

Comparative Evaluation of Hybrid Deep Reinforcement Learning Models with Fixed Risk-Reward Ratios for Intraday Gold Trading

Ravi Prajapati

Department of Electronics And Computer Engineering
Thapathali Campus
Kathmandu, Nepal
ravip121997@gmail.com

Kobid Karkee

Department of Electronics And Computer Engineering
Thapathali Campus
Kathmandu, Nepal
kobid@tcioe.edu.np

Abstract— This study proposes a hybrid trading framework that combines a Rule-Based Trading Strategy with a Deep Reinforcement Learning (DRL) agent for intraday gold trading on a 15-minute timeframe. The Rule-Based Trading Strategy generates rule-based signals, while the DRL agent determines whether to execute or skip the generated trading opportunities. The framework evaluates three separate DRL models using fixed risk-reward (RR) ratios of 1:1, 1:2, and 1:3. Volatility-adjusted stop-loss levels are determined using the Average True Range (ATR), with a fixed risk of 0.5% per trade. Each model is trained and evaluated independently under identical trading conditions, enabling a systematic comparison of the effects of different fixed RR ratios on DRL-based trading performance. The trading environment incorporates transaction costs and risk-based position sizing, while the reward mechanism reflects the realized outcome of each trade under the corresponding fixed RR configuration. Model performance is evaluated using cumulative return, maximum drawdown, win rate, profit factor, and average return per trade. The study compares the three fixed-RR models to determine which risk-reward configuration provides the most favorable balance between profitability, risk, and trading consistency in intraday gold trading.

Keywords—Backtesting, Deep Reinforcement Learning (DRL), Proximal Policy Optimization (PPO), Intraday Gold Trading, Rule-Based Trading Strategy.

I. INTRODUCTION

Financial markets have increasingly adopted artificial intelligence (AI) techniques, particularly for algorithmic trading. Deep Reinforcement Learning (DRL) has gained attention because it combines the feature-learning capability of deep learning with the sequential decision-making ability of reinforcement learning, enabling agents to learn trading policies through interaction with market environments [1]. However, many DRL-based trading systems focus primarily on profitability while giving limited attention to risk management, stability, and interpretability.

Rule-based trading strategies based on technical indicators remain widely used because of their structured and interpretable nature, but they often lack adaptability to changing market conditions. The gold market (XAUUSD), with its high volatility and sensitivity to economic factors, presents challenges for both approaches. Combining rule-based signals with DRL therefore provides an opportunity to improve trade selection and risk-aware decision-making.

A. Motivation

The motivation for this research arises from the limitations of both rule-based and standalone DRL approaches. Rule-based systems may perform poorly when market conditions change, while DRL models can suffer from unstable training, large decision spaces, and limited interpretability. This motivates the development of a hybrid framework in which a rule-based strategy provides structured trade signals while a DRL agent evaluates these signals and adapts decisions according to market conditions.

B. Problem Statement

Existing DRL trading systems face challenges related to inefficient exploration, unstable learning, and limited risk management. Therefore, this research addresses the problem of developing a trading framework that combines rule-based signal generation with adaptive DRL decision-making while incorporating multiple Risk Reward and controlled risk management for intraday gold trading.

C. Objective

The primary objective of this research is to develop and evaluate a hybrid trading framework that combines a Rule-Based Trading Strategy with a DRL agent for gold trading. The proposed system allows the DRL agent to accept or reject candidate trade signals for multiple RR ratio from 1:1, 1:2, and 1:3 while operating under predefined risk-management constraints. The main objective of this paper is to answer,

“Can a PPO-based Deep Reinforcement Learning agent improve the performance of a predefined rule-based strategy for intraday gold trading?”

II. LITERATURE REVIEW

The literature review examines existing approaches to algorithmic trading, with particular emphasis on Deep Reinforcement Learning (DRL), rule-based strategies, adaptive trading, and risk management. The reviewed studies highlight the potential of DRL while identifying limitations in action-space design, adaptability, and risk management that motivate the proposed hybrid framework.

A. Related Works

Sim and Kirk (2023) propose a generalized DRL framework using a Deep Q-Network (DQN) with a multivariate state space and adaptive synchronization for

trading multiple S&P 500 stocks [2]. The model achieves an average annualized return of 23.75%, a Sharpe ratio of 0.6622, and a maximum drawdown of 48.70%, outperforming a long-only benchmark.

Lal (2022) applies Double Deep Q-Learning (DDQL) with a Convolutional Neural Network (CNN), experience replay, and a target network to trade four commercial banks on the Nepal Stock Exchange (NEPSE) [3]. The proposed agent outperforms traditional strategies, achieving an annualized expected return of 114.44% on NABIL stock. Despite its profitability, the framework focuses primarily on returns and uses a conventional buy/sell/hold action space. It also focuses on daily stock trading and does not address the volatility and intraday characteristics of gold markets.

