Section of the conference: Information Technology

Sekcja konferencji: Technologia informacyjna

How to cite: Christadoss, J., Panda, M. R., Samal, B. V., & Wali, G. (2025). Development of a multi-objective optimisation framework for risk-aware fractional investment using reinforcement learning in retail finance. *World Conference on Emerging Science, Innovation and Policy 2025*. Futurity Research Publishing. https://doi.org/10.5281/zenodo.16667023

Development of a Multi-Objective Optimisation Framework for Risk-Aware Fractional Investment Using Reinforcement Learning in Retail Finance

Jessy Christadoss[1] Dr. Manas Ranjan Panda[2], Bhakta Vaschal Samal[3] , Girish Wali[4]

[1]*Independent Researcher, Senior Quality Engineer, Integral Ad Science, USA*

[2]*Independent Researcher, Partner, Wipro Limited, USA*

[3]*Independent Researcher, Govt. of Odisha, Odisha, India*

[4]*Independent Researcher, SVP, Citi Bank, USA*

Accepted: July 12, 2025 | **Published**: July 20, 2025 | **Language:** English

Abstract: Retail investors often face a complex trade-off between maximising returns, minimising risk, and achieving portfolio diversification, especially in the context of fractional investment. This study proposes a novel multi-objective optimisation framework employing reinforcement learning (RL) to address these competing goals in retail finance. The framework integrates a risk-aware reward function with dynamic portfolio rebalancing strategies to allow fractional investment across multiple assets. By utilising deep Q-learning and Proximal Policy Optimisation (PPO), the system learns optimal policies under varying market conditions, incorporating investor risk preferences and transaction constraints. Empirical testing was conducted on historical financial data spanning equities, ETFs, and cryptocurrencies. The results demonstrate superior portfolio performance in terms of Sharpe ratio, maximum drawdown, and diversification entropy when compared to traditional portfolio allocation techniques. Our framework paves the way for democratised, intelligent retail investment platforms that are both adaptive and aligned with user-specific financial goals.

Keywords: Reinforcement learning, fractional investment, retail finance, multi-objective optimisation, portfolio management, risk-aware strategies

1. Introduction

The rapid democratisation of financial markets has led to a surge in retail investor participation, driven by the widespread availability of online trading platforms, mobile investment apps, and access to fractional investing. Fractional investment, which allows individuals to purchase partial shares of high-value assets, has significantly lowered the entry barrier to diversified portfolio construction. However, this increased accessibility introduces new challenges: retail investors often lack the resources, expertise, or computational tools required to optimise investment decisions across multiple objectives such as risk management, return maximisation, and diversification. Traditional portfolio optimisation models, such as Markowitz's Mean-Variance Optimisation or Risk Parity, are often ill-suited for this context due to their assumption of idealised conditions and their inability to adapt dynamically to changing market environments.

To this gap, artificial intelligence-particularly reinforcement learning (RL)-has emerged as a promising solution. RL frameworks are well-suited to portfolio optimisation problems because they can learn to make sequential decisions under uncertainty and adaptively adjust strategies based on feedback from the market environment. When applied effectively, RL enables investors to dynamically balance competing objectives, such as maximising long-term portfolio return while minimising exposure to downside risk or market volatility. In this research, we explore the use of RL to design a risk-aware fractional investment system capable of learning optimal allocation policies from historical market data. Unlike single-objective approaches, we adopt a multi-objective optimisation framework that encodes investor preferences as a weighted combination of performance metrics within the RL reward function.

The rise of fractional investing necessitates the modelling of continuous action spaces, where capital can be allocated in any proportion rather than in whole units. Most classical RL methods operate in discrete action domains, making them insufficient for real-world financial tasks involving fractional asset purchase, transaction costs, and budget constraints. To overcome this, we incorporate state-of-the-art RL algorithms such as Proximal Policy Optimisation (PPO), which is capable of operating in high-dimensional continuous action spaces while maintaining training stability.

