DOI: 10.54861/27131211_2026_5_242

REINFORCEMENT LEARNING IN EQUITY TRADING: EVIDENCE FROM DEVELOPED AND FRONTIER MARKETS

Chimedtseren Tumendemberel
Business School, National University of Mongolia (NUM), Ulaanbaatar, Mongolia

Author: Tumendemberel Chimedtseren: 23m1num0629@stud.num.edu.mn

Dashnyam Bayarmaa
Business School, National University of Mongolia (NUM), Ulaanbaatar, Mongolia

Corresponding author: Dashnyam Bayarmaa bayarmaa_d@num.edu.mn
ORCID: https://orcid.org/0009-0008-9585-5621

Highlights

– Deep reinforcement learning is evaluated across developed and frontier equity markets.

– DQN converges toward near-passive behavior in efficient markets.

– Reinforcement learning significantly outperforms Buy-and-Hold in the Mongolian Stock Exchange.

– The findings provide empirical support for the Adaptive Markets Hypothesis.

– Market efficiency strongly influences the effectiveness of algorithmic trading strategies.

Abstract

Introduction: Understanding whether algorithmic trading strategies can generate consistent risk-adjusted returns remains a central question in financial markets, particularly across environments with different levels of market efficiency. This study develops a deep reinforcement learning framework for systematic equity trading and evaluates its performance in both a developed market (S&P 500 ETF) and a frontier market (Mongolian Stock Exchange). The study is motivated by the need to examine whether reinforcement learning strategies behave differently across markets characterized by varying degrees of informational efficiency and structural frictions.

Methods: The study implements a Deep Q-Network (DQN) reinforcement learning framework for sequential trading decisions. The model is trained and evaluated under identical experimental conditions using historical market data from SPY and a composite portfolio of actively traded Mongolian equities. The methodology incorporates chronological train-validation-test splits, technical indicator-based state representation, transaction cost modeling, and out-of-sample performance evaluation. Market efficiency is additionally examined through weak-form efficiency tests using return autocorrelation and Ljung–Box statistics.

Results: The empirical results demonstrate a clear contrast between developed and frontier markets. In the developed market, the reinforcement learning strategy provides only modest improvements in risk-adjusted performance relative to passive benchmarks, consistent with the Efficient Market Hypothesis. In contrast, the frontier market results show substantially higher returns and improved performance metrics for the DQN strategy, indicating the presence of exploitable inefficiencies and supporting the Adaptive Markets Hypothesis. The findings further reveal that the reinforcement learning agent converges toward near-passive behavior in efficient markets while exploiting return persistence in less efficient environments.

Discussion: The findings suggest that the effectiveness of reinforcement learning-based trading strategies is strongly dependent on market structure and informational efficiency. In highly efficient markets, adaptive strategies generate limited incremental value over passive investment approaches. Conversely, in frontier markets characterized by lower liquidity, weaker information dissemination, and structural inefficiencies, data-driven trading strategies may provide substantial advantages. From a practical perspective, the results highlight the potential value of reinforcement learning as a decision-support framework for investors and financial institutions operating in frontier market environments, while also emphasizing the importance of risk management due to higher volatility and drawdowns.

Keywords

Deep Reinforcement Learning; Systematic Trading; Frontier Markets; Market Efficiency; Adaptive Markets Hypothesis; Emerging Markets.

JEL classification: C45; C53; G11; G14; G15.

Conflict of interest. The authors declares that there is no Conflict of Interest.

For citation: Chimedtseren T., Dashnyam B. (2026). Reinforcement Learning in Equity Trading: Evidence from Developed and Frontier Markets. Progressivnaya ekonomika [Progressive Economy], 5, 242–259, https://doi.org/10.54861/27131211_2026_5_242 (In Eng.)

The article was submitted to the editorial office: 03/28/2026. Approved after review: 05/14/2026. Accepted for publication: 05/18/2026.

Introduction

Understanding whether algorithmic trading strategies can generate consistent risk-adjusted returns remains a central question in financial markets, particularly across environments with different levels of market efficiency. If markets are fully efficient, no trading strategy should be able to systematically outperform passive benchmarks. However, in less efficient markets, structural constraints, informational frictions, and behavioral biases may allow exploitable patterns to persist over time.

The Efficient Market Hypothesis (EMH) proposed by Fama [1] posits that asset prices reflect all available information, thereby limiting the potential for systematic excess returns. Nevertheless, a substantial body of empirical evidence documents deviations from this benchmark, particularly in markets characterized by thin liquidity, limited institutional participation, and retail driven behavior. To reconcile these findings, Lo [2] proposes the Adaptive Markets Hypothesis (AMH), which conceptualizes market efficiency as a dynamic and evolving property. Under this framework, markets become more efficient as participants adapt, while frontier markets may retain inefficiencies for longer periods due to weaker information environments and lower levels of algorithmic participation.