Guevara (2025) proposes an actor-critic reinforcement learning framework using adaptable data windows and specialized agents for different market periods [4]. The approach improves adaptation across market conditions and reports an average loss reduction of 0.25% in gold, with substantial improvements over single-agent models.

Kwon and Lee (2024) propose a hybrid framework combining a rule-based Donchain Channel breakout strategy with DRL for trade execution [5]. The rule-based component generates entry signals, while the DRL agent optimizes execution and incorporates volatility-based position sizing. The hybrid approach improves consistency compared with standalone DRL and traditional strategies. This separation of signal generation and execution directly support the approach adopted in this thesis.

Zhong et al. (2024) propose a DRL framework that dynamically switches between predefined strategies, including RSI, MACD, and Bollinger Bands [6]. The agent selects strategies according to market conditions, improving risk-adjusted performance and robustness compared with fixed-strategy approaches. However, the framework requires multiple predefined strategies, increasing complexity and dependence on handcrafted components. Nevertheless, it demonstrates the value of adaptive decision-making across different market regimes.

B. Research Gap

Most existing DRL trading systems formulate trading as a direct mapping from market states to buy, sell, or hold actions. This can create a large exploration space and contribute to unstable or inefficient policy learning. Furthermore, many studies focus on daily or traditional financial markets rather than intraday gold trading. Although, some studies combine rule-based strategies with DRL, limited research explicitly treats the DRL agent as a selective decision layer that validates candidate signals while also optimizing risk--reward settings.

Therefore, a research gap exists in developing a hybrid DRL framework that combines structured signal generation with adaptive trade selection and discrete risk--reward optimization. Such an approach can reduce the effective action space, improve learning stability, and provide greater adaptability to changing intraday market conditions.

III. METHODOLOGY

A. Dataset

For this research, the dataset is sourced from MetaQuotes Ltd. and consists of XAU/USD gold price data at a 15-minute timeframe. The dataset contains DateTime and OHLC (Open,

High, Low, Close) prices, providing the essential information required to analyze short-term intraday market movements.

The dataset is divided chronologically into training, validation, and test sets to prevent data leakage. The model is trained on earlier data (2020-01-01 to 2024-12-31), validated on (2025-01-01 to 2025-12-31) data, and tested on unseen data of (2026-01-10 to 2026-06-17). This time-based split ensures that future information does not influence past predictions and better reflects real-world trading conditions.

B. Preprocessing

The raw XAU/USD OHLC data is first cleaned by standardizing column names and sorting the data chronologically. Technical indicators including RT, MACD, Signal, ATR, DI, and ADX are then calculated. The data is normalized to improve model stability, and the selected features are high relative to the current price (high_rel), low relative to the current price (low_rel), close relative to the current price (close_rel), close return (close_return), ATR percentage (ATR_pct), MACD percentage (MACD_pct), ADX, positive DI (DI_pos), negative DI (DI_neg), and the rule-based trading signal (DSG). Unnecessary intermediate columns are removed before the data is provided to the DRL environment.

The Rule-Based Trading Strategy (DSG) generates candidate signals using MACD crossovers with an ADX filter. A buy signal (1) is generated when MACD crosses above the Signal line and ADX[12] is greater than 20. Otherwise, no signal (0) is generated.

For each generated trade, the stop-loss is determined using 1.5 times the ATR from the entry price. Separate PPO agents are trained for three fixed risk--reward ratios: 1:1, 1:2, and 1:3. Each agent learns whether to accept or reject the candidate signals while operating under its assigned RR ratio. This approach allows the performance of different risk--reward configurations to be evaluated independently.

C. Model Used

The Proximal Policy Optimization (PPO)[10] is used as the reinforcement learning algorithm in this study as previous studies suggest that PPO is superior to the DQN and A2C algorithms, particularly in terms of total profit and risk management[7] [8] [9]. PPO is a policy-gradient method that combines deep neural networks with reinforcement learning to learn trading decisions through interaction with a simulated market environment [10]. It is selected because of its stable policy updates and suitability for volatile markets such as gold.

At each decision point, the agent receives an observation consisting of the market information from the previous 32 candles along with the current account balance. The agent starts with an initial balance of \$10,000 and risks 0.5% of the current account balance per trade. The position size is therefore adjusted according to the account balance and the predefined stop-loss.

The Rule-Based Trading Strategy generates the initial buy or sell signals, while the PPO agent acts as a decision layer that determines whether to accept or reject each candidate trade. A volatility-adjusted stop-loss of 1.5 ATR from the entry price is used for each trade.