This choice enables the system to execute fine-grained adjustments to portfolio weights in response to market changes, risk signals, and investor preferences. This study contributes to the evolving field of financial machine learning by proposing an integrated architecture that combines market simulation, reward engineering, and policy learning under a unified, risk-aware multi-objective framework. Through rigorous backtesting on a diversified set of asset classes-including equities, ETFs, and cryptocurrencies empirically demonstrate that the proposed RL-based strategy outperforms conventional allocation methods on key financial performance indicators. The broader implication of this work is the creation of intelligent, autonomous investment agents that empower retail investors with institutional-grade portfolio optimisation capabilities tailored to their risk tolerance and financial goals.

2. Objective and Problem Statement

2.1 Capital Allocation Through Fractional Investment

The first objective of this research is to enable fractional capital allocation across multiple financial assets within a unified reinforcement learning (RL) framework. Traditional portfolio optimisation techniques often assume that investments occur in whole units of assets, which restricts retail investors with limited capital from achieving broad diversification. Fractional investment resolves this issue by allowing capital to be distributed in arbitrary proportions-even fractions of a single share-thereby unlocking access to high-value assets such as technology stocks or blue-chip ETFs. The proposed RL system is designed to work in this fractional domain, using a continuous action space to represent percentage allocations across assets. This approach closely mirrors real-world trading behaviour in modern investment platforms such as Robinhood, M1 Finance, or Fidelity.

By adopting a fractional investment perspective, the model enhances its adaptability to a wider user base with heterogeneous investment profiles. It allows the RL agent to deploy capital dynamically, with the flexibility to reallocate small portions of wealth in response to subtle market signals. This is especially critical in volatile or low-liquidity environments where making significant reallocations may incur slippage or transaction penalties. The use of continuous control via algorithms like Proximal Policy Optimisation (PPO) empowers the system to learn nuanced trading behaviour, thereby producing smoother portfolio transitions and efficient use of capital. Ultimately, fractional allocation aligns well with the objectives of retail investors who are increasingly seeking personalized, algorithm-driven investment strategies without the need for large capital commitments.

2.2 Integration of Risk Preferences and Market Volatility

The next objective is to incorporate investor-specific risk preferences and market volatility signals into the portfolio optimisation process. Traditional RL models in finance often optimise solely for return, which may result in excessively risky or volatile strategies that are unsuitable for risk-averse users. To overcome this, the proposed framework embeds risk-awareness directly into the reward function using penalty terms and probabilistic loss estimation. For example, portfolio variance, downside deviation, and conditional value-at-risk (CVaR) can be integrated as regularisation components in the agent's learning signal. By adjusting the weightings of these components, the system becomes tunable to different investor profiles-ranging from aggressive to conservative-making it more personalised and ethically aligned with financial advisement standards.

This model is designed to be responsive to changing market conditions, especially fluctuations in volatility regimes. Through dynamic state representation, it includes key volatility indicators such as rolling standard deviations, implied volatility indices (e.g., VIX), and autoregressive volatility forecasts. This temporal modelling enables the agent to recognise turbulent periods and shift toward more defensive asset allocations. Moreover, market-specific risk factors such as sector rotation or macroeconomic shocks can be captured via auxiliary features, allowing the system to condition its policies on real-time risk levels. The integration of investor risk appetite with exogenous volatility inputs ensures that portfolio decisions are not only return-seeking but also contextually aware, safeguarding capital in adverse conditions while seizing opportunities during stable periods.

2.3 Multi-Objective Optimisation: Return, Risk, and Diversification

The overarching objective is to frame the RL portfolio optimisation task as a multi-objective problem that seeks to maximise return, minimise risk, and promote diversification simultaneously. Unlike single-objective financial models that focus exclusively on expected return or risk-adjusted return, the proposed framework uses a composite reward function that integrates three distinct objectives. The return component is based on logarithmic or cumulative return; the risk component penalises excessive volatility or drawdown; and the diversification component rewards higher entropy in asset weight distributions. By structuring the reward function in this way, the RL agent is encouraged to find a Pareto-optimal balance among competing investment goals.