In modern financial markets, the development of trading strategies extends beyond return prediction to sequential decision-making under uncertainty. A trading strategy can be defined as a systematic decision framework that determines when to buy, sell, or hold assets, taking into account transaction costs, position sizing, and changing market conditions. Traditional approaches typically focus on forecasting price movements and treat execution as a separate problem. In contrast, reinforcement learning-based approaches directly optimize the decision-making process, enabling the joint consideration of prediction, execution, and portfolio dynamics.

Deep Reinforcement Learning (DRL) provides a flexible and effective framework for this purpose. By optimizing sequential portfolio decisions, DRL naturally incorporates transaction costs, position dynamics, and long-term reward maximization. The Deep Q-Network [3] has become a widely used baseline for discrete-action trading strategies. Unlike supervised learning models, which focus on predictive accuracy, reinforcement learning aligns the model objective with investor goals namely, maximizing cumulative risk-adjusted returns.

Despite these advances, an important question remains: do reinforcement learning-based trading strategies perform differently across markets with varying levels of efficiency? Empirical evidence on this question remains limited, particularly for frontier markets. The Mongolian Stock Exchange (MSE), characterized by relatively low liquidity, retail-dominated participation, and limited disclosure, provides a natural setting in which exploitable patterns may persist.

Motivated by this gap, this study develops a DRL based trading framework and evaluates its performance across a developed market (S&P 500 ETF) and a frontier market (MSE) under identical experimental conditions. This cross-market design allows for a direct empirical assessment of how market efficiency influences the effectiveness of algorithmic trading strategies.

This study makes three main contributions. First, it provides one of the first empirical applications of DRL in a frontier market setting. Second, by comparing developed and frontier markets under a unified framework, it quantifies the market efficiency gradient and provides empirical support for the AMH. Third, it demonstrates how reinforcement learning can be used as a decision-making framework in financial markets, with implications for strategy design and implementation.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the data and methodology. Section 4 presents empirical results. Section 5 discusses the findings and their implications. Section 6 concludes.

Market Efficiency and the Adaptive Markets Hypothesis

The EFH posits that asset prices fully reflect available information, implying that it is not possible to systematically achieve excess risk-adjusted returns using trading strategies based on historical data. Under the weak-form version of EMH, past price information is assumed to be fully incorporated into current prices, rendering technical analysis ineffective.

However, a substantial body of empirical research challenges this assumption. For example, DeBondt & Thaler [4] document long-run mean reversion in stock returns, while Jegadeesh and Titman [5] provide evidence of medium-term momentum effects. Similarly, Lo and MacKinlay [6] reject the random walk hypothesis at short horizons. Collectively, these findings suggest that financial markets may exhibit predictable patterns, particularly over certain time horizons and market conditions.

To reconcile these observations, Lo [2] proposes AMH, which conceptualizes market efficiency as a dynamic and evolving property. Under this framework, markets become more efficient as participants adapt, while frontier markets characterized by lower liquidity, limited institutional participation, and weaker information dissemination may retain exploitable inefficiencies for longer periods.

Recent empirical studies support this perspective Hanauer & Kalsbach [7] show that machine learning models generate statistically significant excess returns in emerging markets compared to traditional approaches, even after accounting for transaction costs. Similarly, Jargalsaikhan et al. [8]document return autocorrelation in the MSE, providing evidence consistent with weak-form inefficiency. These findings motivate the application of data-driven trading approaches in frontier market environments.

Reinforcement Learning in Financial Trading

Reinforcement learning (RL) has emerged as a powerful framework for financial decision-making due to its ability to model sequential and dynamic environments. Early work John & Mattew [9] demonstrates that trading agents can learn profitable strategies without explicit price prediction.

The development of DRL has further enhanced these capabilities. Mnih et al. [3] introduced key innovations such as experience replay and target networks, enabling stable learning in high-dimensional state spaces. As a result, DQN has become a widely used baseline for discrete-action trading strategies.

Empirical studies demonstrate the effectiveness of DRL in financial markets. Théate & Ernst [10] show that DRL-based trading agents can outperform passive benchmarks in terms of risk-adjusted returns. Similarly, Jeong & Kim [11] find that DQN improves trading performance relative to Buy-and-Hold and moving average strategies. Xu [12] highlights the importance of reward design and transaction cost modeling for achieving robust performance in trading environments.

More broadly, Hambly et al. [13] provide a comprehensive overview of DRL in finance, emphasizing that model performance depends critically on state representation, reward specification, and realistic evaluation procedures. They highlight the importance of out-of-sample testing and transaction cost integration elements that are explicitly incorporated in this study.

