Forecasting S&P 500 Implied Volatility with Deep Reinforcement Learning

Your Name
University of XYZ

June 12, 2025

Abstract

This paper proposes a deep-reinforcement-learning (DRL) framework for one-day-ahead forecasting of the S&P 500 implied-volatility surface. We show that policy-gradient agents (PPO, A2C) outperform a comprehensive set of econometric and machine-learning baselines.

1 Introduction

2 Data and Feature Engineering

Briefly summarise the OptionMetrics dataset and the derived feature blocks (surface, realised, macro, FPCA).

2.1 New Features and VVIX Splice

- **New Features:**
  - *Realised Volatility:* Calculated from the underlying stock price.
  - *Macro Factors:* Derived from economic indicators such as GDP, inflation, and unemployment.
  - *FPCA:* Principal Component Analysis applied to the implied volatility surface.
- **VVIX Splice:** A method to estimate the VVIX index using the S&P 500 index and the VIX index.

3 Methodology

This section presents the econometric and machine–learning baselines (HAR-RV, ridge-OLS, LSTM) and the DRL environment (state, action, reward with static-arbitrage penalty).

3.1 Hyper-parameter tuning

All learnable models are optimised with *Optuna* [Akiba et al., 2019]. Stage 1 draws 30 trials from a log-uniform search space covering the learning rate $\alpha \in [10^{-5}, 10^{-2}]$, entropy coefficient $\beta \in [0, 10^{-2}]$, mini-batch size $\{64, 128, 256\}$, and discount factor $\gamma \in [0.90, 0.999]$. We employ `MedianPruner` early-stopping with a patience of five evaluation windows; unpromising trials are terminated to conserve compute. Stage 2 "narrow search" re-samples a further ten trials using truncated priors centred on the best quartile of Stage 1. The final configuration is the global best across both stages. A complete sweep for PPO, A2C, and the LSTM baseline takes ˜90 minutes on a 16-core CPU workstation.

4 Results

4.1 Out-of-sample Accuracy

Table 1 reports RMSE, MAE, MASE, MAPE and QLIKE for all models.

model	RMSE	MAE	MASE	MAPE(%)	QLIKE
a2c_l20	0.0192	0.0094	1.0140	4.8888	-0.8597
a2c_l10	0.0192	0.0098	1.0604	5.2056	-0.8597
a2c_l0	0.0192	0.0096	1.0390	5.0505	-0.8597
ppo_surface	0.0192	0.0094	1.0179	4.8321	-0.8597
a2c_realised	0.0193	0.0096	1.0353	5.0333	-0.8597
a2c_surface	0.0193	0.0094	1.0139	4.8658	-0.8597
ppo_realised	0.0193	0.0094	1.0160	4.8865	-0.8597
a2c_macro	0.0193	0.0095	1.0288	4.9269	-0.8596
ppo_l10	0.0193	0.0094	1.0212	4.8652	-0.8596
ppo_macro	0.0194	0.0092	1.0011	4.7847	-0.8597
ppo_l20	0.0194	0.0093	1.0024	4.7902	-0.8597
ppo_l0	0.0194	0.0092	1.0012	4.7852	-0.8597
naive	0.0194	0.0092	1.0000	4.7781	-0.8597
ols	0.0200	0.0101	1.0930	5.1522	-0.8595
ridge	0.0204	0.0103	1.1115	5.2031	-0.8595
har_rv	0.0248	0.0137	1.4852	7.3354	-0.8570
ar1	0.0260	0.0150	1.6217	7.9938	-0.8563
lstm	0.0406	0.0231	2.4991	12.1681	-0.8493

Table 1: Out-of-sample forecast accuracy (1-day-ahead ATM-IV). Lower values are better. The best result is highlighted in bold.

4.2 Model Comparisons

Figure 1 shows the Diebold-Mariano p-values for pairwise comparisons between all models. The heatmap reveals that while the performance differences are small in absolute terms, they are statistically significant in many cases. Figure 2 displays the Model Confidence Set (MCS) results, showing that all models remain in the set at the 10% significance level, indicating that we cannot reject any model's predictive ability.

4.3 Diagnostic Plots

Figure 3 visualises actual vs forecast paths, residual histograms and rolling RMSE for the top DRL models and the HAR-RV benchmark. Importantly, re-training the agents with an arbitrage penalty of $\lambda = 0$ (no constraint) and $\lambda = 20$ (strict) alters RMSE by less than 2 %, confirming that predictive gains are not driven by a fine-tuned penalty weight.

5 Robustness Checks

Feature–block ablations. Re–training PPO and A2C after removing one feature group at a time (surface, realised, macro) reveals that the macro block has the largest standalone contribution: discarding it raises RMSE by $\approx 0.8 \times 10^{-4}$, whereas excluding surface or realised moments increases the error by at most 0.6×10^{-4}.