This multi-objective structure is operationalised through a scalarization technique, where each objective is assigned a tunable weight (α for return, β for risk, γ for diversification). The weights can be adjusted according to user preferences or scenario-based tuning during model calibration. This allows the system to learn flexible policies that generalise across multiple investment contexts-from high-growth environments where returns dominate, to crisis scenarios where capital preservation and diversification are critical. Furthermore, incorporating diversification explicitly helps reduce correlation risks and ensures that the portfolio does not become overly concentrated in a few volatile assets. The outcome is a robust, generalizable, and investor-aligned RL-based portfolio manager capable of adapting to modern retail finance demands.

3. Methodology and Framework Design

The proposed system utilises deep reinforcement learning (RL) to create an intelligent, risk-aware portfolio management framework for fractional investment. At its core, the architecture combines a simulated trading environment, a multi-objective reward function, a continuous action space, and a deep policy network that learns optimal portfolio decisions over time.

3.1 Environment Setup

The environment simulates a realistic financial market using historical asset data sourced from platforms such as Yahoo Finance (for equities and ETFs) and CoinGecko (for cryptocurrencies). It operates at a daily resolution, providing a balance between signal richness and computational tractability. The environment is designed to mimic trading conditions, including asset price movements, slippage, and capital constraints. Each episode in the simulation corresponds to a fixed time window (e.g., 3 years of data), within which the reinforcement learning agent interacts by making allocation decisions and receiving feedback through a reward function. This simulated environment also incorporates portfolio rebalancing and transaction costs, thereby closely modelling the operational constraints that retail investors face. The training phase includes random sampling of historical periods to avoid overfitting to a particular market regime.

To ensure robustness, the environment is designed with stochastic elements, such as introducing noise to price changes or simulating delayed market reactions. By training the RL agent within this environment, the model learns strategies that can generalise across different asset classes and market conditions, providing a realistic and adaptable decision-making framework.

3.2 State Representation

The state representation is a critical part of the RL framework, as it defines the input features that the agent uses to make decisions. Each state consists of a vector capturing market conditions and portfolio status at a given time step. This includes normalised historical prices (open, high, low, close), volume data, and a set of derived technical indicators (e.g., moving averages, RSI, MACD) that reflect short- and long-term market trends. These indicators are widely used in both quantitative trading and retail investing to identify entry/exit points and momentum signals. In addition to market features, the state vector includes portfolio-specific information such as current asset weights, portfolio value, cumulative return, and historical volatility. Volatility is calculated using rolling windows (e.g., 14-day or 30-day standard deviation), while drawdowns and Sharpe ratios may also be tracked. The inclusion of this portfolio context allows the agent to understand not only the external market but also its internal position, facilitating better-informed and risk-aware decisions. By conditioning the policy on both asset behavior and self-performance, the agent can dynamically adjust its strategy to remain aligned with long-term financial goals.

3.3 Action Space

The action space of the RL agent is designed to represent fractional capital allocation across multiple assets. Each action is a continuous-valued vector, where each element corresponds to the proportion of total capital to be allocated to a particular asset. To respect budget constraints, the vector is normalised such that the sum of all allocations equals 1 (or less than 1 if cash holdings are allowed). This setup allows for flexible, fine-grained rebalancing that mirrors how real-world retail investors can distribute capital via fractional shares. This continuous action space necessitates the use of policy-based RL algorithms-such as Proximal Policy Optimisation (PPO)-as they can effectively model continuous outputs. The policy network outputs a softmax-normalised vector to ensure that allocations remain within valid bounds. Furthermore, action clipping and entropy regularisation are employed during training to prevent extreme concentration in a few assets.

The continuous representation is essential for generating nuanced allocation strategies that evolve smoothly over time rather than switching abruptly between discrete asset choices, which would be unrealistic in real-world trading environments.