To evaluate different risk--reward configurations, three separate PPO agents are trained independently with fixed RR ratios of 1:1, 1:2, and 1:3. Each agent learns a policy that maximizes cumulative rewards by selecting which candidate trades to execute under its assigned RR configuration.

During training, the agents explore different trading decisions and gradually learn to exploit profitable patterns. The trained agents are then evaluated on unseen validation and test data to compare their performance across the three risk-reward configurations.

Experiment Parameter Configuration: To ensure consistency and reproducibility, the proposed trading framework is evaluated using a fixed experimental configuration. The experimental parameters define the characteristics of the trading environment, including the initial account balance, risk exposure, stop-loss, transaction costs, observation window, and risk-reward configurations. These parameters determine how trades are executed and how financial performance is measured.

TABLE I. EXPERIMENT PARAMETERS

Parameters	Values
Initial Balance	\$10,000
Risk Per Trade	0.5%
Lookback Window	32
Stoploss	1.5 ATR
Risk Reward Ratio	1:1, 1:2, 1:3
Spread	20 points
Comission	\$5 per lot

Hyperparameter Configuration: The PPO-specific training settings are defined separately as model hyperparameters. These parameters control the learning process, including the learning rate, discount factor, batch size, number of training steps, policy clipping range, and other PPO optimization settings.

TABLE II. HYPERPARAMETER VALUES

Hyperparameters	Values
Learning Rate	0.0003
Number of Steps	2048
Batch Size	64
No of Epoch	10
Discount Factor	0.99
GAE Parameter	0.95
Clip Range	0.2
Entropy Coefficient	0
Value Function Coefficient	0.5
Random Seed	42

D. Working Principle

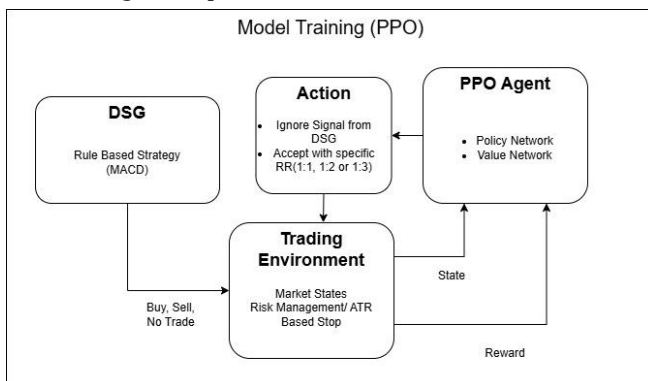


Fig. 1. Working of the Model

1) State Space

At each decision point, the observation contains data from the previous 32 candles using the following features: high_rel,

low_rel, close_rel, close_return, ATR_pct, MACD_pct, ADX, DI_pos, DI_neg, and DSG, together with the current account balance. These features provide information about recent price movement, volatility, momentum, trend strength, market direction, and the current trading signal. The final feature set was selected after evaluating multiple feature combinations during the validation process. Among the feature sets tested, this combination produced the best validation performance and was therefore used for the final model.

2) Action Space

The PPO agent has a discrete action space with two actions: Take (1) or Skip (0) the trades determined by the rule-based signal(DSG).

Deterministic Signal Generator (DSG):

The Deterministic Signal Generator (DSG) is a predefined rule-based trading strategy used to generate candidate trading signals for the PPO agents. The DSG is based on the Moving Average Convergence Divergence (MACD) indicator. It determines the potential trading direction, while the PPO agent subsequently decides whether to execute or skip the generated trading signal.

The MACD indicator is calculated using a 12-period fast exponential moving average (EMA) and a 26-period slow EMA.[11] The MACD line is calculated as the difference between these two exponential moving averages:

$$MAX\Delta = EMA_{12} - EMA_{26} \quad (1)$$

A 9-period EMA of the MACD line is used as the signal line:

$$\Sigma\gamma\nu\alpha\lambda = EMA_9(MAX\Delta) \quad (2)$$

A BUY signal is generated when the MACD line crosses above the signal line and ADX is greater than 20. This occurs when the MACD line is less than or equal to the signal line on the previous candle and becomes greater than the signal line on the current candle:

$$MAX\Delta_{t-1} \leq \Sigma\gamma\nu\alpha\lambda_{t-1} \text{ and } MAX\Delta_t > \Sigma\gamma\nu\alpha\lambda_t \quad (3)$$

3) Reward Function

The agent receives a positive reward(+1) when an executed trade reaches its take-profit level and a negative reward(-1) when it reaches the stop-loss for a 1:1 Risk reward agent. Similarly, for 1:2 the agent receives +2 and for 1:3 the agent receives +3 when the executed trade reaches its take profit.