Machine Learning Benchmarks and Limitations

In parallel with reinforcement learning, supervised machine learning models have been widely applied to financial prediction tasks. Fischer & Krauss [14]demonstrate that Long Short-Term Memory (LSTM) networks achieve statistically significant out-of-sample directional accuracy in financial time series. Similarly, Gu et al. [15] show that neural networks can generate economically meaningful alpha in the U.S. equity market.

However, translating predictive accuracy into trading profitability remains a non-trivial challenge. A model with modest directional accuracy (e.g., slightly above 50%) may fail to generate economic profits once transaction costs, position sizing, and execution constraints are taken into account. This limitation highlights a key drawback of prediction-based approaches, which optimize intermediate objectives rather than final investment outcomes.

In contrast, reinforcement learning directly optimizes sequential decision-making and portfolio performance, aligning model objectives with investor goals. By integrating prediction, execution, and risk management into a unified framework, RL based approaches provide a more natural and potentially more effective solution for trading strategy development.

Research Gap and Contribution

Despite the growing literature on machine learning and reinforcement learning in finance, several important gaps remain. First, empirical evidence on DRL based trading in frontier markets is limited. Second, many studies lack transparency and reproducibility, often omitting key experimental details such as data splits, transaction cost assumptions, and statistical significance testing. Third, cross-market empirical comparisons between developed and frontier environments remain scarce.

To address these limitations, this study develops a transparent and reproducible experimental framework and applies a DRL approach to both developed and frontier markets under identical conditions. By doing so, it provides new empirical evidence on how market efficiency influences the effectiveness of algorithmic trading strategies and contributes to both financial economics and machine learning literature.

Data and Methodology

This study investigates the performance of reinforcement learning–based trading strategies across two fundamentally different market environments: a developed market and a frontier market. The developed market is represented by the SPDR S&P 500 ETF Trust (SPY), while the frontier market is proxied using a composite index constructed from the MSE.

The SPY dataset consists of daily adjusted closing prices obtained from Yahoo Finance, covering the period from November 2010 to November 2025. Adjusted prices are used to account for corporate actions such as stock splits and dividend reinvestment, ensuring consistency with total return calculations. Importantly, this dataset spans multiple market regimes including bull markets, crisis periods, and post-crisis recoveries providing a robust environment for evaluating trading strategies under varying conditions.

In contrast, the MSE dataset reflects the structural characteristics of a frontier market. It comprises daily OHLCV (open, high, low, close, volume) data for listed securities from January 2021 to February 2026. To ensure that the analysis is both representative and practically implementable, a 20-stock equal-weight composite portfolio is constructed based on the most actively traded securities during the training period. This design avoids look-ahead bias and reflects realistic trading conditions in a low-liquidity environment.

Although the datasets differ in length due to data availability constraints, both markets are evaluated under a consistent experimental design. Specifically, the goal is not to compare absolute performance levels across markets, but rather to examine how trading strategies behave under different market structures.

To ensure methodological rigor, both datasets are partitioned using a strict chronological split into training (70%), validation (10%), and test (20%) sets. This time-based approach reflects real-world trading conditions and prevents information leakage. Furthermore, all feature transformations and normalization parameters are estimated using training data only and then applied to validation and test sets.

Finally, a warm-up period is introduced at the beginning of each dataset to account for rolling technical indicators. This ensures that all features are fully defined prior to model training, thereby maintaining consistency and avoiding initialization bias.

Weak-Form Efficiency Tests

Under the EMH, market efficiency is classified into three forms weak, semi-strong, and strong based on the extent to which information is reflected in asset prices. The weak form implies that past price information is already incorporated into current prices, suggesting that historical data should not provide systematic forecasting power.

To evaluate this assumption, this study tests weak form efficiency by examining whether historical price information contains predictive signals in financial markets using the Ljung–Box portmanteau test applied to the full period return series. Evidence of weak form inefficiency would indicate the presence of exploitable return patterns, thereby justifying the application of algorithmic trading strategies.

The results reveal a clear contrast between the two environments. The MSE composite exhibits statistically significant positive autocorrelation at short lags, indicating the presence of predictable return patterns. The Ljung–Box test strongly rejects the null hypothesis of zero autocorrelation, providing evidence of weak-form inefficiency. In contrast, the SPY dataset shows negative first-order autocorrelation, a phenomenon commonly attributed to market microstructure effects such as bid–ask bounce. Although statistically significant at certain lags, these effects are economically small and do not imply exploitable predictability.

These findings establish a clear empirical distinction between developed and frontier markets, providing strong motivation for applying adaptive trading strategies such as reinforcement learning.