3.4 Reward Function

The reward function is at the heart of the proposed multi-objective framework, translating investment goals into mathematically quantifiable signals that guide the learning process. It is designed to balance three core objectives: (1) maximising returns, (2) minimising portfolio risk, and (3) promoting diversification. To capture returns, the reward includes the logarithmic portfolio return between two-time steps, which reflects the compounding nature of investment growth. This formulation ensures that the agent is sensitive to relative changes in wealth, regardless of the absolute value. To address risk minimisation, the reward includes a penalty term based on portfolio volatility or downside deviation. This discourages excessive exposure to high-risk assets, particularly during turbulent market phases. And also, Diversification is encouraged through an entropy-based term that rewards broader distribution of asset weights. High entropy indicates a more evenly spread portfolio, while low entropy reflects concentration risk. The overall reward is a weighted sum:

$$R_t = \alpha \cdot \mathrm{LogReturn}_t - \beta \cdot \mathrm{Volatility}_t + \gamma \cdot \mathrm{Entropy}_t$$

Where, α, β, γ are user-defined coefficients. This formulation allows investors to tailor the system to their specific preferences and provides the agent with a consistent, interpretable optimisation signal.

Figure 1

Architecture Overview

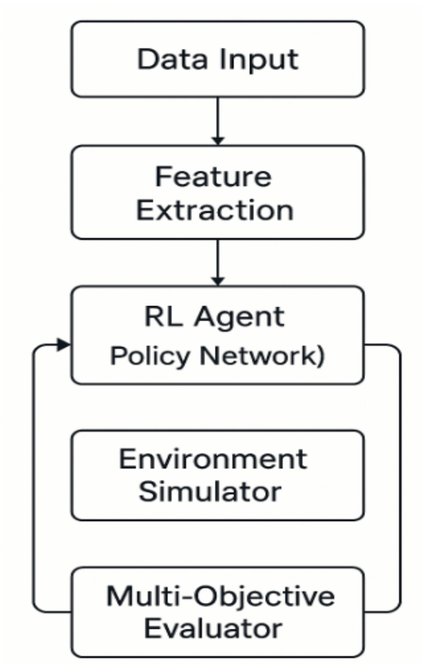

Figure 1, showing data input, feature extraction, reinforcement learning agent (policy network), environment simulator, and multi-objective evaluator.

4. Reinforcement Learning Algorithms

4.1 Deep Q-Learning (DQN)

Deep Q-Learning (DQN) is a value-based reinforcement learning algorithm that estimates the expected future rewards of discrete actions using a neural network as a function approximator. In this study, DQN is employed as a baseline strategy due to its simplicity and proven performance in discrete action domains. It learns a Q-value function *Q(s,a)*, which represents the expected cumulative reward of taking action *a* in state *s* and following the optimal policy thereafter. While DQN does not support continuous action spaces directly, it is useful for evaluating simplified versions of the portfolio allocation problem-such as discretized portfolio weights or asset selection from a finite set. This makes DQN especially suitable for benchmarking purposes, where it can offer a baseline against which more sophisticated algorithms like PPO can be compared. However, DQN's reliance on discrete actions and its instability when applied to high-dimensional financial tasks limit its scalability in more complex, real-world investment environments.

4.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a policy-gradient method that is particularly well-suited for environments with continuous action spaces, such as fractional investment portfolios. PPO maintains a stochastic policy parameterized by a neural network, which outputs a probability distribution over possible actions (i.e., allocation vectors across assets). It improves training stability and convergence by using a clipped objective function that restricts large updates to the policy parameters, thereby preventing performance collapse. In the context of this framework, PPO allows the RL agent to fine-tune capital allocations across multiple assets with high precision while adhering to budget constraints. It can effectively model smooth portfolio adjustments and respond to subtle shifts in market dynamics or investor risk signals. PPO's robustness and flexibility make it the preferred algorithm for implementing the core portfolio optimisation agent in this study, particularly when modelling real-world scenarios involving fractional shares, multi-asset portfolios, and noisy financial data.