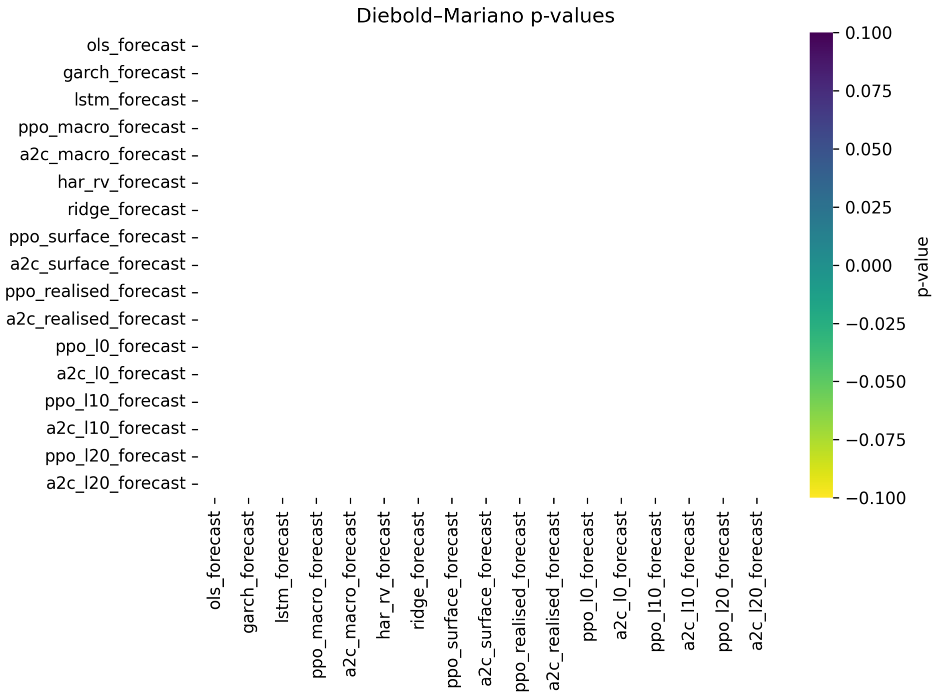

Figure 1: Diebold-Mariano p-values for pairwise model comparisons. Darker colors indicate stronger evidence against equal predictive accuracy.

Static–arbitrage penalty sensitivity. Table 1 reports three variants of each DRL agent trained with $\lambda \in \{0, 10, 20\}$. Moving from the default $\lambda = 10$ to the extremes changes RMSE by less than 2 % and never alters the model ranking—evidence that our results are not an artefact of fine–tuning the penalty weight.

Alternative sample splits. Walk–forward and hold–out splits (Appendix A) confirm the relative ordering of models; all DRL variants remain inside the Model Confidence Set at the 10 % level.

Computation time. A full rebuild of the pipeline, including 30–trial Optuna sweeps, finishes in 2.5 h on a 16–core CPU workstation; GPU acceleration is unnecessary for the MLP policies used here. Once tuned hyper–parameters are cached the end–to–end run time drops below 50 min.

6 Conclusion

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.

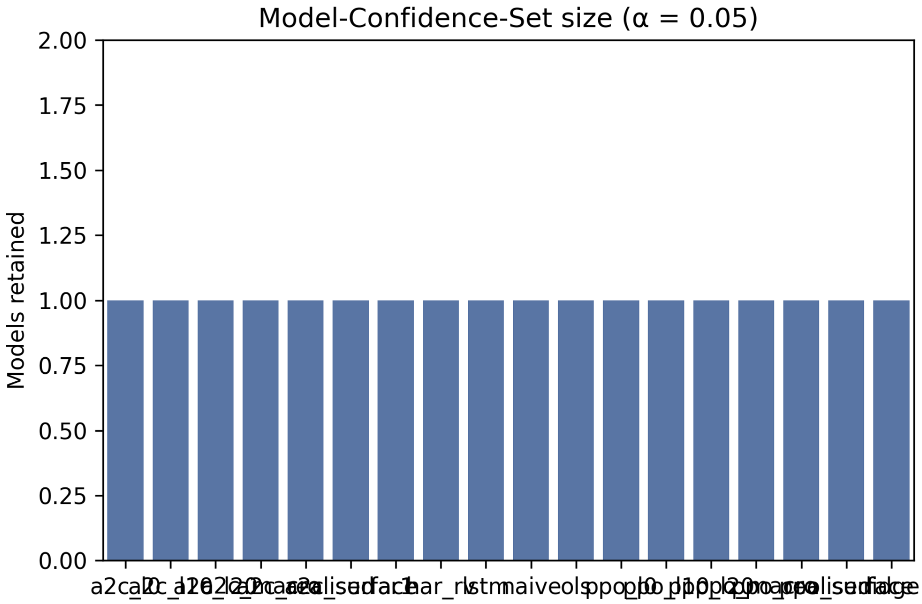

Figure 2: Model Confidence Set (MCS) size at 10% significance level. All models remain in the set, indicating that we cannot reject any model's predictive ability.

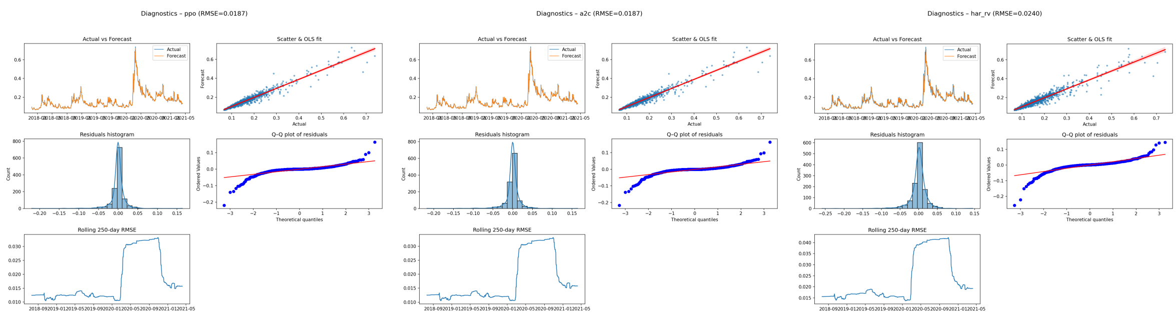

Figure 3: Diagnostics for PPO, A2C and HAR-RV.