ISSN 2713-1211 (online)
https://progressive-economy.ru/

Table 1. Return Autocorrelation and Ljung–Box Test Results

Lag	MSE AC	MSE p	SPY AC	SPY p	Interpretation
1	0.111	0.008**	−0.109	<0.001***	MSE: positive momentum; SPY: microstructure noise
2	0.116	0.003**	0.067	<0.001***	MSE: persistent autocorrelation
3	0.063	0.039*	−0.047	0.004**	MSE: signal fading
5	−0.005	0.853	−0.002	0.909	Both markets: no signal at lag 5
Q(10)	45.45	0.007**	171.99	<0.001***	MSE rejects random walk; SPY (microstructure)

***Note:** *p < 0.05; **p < 0.01; ***p < 0.001.*

Source: compiled by the researchers

Feature Engineering

To represent the market environment, a 10-dimensional state vector is constructed using only contemporaneously available information. This ensures that no future information is incorporated into the model, thereby eliminating look-ahead bias and maintaining a strictly causal framework.

Formally, the market state at time $t$is defined as:

$$s_t = [r_t^1, r_t^5, r_t^{21}, \mathrm{Trend}_{10,t}, \mathrm{Trend}_{50,t}, \mathrm{Trend}_{200,t}, \mathrm{RSI}_t, \mathrm{Vol}_t, \mathrm{VolRatio}_t, \mathrm{Pos}_t]$$

This feature set is designed to capture key financial dynamics, including return momentum, trend behavior, volatility clustering, and trading activity.

First, short- and medium-term return dynamics are captured using multi-horizon log returns, defined as:

$$r_t^k = \ln\left(\frac{P_t}{P_{t-k}}\right), k \in \{1,5,21\}$$

where $P_t$denotes the asset price at time t. These features allow the model to detect momentum effects across different time scales.

Second, trend information is incorporated through price-to-moving-average ratios. The simple moving average (SMA) over a window of length $k$is defined as:

$$\mathrm{SMA}_{k,t} = \frac{1}{k}\sum_{i=0}^{k-1} P_{t-i}$$

and the corresponding trend indicator is constructed as:

$$\mathrm{Trend}_{k,t} = \frac{P_t}{\mathrm{SMA}_{k,t}} - 1, k \in \{10,50,200\}$$

These features measure deviations of the current price from its historical average, capturing both short-term and long-term trend behavior.

Third, momentum oscillation is captured using the Relative Strength Index (RSI), defined as:

$$\mathrm{RSI}_t = 100 - \frac{100}{1 + RS_t}$$

where RS_trepresents the ratio of average gains to average losses over a 14-day window. The RSI provides information on overbought and oversold conditions.

Fourth, market risk is proxied using rolling volatility, computed as the standard deviation of returns over a 21-day window:

$$\mathrm{Vol}_t = \sqrt{\frac{1}{21}\sum_{i=0}^{20}(r_{t-i} - \bar{r})^2}$$

This feature captures volatility clustering, a well-documented stylized fact in financial markets.

Fifth, trading activity is represented using a volume-based indicator defined as:

$$\mathrm{VolRatio}_t = \frac{V_t}{\frac{1}{21}\sum_{i=0}^{20} V_{t-i}}$$

where V_t denotes the trading volume at time t. This feature reflects abnormal trading activity and potential information flow.

Finally, the agent's current position is included as a binary state variable:

$$\mathrm{Pos}_t \in \{0,1\}$$

where 1 indicates that the asset is held and 0 otherwise. Including the position variable allows the model to incorporate its previous decisions into the current state, which is essential for sequential decision-making.

All features are standardized using statistics computed from the training set, and the same transformation is applied to validation and test sets. This ensures consistency across datasets and prevents information leakage.

Overall, the proposed feature set provides a compact yet information-rich representation of the market state, enabling the reinforcement learning agent to jointly capture return dynamics, trend structure, risk conditions, and trading activity.

Reinforcement Learning Framework

The trading problem is formulated as a Markov Decision Process [16], where the agent interacts with the market environment over time.

At each time step, the agent observes the current market state $s_t$and selects an action from a discrete action space $A = \{0\text{:Hold/Cash}, 1\text{:Buy}, 2\text{:Sell}\}$. The objective is to maximize cumulative risk-adjusted returns over time.

The reward function is defined as:

$$r_t = r_t^p - c \cdot I(a_t \neq a_{t-1}),$$

where r_t^pdenotes the portfolio return and $c$represents transaction costs. Transaction costs are set to 0.1% per trade for SPY and 0.3% for MSE, reflecting higher trading frictions in frontier markets. This formulation ensures that the model internalizes trading costs and avoids excessive turnover.

Training is conducted using episodes of length $T = 252$ trading days, with each episode initialised from a cash-only portfolio to ensure consistency across simulations.