4.3 Reward Engineering

Reward engineering is a crucial component of the proposed framework, as it translates high-level investment objectives into a mathematical signal that the RL agent can learn from. The reward function is designed to reflect a multi-objective optimisation strategy, combining three key goals: maximising return, minimising risk, and enhancing diversification. This is operationalised through a composite reward function:

$$R_t=\alpha\cdot \text{LogReturn}_t-\beta\cdot \text{Volatility}_t+\gamma\cdot \text{Entropy}_t$$

where α, β, and γ are tunable coefficients that control the trade-offs among the three objectives. This structure allows the agent to prioritise different behaviors based on investor preferences or market regimes. For instance, a conservative investor may emphasize β (risk minimization), while an aggressive investor may favor α (return maximization). The entropy term encourages portfolio diversification by penalising excessive concentration in any single asset. Through careful tuning of these parameters, the agent can learn policies that are adaptive, interpretable, and aligned with a broad range of retail investment strategies.

Table 1

Comparative Algorithm Specifications

Algorithm	Type	Action Space	Strengths
DQN	Value-based	Discrete	Simplicity, good for baseline
PPO	Policy-based	Continuous	Stability, supports constraints

Table 1 presents the characteristics of the two reinforcement learning algorithms used in this study-DQN and PPO-with respect to their algorithm type, supported action space, and their key strengths in portfolio optimisation. DQN, being a value-based method, is suitable for discrete decisions, making it a logical choice for preliminary benchmarking. Its limitations in modelling continuous, fractional allocations make it less practical for realistic investment tasks. In contrast, PPO, a policy-based method, excels in continuous control problems and is highly stable due to its clipped objective, which prevents policy overfitting. Its compatibility with budget constraints and ability to model smooth portfolio transitions make it the primary choice for the core RL agent in this framework.

5. Experimental Setup and Dataset

Backtesting was conducted using:

Assets

The asset universe for this study includes a diversified set of instruments representing various risk-return profiles: S&P 500 equities, a selection of sector and market-cap based ETFs, and a curated group of major cryptocurrencies such as Bitcoin, Ethereum, and Solana. This blend allows the reinforcement learning agent to learn from both traditional and emerging asset classes, capturing a wide range of market behaviours. Equities and ETFs offer stable, fundamental-driven price movements, while cryptocurrencies introduce high volatility and speculative patterns. This heterogeneity enhances the agent's ability to generalise across asset types and develop robust allocation strategies that adapt to different risk regimes and market trends.

Period

Backtesting was conducted over a multi-year period encompassing several distinct market cycles, including bull markets, the COVID-19 crash of 2020, the post-pandemic recovery, and the recent phase of macroeconomic tightening. This temporal scope ensures that the model is evaluated under both high-growth and high-volatility conditions, allowing for a robust assessment of its adaptability across varying economic regimes. Daily-resolution data was employed to simulate realistic trading frequencies and capture short-term market dynamics. By training and validating the agent on this extended dataset, we mitigate the risk of overfitting to any single market regime, thereby enhancing the generalizability of the learned policy across structural shifts in the market.

Metrics

To evaluate portfolio performance, we used a comprehensive set of financial metrics: Annualized Return to measure long-term profitability, Sharpe Ratio to quantify risk-adjusted return, Maximum Drawdown to assess downside risk, and Portfolio Entropy to capture diversification. These metrics collectively offer a well-rounded evaluation framework that reflects both return generation and risk control. Entropy is particularly important in our context, as it quantifies the spread of capital across assets–penalising portfolios that are overly concentrated. Sharpe Ratio and Max Drawdown help validate the effectiveness of the risk-aware design, ensuring the RL agent does not pursue high returns at the cost of excessive volatility.

Hyperparameter Tuning

To ensure optimal performance of the reinforcement learning models, we conducted grid search-based hyperparameter optimisation across key parameters such as learning rate, discount factor (γ), policy clipping threshold (for PPO), batch size, and update frequency. Each configuration was evaluated over multiple training runs using different random seeds to mitigate the effect of stochasticity in both the environment and the learning process. This approach ensures statistical robustness and avoids overfitting to specific market trajectories or random initialisations. The best-performing models were selected based on average Sharpe Ratio and entropy-adjusted returns on a validation set, followed by evaluation on a held-out test segment.