4) Trade Execution and Termination Logic

When a valid buy or sell signal is generated, the agent selects Take or Skip. If Take is selected, the trade is opened at the next candle's opening price. The stop-loss is placed at 1.5 ATR from the entry price, with the take-profit determined by the agent's assigned RR configuration of 1:1, 1:2, or 1:3. Position size is calculated so that 0.5% of the current account balance is at risk.

5) Trade Termination

A trade remains active until either the stop-loss or take-profit is reached. The episode continues through the available market data and terminates when the end of the dataset is reached or when no further candles are available for trade execution. The simulation also incorporates a spread of 20

points and a commission of \$5 per lot to account for transaction costs.

If a trade remains open when the end of the available dataset is reached, it is forcibly closed at the closing price of the final available candle. The resulting profit or loss from this forced closure is included in the corresponding performance calculations. Furthermore, because OHLC candle data do not provide the exact intrabar sequence of price movements, if both the stop-loss and take-profit levels are reached within the same candle, the stop-loss is assumed to be triggered first. This conservative assumption is applied consistently throughout the process.

6) Rationale for Methodological Design

The methodological choices in this study were guided by the objective of investigating whether DRL can improve the selective execution of a predefined trading strategy rather than replacing the strategy's signal-generation mechanism. PPO was selected because the resulting decision problem has a discrete action space consisting of trade execution or rejection. The MACD 12/26/9 crossover strategy was used to generate deterministic candidate signals, providing a reproducible baseline and allowing the contribution of the DRL trade-selection component to be isolated.

Three fixed risk-reward configurations (1:1, 1:2, and 1:3) were evaluated to examine the effect of different reward-to-risk objectives while keeping the agent's decision space unchanged. The selected state features capture complementary information regarding price movement, returns, volatility, momentum, and trend strength. The DSG signal is also provided to the agent because it defines the candidate trade direction.

A chronological training-validation-test framework was adopted to preserve the temporal structure of financial data and prevent future observations from influencing model development. The final test period was kept unseen during training and model selection to provide an out-of-sample evaluation of generalization. Finally, the action space was restricted to TAKE or SKIP because the trading direction was already determined by the DSG. This design isolates the specific research question of whether a DRL agent can improve the execution quality of a conventional rule-based trading strategy through selective trade execution.

IV. RESULT AND DISCUSSION

This chapter presents and discusses the performance of the proposed hybrid trading framework. The evaluation focuses on the three PPO agents trained with different risk-reward ratios: 1:1, 1:2, and 1:3. The agents are evaluated on unseen test data and compared using profitability, risk, and trading-performance metrics. The results are also compared with the underlying Rule-Based Trading Strategy to determine whether the PPO-based trade-selection layer provides additional value.

A. Training Performance

The training performance of the three PPO agents was evaluated over 500,000 training timesteps with validating the agent's performance in each 50,000 steps. Separate agents were trained using risk-reward ratios of 1:1, 1:2, and 1:3 under the same training and environment settings. Training performance was assessed using cumulative reward, policy loss, value function loss, and entropy to examine the learning behavior and stability of the agents.

1) Cumulative Reward :

The cumulative reward obtained by PPO agent during training is presented in Figure 2.

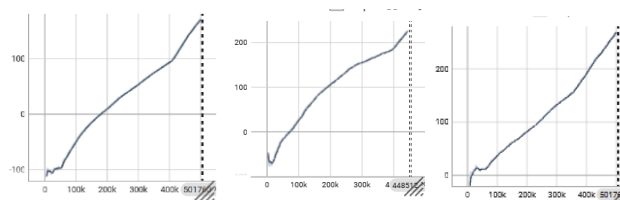


Fig. 2. Cumulative reward obtained by 1:1, 1:2 and 1:3 Risk Reward Agent

The reward progression shows that the cumulative reward for the agent increases as we train the agent through 500k steps. It shows increase in cumulative reward for the agent as we continue training and the value is seen to be increasing on all three cases.

2) Entropy Loss:

The entropy of the PPO policy during training is shown in Figure 3.

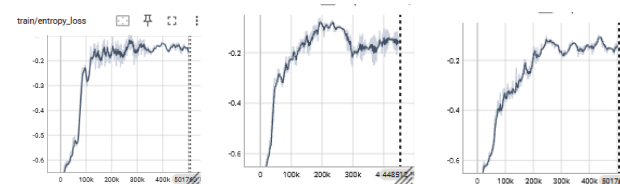


Fig. 3. Entropy Loss of the PPO Policy for 1:1, 1:2 and 1:3 Risk Reward Agent

The entropy increased as the training progresses and after around 200k timestep the entropy started to flat out. This indicates that the agent's exploration behavior became consistent as it learned the Take/Skip policy.