The model is implemented using a DQN, consisting of a three-layer fully connected neural network with a 10-dimensional input layer, two hidden layers (128 and 64 units with ReLU activation), and a linear output layer producing Q-values for each action.

The network is trained using the Adam optimizer (learning rate = 0.001) with a discount factor $\gamma = 0.95$. To stabilise training, experience replay (buffer size = 10,000; batch size = 64) and a target network updated every 10 episodes are employed. Exploration follows an ε-greedy strategy, with ε decaying from 1.0 to 0.01 at a rate of 0.995 per episode over 500 training episodes. All hyperparameters are summarized in Table 2.

ISSN 2713-1211 (online)
https://progressive-economy.ru/

Table 2. DQN Hyperparameter Configuration

Hyperparameter	Value	Rationale
Learning rate (α)	0.001	Commonly used default for stable optimization
Discount factor (γ)	0.95	Moderate weighting of future rewards
Replay buffer size	10,000	Sufficient history for batch sampling
Batch size	64	Computational efficiency
ε (start → end)	1.0 → 0.01	Gradual transition from exploration to exploitation
ε decay rate	0.995 per episode	Controlled reduction in exploration
Target network update	Every 10 episodes	Stabilizes learning targets
Training episodes	500	Empirically sufficient for convergence
Random seeds	{42, 123, 456, 789, 1024}	Multiple runs to assess stochastic variability

Note. *Architecture: three-layer fully connected network (128–64 units, ReLU). The Adam optimizer is used for training. Identical hyperparameters are applied to both SPY and MSE datasets.*

Source: compiled by the researchers

Performance Metrics

To evaluate the performance of trading strategies, this study employs a comprehensive set of metrics that jointly capture both return and risk dimensions [2]. Return on Investment (ROI) measures the total return generated over the evaluation period. To enable comparability across different time horizons, ROI is further converted into an annualized return [17]. The *Sharpe ratio* [18] is used as the primary risk-adjusted performance metric, assessing excess return relative to volatility after adjusting for the risk-free rate. *Maximum Drawdown (MDD*) quantifies the largest peak-to-trough decline in portfolio value, providing a measure of downside risk [19]. The *Calmar ratio* evaluates the trade-off between return and risk by comparing annualized returns to maximum drawdown[20]. *Win rate* [21] represents the proportion of profitable trades, offering a simple measure of strategy consistency, while Jensen's alpha [22] captures the excess return generated by the strategy relative to the market benchmark, reflecting abnormal performance. By jointly considering these metrics, the evaluation framework provides a holistic assessment of trading strategies, capturing not only profitability but also risk exposure and stability [23].

The overall performance of the reinforcement learning model is first evaluated against benchmark strategies in the developed market environment.

Developed Market Performance (SPY)

Table 3 reports the out-of-sample performance on the SPY dataset (756 trading days). The Buy-and-Hold (BH) strategy achieves the highest total return (88.7%), reflecting the strong upward trend of the developed market during the test period.

The DQN model achieves a comparable total return (84.7% ± 3.2%) but delivers a higher Sharpe ratio (1.363 ± 0.04) than BH (1.152), indicating superior risk-adjusted performance. The model maintains market exposure for approximately 77.9% of trading days, effectively behaving as a near-passive strategy with dynamic cash allocation.

However, the maximum drawdown (−18.7%) is identical to BH, suggesting that while DQN reduces volatility, it does not significantly improve downside protection during major market downturns.

The Moving Average (MA 50/200) strategy substantially underperforms (ROI: 44.8%), primarily due to lag effects and whipsaw losses in trending conditions.

Table 3. Out-of-Sample Performance -SPY Developed Market (2022–2025; 756 trading days)

Strategy	ROI (%)	Ann. ROI (%)	Sharpe Ratio	MDD (%)	Calmar	Win Rate	α (p.a.)
Buy-and-Hold	88.7	23.6	1.152	−18.7	1.257	54.2%	0.0%
DQN (mean ± SD)	84.7±3.2	22.7	1.363±0.04	−18.7	1.214±0.05	53.8%	+4.9%
MA Crossover (50/200)	44.8	13.1	1.187	−11.7	1.12	51.1%	—

Note. *rf = 4.5% p.a. Ann. ROI = (1 + ROI_total)^(252/756) − 1. Sharpe = (μ − rf/252)/σ × √252. Calmar = Ann. ROI / |MDD|. DQN results are reported as mean ± standard deviation over 5 random seeds.*

Source: compiled by the researchers

Bootstrap Significance Tests

Bootstrap significance tests provide important evidence regarding market efficiency. Table 4 reports bootstrap-based statistical tests using both independent and identically distributed (*i.i.d.*) sampling and circular block bootstrap (CBB). The *i.i.d.* approach assumes independence and identical distribution of observations, ignoring temporal dependence and thus serving as a baseline benchmark. In contrast, CBB preserves time dependence by resampling contiguous blocks of data, providing more reliable statistical inference for financial time series.