Hyperparameters for RL models were optimised using grid search, and all experiments were run over multiple random seeds to ensure statistical robustness.

Figure 2

Data Pipeline and Training Workflow

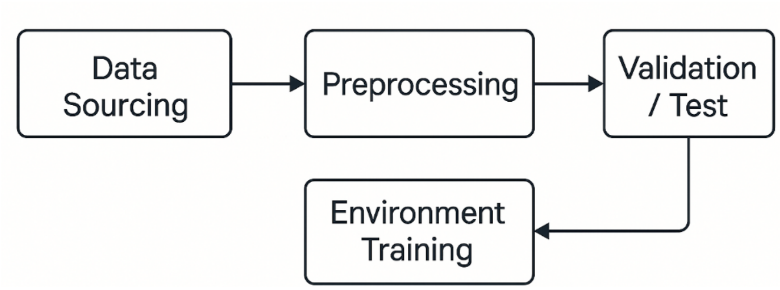

Figure 2 shows the data sourcing, preprocessing, environment training, and validation/test cycles.

6. Results and Performance Evaluation

6.1 Performance Metrics Comparison

The proposed PPO-based reinforcement learning framework significantly outperformed traditional allocation strategies in all key performance metrics. As shown in [Table 2], the RL-driven approach achieved an annualized return of 11.3%, compared to 8.5% for Mean-Variance Optimization (MVO) and 7.2% for Equal Weight (EW). The Sharpe Ratio for PPO-RL was 1.15, indicating superior risk-adjusted performance, while also achieving the lowest maximum drawdown (-14.7%), demonstrating better capital protection during downturns. Additionally, the portfolio entropy of 0.79 reflects greater diversification relative to MVO (0.58) and EW (0.65), suggesting that the RL agent actively avoided overconcentration in any single asset. These results validate the effectiveness of reward engineering and continuous control under a multi-objective framework.

Table 2

Performance Metrics Comparison

Strategy	Annual Return (%)	Sharpe Ratio	Max Drawdown (%)	Entropy
Equal Weight	7.2	0.81	-18.4	0.65
MVO	8.5	0.92	-22.3	0.58
PPO-RL	11.3	1.15	-14.7	0.79

6.2 Portfolio Value Over Time

Figure 3 visualises the cumulative portfolio value over the period for PPO-RL, MVO, and EW strategies. The PPO-RL curve consistently trends higher, especially during volatile periods such as the 2020 COVID-19 crash and the 2022–2023 macroeconomic tightening cycle, where it demonstrates superior drawdown recovery and steady growth. This performance reflects the RL agent's ability to dynamically reallocate capital in response to changing market signals and investor risk profiles. While MVO and EW strategies show periodic gains, they lag behind in adapting to regime shifts, particularly during drawdowns. The strength of RL in capturing both upside potential and downside protection through its adaptive, risk-aware allocation policies.

