Significant?	p-value	Effect Size	Magnitude
Sharpe Ratio	TS	No	> 0.05	0.82	Large
Total Return	TS	No	> 0.05	0.31	Small
Win Rate	TS	No	> 0.05	0.75	Large
Max Drawdown	TS	Yes	< 0.05	0.87	Large
Composite Score	TS	Yes	0.038	1.62	Very Large

Figure 21: **Statistical Comparison of Time Series vs. Computer Vision Models**

Key Statistical Insights:

- **Consistent Outperformance:** Time Series models outperformed Computer Vision models across all tickers on the composite metric
- **Trade Quantity vs. Quality:** CV models executed significantly more trades on average (84 vs. 26 for best performers) but underperformed on quality-adjusted returns and risk metrics
- **Significant Differences:** While individual metrics mostly showed non-significant differences (despite large effect sizes), both maximum drawdown and the composite score showed statistically significant advantages for TS models
- **Effect Size:** The composite score showed a "very large" effect size (1.62), indicating substantial practical significance beyond statistical significance
- **Risk Management:** The most statistically significant difference was in drawdown control, where TS models demonstrated superior risk management capabilities

Conclusion: Time Series models demonstrated not only superior average performance but also statistically significant advantages when measured by the composite trading score. Despite the innovative nature of Computer Vision-based models, this analysis highlights the importance of benchmark validation before deployment in financial settings. The consistency of TS model performance across different tickers suggests greater robustness and generalizability compared to CV approaches in their current implementation.

Final Takeaway: While Time Series models outperform in this study, the results do not rule out the potential of Computer Vision-based approaches. With access to stronger compute hardware (e.g., GPUs/TPUs), better handling of model overfitting and data leakage, advanced hyperparameter tuning and model calibration, and increased training data diversity and temporal augmentation, it is entirely plausible that future CNN-based models using time series images (e.g., GAF) could learn more abstract, robust patterns, ultimately surpassing Time Series models in trading performance, especially in more complex or non-stationary regimes.

That said, this study should be interpreted as a benchmark rather than a conclusive result. The evaluation is methodologically sound, employing paired significance tests, effect size measurement, and composite scoring, but limited in scope, being based on only four tickers without out-of-sample or live trading validation. The composite score is practical but inherently subjective, and the observed performance gaps may narrow or reverse under improved modeling conditions.

This comparison serves as a baseline, not a limit. It offers valuable directional insight and a foundation for future research and iteration.