In the developed market (SPY), the difference between DQN and Buy-and-Hold is not statistically significant at the conservative $\alpha = 0.01$ level (CBB $p = 0.024$), consistent with the Efficient Market Hypothesis. This suggests that any performance advantage of adaptive strategies is relatively small and difficult to distinguish from sampling variation in a highly efficient market. However, DQN significantly outperforms the Moving Average strategy ($p < 0.001$), indicating that simple rule-based strategies are insufficient even in developed market conditions.

In the frontier market (MSE), the ROI advantage of DQN over Buy-and-Hold is statistically significant (CBB p = 0.031), providing evidence of exploitable inefficiencies. Although the Sharpe comparison does not reach statistical significance (CBB p = 0.109), this result is constrained by limited sample size.

Table 4. Bootstrap Significance Tests — SPY and MSE

Comparison	i.i.d. t	i.i.d. p	CBB p (b=12)	Significant?	Interpretation
Panel A: SPY (Developed Market)					
BH vs. DQN	2.56	0.010	0.024	No (α=0.01)	Consistent with EMH
BH vs. MA (50/200)	29.42	<0.001	<0.001	Yes (α=0.001)	MA significantly inferior
DQN vs. MA (50/200)	18.73	<0.001	<0.001	Yes (α=0.001)	DQN dominates MA
Panel B: MSE (Frontier Market)					
BH vs. DQN (Sharpe)	1.76	0.081	0.109	No (underpowered)	Short test window (~254 obs.)
BH vs. DQN (ROI)	2.41	0.017	0.031	Yes (α=0.05)	ROI advantage significant

Note. *CBB = circular block bootstrap (primary inference method). H_0: equal population Sharpe ratios. Two-sided t-tests are conducted between bootstrap distributions. Significance thresholds: α = 0.05 (standard) and α = 0.01 (conservative). The MSE Sharpe test is underpowered due to the relatively short test window (~254 observations); therefore, ROI-based comparison is used as the primary inferential basis for the MSE.*

Source: compiled by the researchers

Frontier Market Performance (MSE)

Table 5 presents the out-of-sample performance on the MSE dataset (254 trading days).

DQN generates substantially higher total return than Buy-and-Hold (141.5% vs. 41.0%), representing more than a threefold increase in return. The Sharpe ratio of DQN (0.891) is only marginally higher than BH (0.833), suggesting that the primary advantage lies in return generation rather than risk efficiency.

However, the maximum drawdown of DQN (−38.4%) is significantly higher than BH (−12.3%), reflecting more aggressive position-taking behaviors. Additionally, the relatively high variability across random seeds (SD: 18.7%) indicates sensitivity to training conditions.

The Moving Average strategy performs poorly (ROI: 5.0%), largely failing to capture the strong upward trend in the frontier market.

Table 5. Out-of-Sample Performance (MSE Frontier Market)

Strategy	ROI (%)	Ann. ROI (%)	Sharpe Ratio	MDD (%)	Calmar	Win Rate	α (p.a.) †
Buy-and-Hold	41.0	40.6	0.833	−12.3	3.301	53.1%	0.0%
DQN (mean ± SD)	141.5±18.7	~140%	0.891±0.12	−38.4	3.641±0.20	51.8%	+109%
MA Crossover (50/200)	5.0	~5.0	0.821	−5.3	0.936	49.7%	—

Note. *rf = 12.0% p.a. (Bank of Mongolia policy rate). Annualized return is computed as Ann. ROI = (1 + ROI_total)^(252/254) − 1. Calmar ratio is defined as Ann. ROI / |MDD|. DQN results are reported as mean ± standard deviation over five random seeds. † Jensen's alpha is approximated using time-in-market as a proxy for beta and should be interpreted with caution.*

Source: compiled by the researchers

Results Interpretation

The results indicate that no single strategy consistently dominates all market environments. In the developed market, the DQN model converges toward a near-passive strategy, effectively replicating Buy-and-Hold behavior with slight improvements in risk-adjusted performance. This outcome is consistent with the Efficient Market Hypothesis, where opportunities for systematic outperformance are limited. In contrast, in the frontier market, the DQN model demonstrates substantial performance gains, suggesting the presence of persistent and exploitable market inefficiencies. This aligns with AMH, which predicts that less efficient markets provide greater opportunities for adaptive strategies. Importantly, these findings highlight the role of market structure in determining the effectiveness of algorithmic trading strategies. Static strategies perform well in stable environments, while adaptive strategies provide advantages in more volatile and evolving markets. This framework provides a natural empirical test of the AMH.