3) Value Function Loss

The value loss of the PPO policy during training is shown in Figure 4.

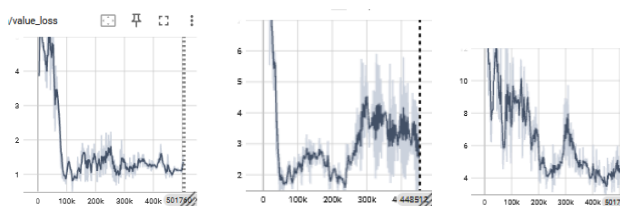


Fig. 4. Value Loss of the PPO Policy for 1:1, 1:2 and 1:3 Risk Reward Agent

The value loss of the PPO policy decreased as the training progresses and stabilized and then fluctuated as it reached later phases of training.

4) Summary of Training Behavior

Overall, the training results indicate that the PPO agents showed stable learning and after around 100k – 200k timesteps the value function loss and entropy all stabilized or was at near its lowest point of the training. However, the cumulative reward was increasing till the last timestep which might have happened as the model overfit the data as we train more. There are also differences in training behavior among the three agents which may be attributed to the different risk-reward configurations, which produce different reward distributions during learning.

B. Validation Performance

During the training process, the three PPO agents were periodically evaluated on the validation dataset to monitor their generalization performance and identify potential overfitting. Model checkpoints were saved at every 50,000 training timesteps. After each checkpoint, the corresponding PPO agent was evaluated on the validation dataset using the same trading environment, risk management settings, and transaction costs. This resulted in validation evaluations at 50,000, 100,000, 150,000, and 200,000 timesteps and so on up to 500,000 timesteps.

The validation performance showed continuous improvement up to approximately 150,000 training timesteps, after which the performance largely plateaued. This indicates that additional training beyond 150,000 timesteps provided limited improvement in the agent's ability to generalize to unseen validation data. Therefore, the model checkpoint at 150,000 timesteps was selected as the optimal model for subsequent testing. This selection is also supported by the training diagnostics, particularly the entropy and value function loss, which indicate that the learning process had largely stabilized around this stage. Selecting the model based on validation performance rather than the final training checkpoint helps reduce the risk of overfitting to the training data.

The validation performance of each checkpoint was compared using key financial metrics, including final account balance, total return, number of trades, win rate, average R per trade, profit factor and maximum drawdown [13]. The checkpoint that demonstrated the most favorable validation performance was selected as the optimal model for subsequent evaluation.

This periodic validation process helps identify the point at which further training no longer improves performance on unseen data. It also helps reduce the risk of selecting a model that performs well on the training data but fails to generalize to new market conditions. The same procedure was applied independently to the three PPO agents trained with 1:1, 1:2, and 1:3 risk-reward ratios.

1) Overall Validation Performance

The validation results for the three PPO agents are presented in Table III.

TABLE III. VALIDATION RESULTS

Metric	PPO 1:1	PPO 1:2	PPO 1:3
Initial Balance	\$10,000.00	\$10,000.00	\$10,000.00
Final Balance	\$10,177.76	\$10,718.17	\$12,992.52
Total Return	1.78%	7.18%	29.93%
Number of Trades	205	67	259
Wins	106	27	80
Losses	99	40	179
Win Rate	51.71%	40.30%	30.89%
Net Profit	\$177.76	\$718.17	\$2,992.52
Profit Factor	1.04	1.36	1.31
Expectancy	\$0.87	\$10.72	\$11.55
Maximum Drawdown	4.45%	4.98%	10.34%

The validation results of the three PPO agents are presented in Table III. PPO 1:3 achieved the highest profitability, increasing the initial balance from \$10,000.00 to \$12,992.52, with a return of 29.93%, net profit of \$2,992.52, and expectancy of \$11.55 per trade. However, this was accompanied by a lower win rate of 30.89% and a maximum drawdown of 10.34%. PPO 1:2 achieved a 7.18% return with

the lowest maximum drawdown of 4.98%, indicating the most balanced risk-return profile. PPO 1:1 recorded the highest win rate (51.71%) and lowest drawdown (4.45%) but generated the lowest return of 1.78%. Overall, PPO 1:3 provided the highest profitability, PPO 1:2 the most balanced performance, and PPO 1:1 the highest win rate with the lowest drawdown.