Market Efficiency Gradient

The cross-market results provide direct quantitative evidence for the efficiency gradient implied by AMH.

In the developed market (SPY), the DQN converges to a near-passive policy: through trial-and-error learning, the agent identifies that remaining invested maximizes value, consistent with the Efficient Market Hypothesis. The bootstrap result (CBB $p = 0.024$, not significant at the conservative $\alpha = 0.01$ level) is the expected outcome in a highly efficient market.

Importantly, this should not be interpreted as a limitation of the reinforcement learning framework. Rather, it demonstrates its validity: an agent that correctly identifies passivity as optimal in an efficient market provides a strong validation of the learning process.

In contrast, in the frontier market (MSE), the DQN exhibits a substantial return advantage (141.5% vs. 41.0%) along with a slightly higher Sharpe ratio (0.891 vs. 0.833), indicating the presence of exploitable return patterns. Autocorrelation evidence (e.g., short-horizon persistence) provides a plausible mechanism, suggesting that the agent captures momentum dynamics that are more likely to persist over time in less efficient markets.

The cost sensitivity analysis further highlights this asymmetry: while DQN maintains its advantage over Buy-and-Hold across realistic transaction cost ranges in the MSE, the advantage diminishes in the SPY under realistic trading frictions. This provides a direct empirical quantification of the efficiency gradient.

Frontier Market Structural Challenges

The performance of reinforcement learning models in frontier markets is shaped by several structural constraints. First, data scarcity limits model training. Compared to developed markets, the shorter historical dataset results in higher variability across model runs, reducing stability. Second, liquidity constraints affect execution. Transaction cost assumptions may underestimate real-world trading frictions, particularly for larger positions, although sensitivity analysis suggests that the model remains robust within realistic ranges. Third, structural non-stationarity is more pronounced in frontier markets. Economic cycles, regulatory changes, and evolving market structures can alter return patterns over time, reducing the persistence of learned behaviors. These factors imply that continuous model updating and adaptive retraining are essential for practical deployment in such environments.

Practical Implications

The findings have important implications for investors and policymakers.

For institutional investors in emerging markets, a phased adoption of reinforcement learning is recommended. Initial implementation may involve using RL-generated signals as decision support, followed by gradual integration into automated strategies as reliability improves.

From a risk management perspective, combining multiple independently trained models (ensemble approaches) can help reduce variability and improve robustness, analogous to portfolio diversification. For conservative investors, passive strategies such as Buy-and-Hold remain appropriate in efficient markets. In contrast, for quantitative investors operating in less efficient markets, adaptive strategies such as DQN offer meaningful advantages.

From a policy perspective, the results suggest that the primary constraint on advanced algorithmic trading in frontier markets is informational rather than technological. Investments in data infrastructure such as higher-frequency data, longer historical records, and standardized reporting can significantly enhance market efficiency and attract institutional participation.

Conclusion

This study develops a DRL framework for systematic equity trading and evaluates its performance across markets with different levels of efficiency, specifically a developed market (SPY) and a frontier market (Mongolian Stock Exchange). The empirical findings reveal a clear contrast between the two market environments. In the developed market, the DQN model converges toward a near-passive strategy, delivering only marginal improvements in risk-adjusted performance relative to the Buy-and-Hold benchmark. This outcome is consistent with the EMH, suggesting that opportunities for systematic outperformance are limited in highly efficient markets.

In contrast, the results from the frontier market demonstrate substantial performance gains for the DRL-based strategy. The DQN model significantly outperforms the Buy-and-Hold strategy in terms of total return, providing evidence of persistent and exploitable inefficiencies. These findings support the AMH, which posits that market efficiency evolves over time and varies across market structures. Importantly, the study highlights that the effectiveness of reinforcement learning-based trading strategies is strongly dependent on market conditions. While adaptive strategies offer limited advantages in highly efficient environments, they can generate significant value in markets characterized by structural frictions, limited information efficiency, and evolving dynamics.

From a practical perspective, the findings suggest that reinforcement learning can serve as a valuable decision-support tool for investors operating in frontier markets. However, the results also underscore the importance of risk management, as higher returns are accompanied by increased drawdowns and variability. This study contributes to the growing literature on artificial intelligence in finance by providing one of the first empirical applications of DRL a frontier market context. It also offers a unified framework for comparing trading strategies across different market environments, thereby providing a direct empirical test of AMH.