2) Validation Equity Curve

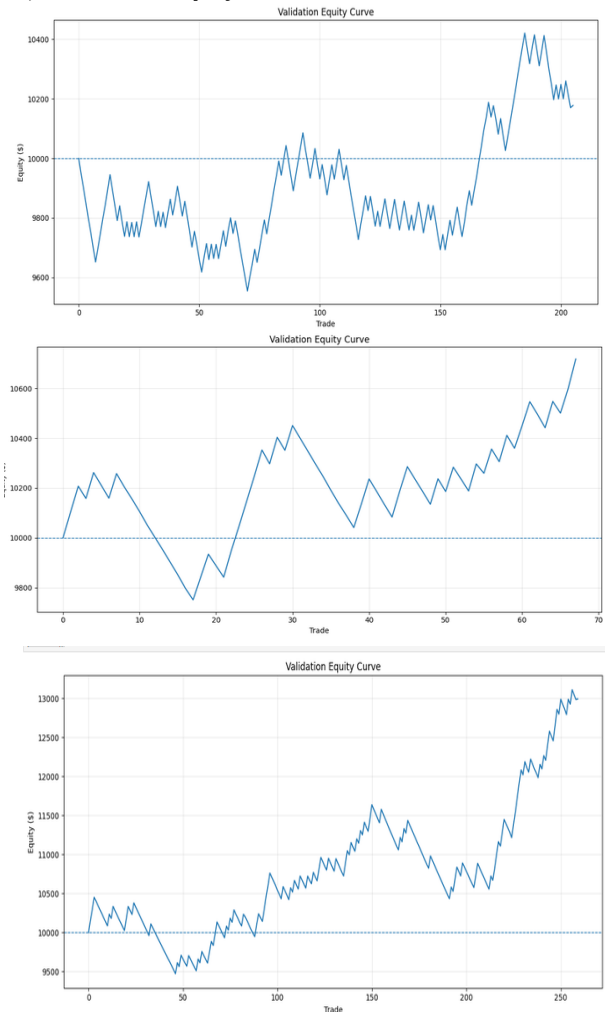


Fig. 5. Validation Equity Curve of 1:1, 1:2 and 1:3 Risk Reward Agent

The equity curves show distinct risk-return characteristics across the three PPO agents. The 1:1 agent ended at \$10,177.76 after fluctuations and recovery, indicating limited profitability. The 1:2 agent showed a more stable upward trend and finished at \$10,718.17, reflecting more selective and consistent trading. The 1:3 agent achieved the strongest growth, reaching \$12,992.52, but with greater fluctuations and drawdowns. Overall, higher RR ratios increased profit potential but also resulted in greater equity variability and risk.

C. Test Performance

After selecting the optimal model based on validation performance, the selected PPO agents were evaluated on the unseen test dataset. The test dataset was not used during model training or model selection and therefore provides an unbiased assessment of the framework's out-of-sample performance.

The evaluation was conducted using the same trading environment and experimental settings used during training and validation, including an initial balance of \$10,000, 0.5% risk per trade, a 1.5 ATR stop-loss, and the corresponding

risk–reward ratio. Spread and commission were also included in the test evaluation.

1) Overall Test Performance and Baseline Performance

The following tables present the performance of the three selected PPO agents on the unseen test dataset. The test set was not used for model training or model selection, providing an out-of-sample evaluation of the learned policies.

TABLE IV. TEST RESULTS

Metric	PPO 1:1	PPO 1:2	PPO 1:3
Initial Balance	\$10,000.00	\$10,000.00	\$10,000.00
Final Balance	\$9,570.83	\$9,795.92	\$9,560.92
Total Return	-4.29%	-2.04%	-4.39%
Number of Trades	92	29	128
Wins	41	8	30
Losses	50	21	96
Win Rate	44.57%	27.59%	23.44%
Net Profit	-\$429.17	-\$204.08	-\$439.08
Profit Factor	0.82	0.79	0.91
Expectancy	-\$4.66	-\$7.04	-\$3.43
Maximum Drawdown	5.36%	4.23%	10.66%

The 1:2 PPO agent achieved the best overall test result in terms of capital preservation, with a -2.04% return and the lowest maximum drawdown of 4.23%. The 1:3 agent achieved the highest expectancy among the three agents at -\$3.43, but also experienced the highest drawdown of 10.66%. All agents recorded a profit factor below 1, indicating that gross losses exceeded gross profits during the test period. This deterioration compared with the validation results highlights the challenges of generalizing reinforcement learning policies to unseen market.

Three baseline strategies were established using the DSG signals with 1:1, 1:2, and 1:3 risk-to-reward ratios. Each baseline executes all trades generated by the DSG without skipping any signals, providing a reference for evaluating the PPO agent's ability to selectively take or skip trades.