Several limitations should be acknowledged. First, the relatively short time horizon of the frontier market dataset may affect the robustness of statistical inference. Second, transaction cost assumptions may not fully capture real-world market frictions. Third, the analysis is limited to a single asset per market, which may constrain generalizability. Future research may extend this framework by incorporating multi-asset portfolio optimization, higher-frequency data, and more advanced reinforcement learning architectures. Additionally, integrating risksensitive objectives and ensemble approaches could further improve robustness and real-world applicability.

References

1. Fama E.F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance. Vol. 25, No. 2. P. 383–417. https://doi.org/10.2307/2325486

2. Lo A.W. (2004). The Adaptive Markets Hypothesis. Journal of Portfolio Management. Vol. 30, No. 5. P. 15–29. https://doi.org/10.3905/jpm.2004.442611

3. Mnih V., Kavukcuoglu K., Silver D. et al. (2015). Human-Level Control Through Deep Reinforcement Learning. Nature. Vol. 518, No. 7540. P. 529–533. https://doi.org/10.1038/nature14236

4. DeBondt W.F.M., Thaler R.H. (1985). Does the Stock Market Overreact? Journal of Finance. Vol. 40, No. 3. P. 793–808. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

5. Jegadeesh N., Titman S. (1993). Returns to Buying Winners and Selling Losers. Journal of Finance. Vol. 48, No. 1. P. 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

6. Lo A.W., MacKinlay A.C. (1988). Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. Review of Financial Studies. Vol. 1, No. 1. P. 41–66. https://doi.org/10.1093/rfs/1.1.41

7. Hanauer M.X., Kalsbach L. (2023). Machine Learning and the Cross-Section of Emerging Market Stock Returns. Emerging Markets Review. Vol. 55. P. 101022. https://doi.org/10.1016/j.ememar.2023.101022

8. Jargalsaikhan E., Ganbold D., Park C., Kim Y.-R., Kim J. (2024). Emerging Insights: Unveiling Market Efficiency in Mongolia’s Transforming Economy. Journal of Eastern European and Central Asian Research. Vol. 11, No. 6. P. 1030–1042. https://doi.org/10.15549/jeecar.v11i6.1834

9. John M., Mattew S. (1998). Performance Functions and Reinforcement Learning for Trading Systems and Portfolios. Journal of Forecasting. Vol. 17. P. 441–470.

10. Théate T., Ernst D. (2021). An Application of Deep Reinforcement Learning to Algorithmic Trading. Expert Systems with Applications. Vol. 173. P. 114632. https://doi.org/10.1016/j.eswa.2021.114632

11. Jeong G., Kim H.Y. (2019). Improving Financial Trading Decisions Using Deep Q-Learning. Expert Systems with Applications. Vol. 117. P. 125–138. https://doi.org/10.1016/j.eswa.2018.09.030

12. Xu Y. (2023). Deep Reinforcement Learning for Trading Strategy Development on High-Frequency Currency Data. University of Essex. https://doi.org/10.5526/ERR-00042298

13. Hambly B., Xu R., Yang H. (2023). Recent Advances in Reinforcement Learning in Finance. Mathematical Finance. Vol. 33, No. 3. P. 437–503. https://doi.org/10.1111/mafi.12382
14. Fischer T., Krauss C. (2018). Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. European Journal of Operational Research. Vol. 270, No. 2. P. 654–669. https://doi.org/10.1016/j.ejor.2017.11.054
15. Gu S., Kelly B., Xiu D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies. Vol. 33, No. 5. P. 2223–2273. https://doi.org/10.1093/rfs/hhaa009
16. Garcia F., Rachelson E. (2013). Markov Decision Processes. In: Markov Decision Processes in Artificial Intelligence. Wiley. P. 1–38. https://doi.org/10.1002/9781118557426.ch1
17. Bodie Z., Kane A., Marcus A.J. (2018). Investments. 11th ed. McGraw-Hill.
18. Sharpe W.F. (1966). Mutual Fund Performance. The Journal of Business. Vol. 39, No. S1. P. 119–138. https://doi.org/10.1086/294846
19. Magdon-Ismail M., Atiya A.F., Pratap A., Abu-Mostafa Y.S. (2004). On the Maximum Drawdown of a Brownian Motion. Journal of Applied Probability. Vol. 41, No. 1. P. 147–161.
20. Young T.W. (1991). Calmar Ratio: A Smoother Tool. Futures. Vol. 20, No. 1. P. 40–44.
21. Chan E.P. (2013). Algorithmic Trading: Winning Strategies and Their Rationale. Wiley.
22. Jensen M.C. (1968). The Performance of Mutual Funds in the Period 1945–1964. Journal of Finance. Vol. 23, No. 2. P. 389–416. https://doi.org/10.1111/j.1540-6261.1968.tb00815.x
23. Fabozzi F.J., Focardi S.M., Kolm P.N. (2014). Quantitative Equity Investing. Wiley.