TABLE V. BASELINE RESULT IN TEST PERIOD

Metric	BASLINE 1:1	BASLINE 1:2	BASLINE 1:3
Initial Balance	\$10,000.00	\$10,000.00	\$10,000.00
Final Balance	\$7,846.61	\$7,672.52	\$8,456.15
Total Return	-21.53%	-23.27%	-15.44%
Trades	362	302	265
Wins	156	83	57
Losses	201	214	204
Win Rate	43.09%	27.48%	21.51%
Net Profit	-\$2,153.39	-\$2,327.48	-\$1,543.85
Profit Factor	0.76	0.76	0.84
Expectancy / Trade	-\$5.95	-\$7.71	-\$5.83
Max Drawdown	25.68%	26.74%	21.12%

The baseline provides the reference performance of the underlying rule-based strategy without DRL filtering. Across all three reward-to-risk configurations, the strategy produces negative returns.

The 1:3 configuration is the strongest baseline, with a final balance of \$8,456.15, corresponding to a 15.44% loss and a maximum drawdown of 21.12%. It also has the highest profit factor among the three configurations (0.84). However, its 21.51% win rate and 25 consecutive losses indicate substantial instability.

The 1:1 configuration achieves the highest win rate at 43.09%, but this does not translate into profitability because the average winning trade is substantially smaller than the average losing trade. It consequently produces a 21.53% loss.

The 1:2 configuration performs worst in terms of total return, losing 23.27%, while also recording the highest maximum drawdown of 26.74%.

2) Comparison of Test Performance with Baseline

The most important result is that the PPO agent substantially reduces losses by selectively filtering DSG signals, even though the underlying strategy itself is unprofitable.

For 1:1, PPO reduces the number of trades from 362 to 92, while improving the total return from -21.53% to -4.29%. Maximum drawdown is also reduced from 25.68% to 5.36%. For 1:2, PPO takes only 29 trades compared with 302 baseline trades. Although the win rate is almost identical, the loss is reduced from -23.27% to -2.04%, and maximum drawdown falls from 26.74% to 4.23%. For 1:3, PPO reduces the trade count from 265 to 128 and improves the profit factor from 0.84 to 0.91. The loss is reduced from -15.44% to -4.39%, while maximum drawdown falls from 21.12% to 10.66%.

The important observation is that the improvement does not primarily come from increasing the win rate. Instead, it comes from trade selection. The PPO agent skips a substantial proportion of DSG signals, avoiding some losing trades while retaining a subset of potentially more favorable signals.

However, none of the three PPO configurations achieves positive expectancy on the test dataset. Therefore, the results demonstrate risk and loss reduction rather than a fully profitable trading system.

V. FINDINGS

The results show that the PPO agent improved upon the baseline strategies that executed every DSG signal. The baseline 1:1, 1:2, and 1:3 strategies produced returns of -21.53%, -23.27%, and -15.44%, respectively, while the PPO agents achieved -4.29%, -2.04%, and -4.39%. This indicates that selectively filtering DSG signals was more effective than executing every generated signal.

The PPO agents substantially reduced the number of executed trades, with 92, 29, and 128 trades for the 1:1, 1:2, and 1:3 configurations, respectively. This resulted in considerably lower maximum drawdowns. For example, the 1:2 agent reduced maximum drawdown from 26.74% to 4.23%, indicating that the agent successfully avoided many unfavorable signals.

The result can be partly attributed to the weak underlying DSG signal quality, as all baseline configurations produced negative expectancy during the test period. In addition, changing market conditions may have made it difficult for the agent to learn a policy that performs consistently across different market regimes.

Overall, the PPO agent demonstrated potential as a trade-filtering mechanism for DSG signals. Its main benefit was reducing unnecessary trades and drawdown rather than generating a profitable strategy. Although the PPO agents outperformed their corresponding baselines, all configurations remained unprofitable on the unseen test dataset.

To better understand the discrepancy between validation and test performance, we first examine the overall market

structure during the three periods. Figure 6 shows the XAUUSD close price along with key market characteristics for the training (2020-2024), validation (2025), and test (2026) periods.

The price chart illustrates that the training period contains a long-term upward trend with several pullbacks, while the validation period (2025) is relatively more bullish with frequent short-term swings. In contrast, the test period (2026) shows a more volatile sharp downside moves, which may have contributed to the decreased performance of the PPO agents.

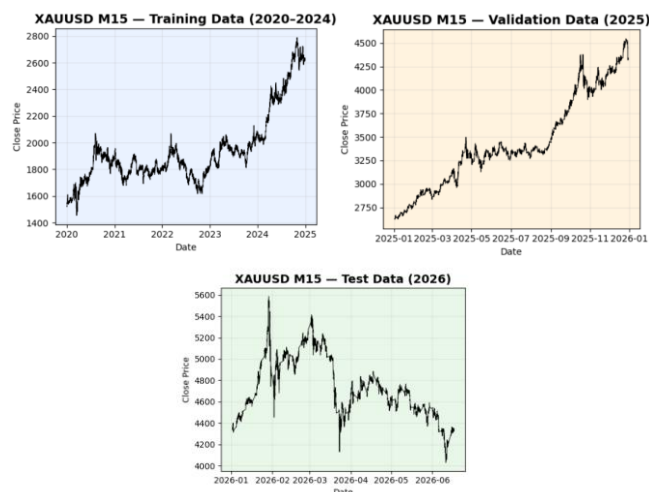


Fig. 6. Market Structure During Training, Validation and Testing

A. Possible Factors Contributing to Limited Generalization

The relatively weak out-of-sample performance may be explained by several interacting factors:

Market non-stationarity: Changes in market trends, volatility, and regimes may have caused patterns learned during training to become less effective during the test period which can be seen in Figure. 6.

Changes in strategy effectiveness: The predefined rule-based strategy have generated less reliable signals under the test-period conditions as we see in Table. V. that the result of only taking the DSG was much worse than the PPO Agents. Since PPO primarily selected or rejected these signals, deterioration in their quality may have limited overall performance.

Limitations of state representation: Although more than 20 features were initially considered, feature selection and ablation reduced the observation space to 10 features. These may not have captured all information relevant to different market conditions, such as additional price-action, volatility, trend, or higher-timeframe characteristics.

Potential validation overfitting: The validation dataset was repeatedly used for feature selection, hyperparameter tuning, reward design, training duration, and model selection. Despite the test set remaining unseen, this repeated optimization may have caused the final policy to adapt to characteristics specific to the validation period.

Overall, the limited generalization likely resulted from the combined effects of changing market conditions, reduced strategy effectiveness, limitations in the state representation, and potential overfitting to the validation period.

VI. CONCLUSION

This study developed a hybrid trading framework in which a PPO agent selectively executes signals generated by a predefined Rule-Based Trading Strategy. The PPO agent substantially improved upon the all-signal baselines by reducing the number of trades, losses, and maximum drawdown. However, the weak underlying DSG signal quality, changing market conditions, and limited generalization to unseen data may have contributed to the continued negative expectancy. Therefore, while PPO-based trade selection shows potential for improving the risk characteristics of a rule-based strategy, further improvements are required to achieve consistent profitability.

The PPO agent demonstrates improved trade selection relative to the unfiltered rule-based baseline, substantially reducing cumulative losses and maximum drawdown, although it does not achieve positive profitability on the test dataset.

REFERENCES

- [1] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincui Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2024.
- [2] Junyoung Sim and Benjamin Kirk. Generalized deep reinforcement learning for trading. *Journal of Student Research*, 12(1), Feb. 2023.
- [3] Janak Kumar Lal. Design of Stock Trading Agent Using Deep Reinforcement Learning. PhD thesis, IOE Pulchowk Campus, 2022.
- [4] Cesar Guevara. Stock market trading via actor-critic reinforcement learning and adaptable data structure. *PeerJ Computer Science* 11:e2690, 2025.
- [5] Yuhee Kwon and Zoonky Lee. A hybrid decision support system for adaptive trading strategies: Combining a rule-based expert system with a deep reinforcement learning strategy. *Decision Support Systems*, 177:114100, 2024.
- [6] Xingyu Zhong, Jinhui Wei, Siyuan Li, and Qingzhen Xu. Deep reinforcement learning for dynamic strategy interchange in financial markets. *Applied Intelligence*, 55(1):30, Nov 2024.
- [7] Chaimae Khaled, Badr Hirschoua, Aziz Lmakri, Hicham Moutachouik, and Mustapha Hain. Optimizing deep reinforcement learning for financial markets: A comparative analysis of dqn, a2c, and ppo. In Saad Motahhir, Badre Bossofi, and Josep M. Guerrero, editors, *Digital Technologies and Applications*, pages 95–106, Cham, 2026. Springer Nature Switzerland.
- [8] Aayansh Vashisht. Comparative analysis of deep reinforcement learning algorithms for financial trading. 04 2026.
- [9] Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. Deep reinforcement learning for automated stock trading: an ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance, ICAIF '20*, New York, NY, USA, 2021. Association for Computing Machinery.
- [10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- [11] Pat Tong Chio. A comparative study of the macd-base trading strategies: evidence from the us stock market, 2022.
- [12] Ikhlaas Gurrib. Performance of the average directional index as a market timing tool for the most actively traded usd based currency pairs. *Banks and Bank Systems*, 13(3):58–70, 2018.
- [13] Malik Magdon-Ismail and Amir F Atiya. Maximum drawdown. *Risk Magazine*, 17(10):99–102, 2004.