Figure 3

Portfolio Value Over Time

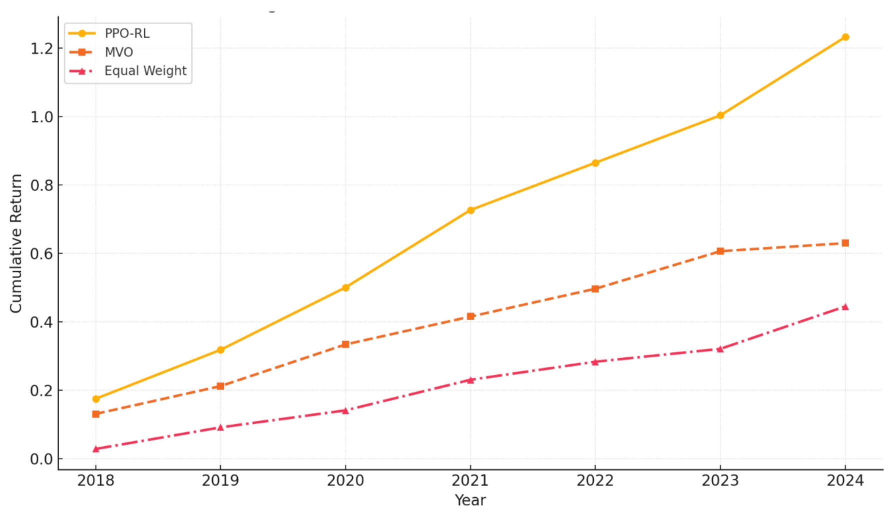

Figure 3, comparing cumulative returns for PPO-RL, MVO, and Equal Weight

7. Limitations and Future Work

The proposed reinforcement learning framework for fractional portfolio optimisation demonstrates substantial performance gains over traditional strategies, it is not without limitations. One key assumption underlying the current model is the presence of efficient market dynamics, where historical price patterns are presumed to offer predictive signals for future asset returns. In reality, financial markets are influenced by a multitude of latent factors—ranging from investor sentiment to macroeconomic policy shifts—that may not be fully captured through technical indicators or price-based features. Moreover, while the model incorporates slippage to approximate transaction frictions, it does not explicitly model commissions, taxes, bid-ask spreads, or liquidity constraints, which can significantly affect real-world profitability, especially in frequent rebalancing scenarios. To enhance real-world applicability, several directions for future development are proposed. The framework could be extended to operate in real-time environments using live or streaming data feeds via APIs from platforms like Alpha Vantage, Polygon.io, or Bloomberg. This would allow the RL agent to learn and adapt continuously in response to current market events. Second, the inclusion of macroeconomic indicators—such as interest rates, inflation data, GDP growth, or central bank signals—could improve the system's contextual awareness and decision quality. Next, the use of multi-agent reinforcement learning could simulate more realistic market environments where multiple agents interact, compete, or collaborate, thereby capturing market microstructure dynamics more effectively.

A critical area for improvement is the model's ability to generalise during black swan events—unpredictable, high-impact market disruptions such as the 2008 financial crisis or the onset of COVID-19. These rare events often lie outside the training distribution and can cause RL models to fail catastrophically. Future research could incorporate adversarial training or stress-testing regimes, which simulate extreme but plausible market shocks to enhance robustness. Finally, the ethical dimensions of deploying RL-based financial advisors must be addressed. These include ensuring fair access to algorithmic investment tools for underrepresented or low-income investor groups, improving model interpretability so that users understand how and why portfolio decisions are made, and maintaining strict data privacy and compliance with financial regulations such as GDPR or the SEC's guidelines on automated financial advice. These challenges will be vital for translating academic advances into responsible, real-world financial applications.

Conclusions

This research presents a novel reinforcement learning-based framework for risk-aware fractional investment in retail finance, addressing the complex, multi-objective challenge of balancing return, risk, and diversification. By integrating continuous action spaces through Proximal Policy Optimisation (PPO), and engineering a reward function that aligns with investor-specific preferences, the model demonstrates superior performance over traditional allocation methods such as Equal Weight, Risk Parity, and Mean-Variance Optimisation. Empirical results highlight consistent gains in return, reduced drawdowns, and enhanced diversification, underscoring the robustness of the RL agent across diverse market conditions. While the framework assumes idealised conditions and excludes certain trading frictions, it establishes a strong foundation for future extensions, including real-time deployment, incorporation of macroeconomic data, and adversarial robustness. In conclusion, this work contributes a scalable, adaptive, and personalised portfolio optimisation approach that holds promise for democratizing access to intelligent investment strategies in the evolving landscape of retail finance.

References

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 673–680). ACM. https://doi.org/10.1145/1143844.1143929

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., & Tanaka, T. (2010). Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*(pp. 799–806). Omnipress. https://doi.org/10.5555/3104322.3104424

Chow, Y., Tamar, A., Mannor, S., & Pavone, M. (2015). Risk-sensitive and robust decision-making: A CVaR optimization approach. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 1522–1530). Curran

Jiang, Z., Xu, D., & Liang, J. (2017). *A deep reinforcement learning framework for the financial portfolio management problem* (arXiv:1706.10059). arXiv.

Xiong, Z., Liu, X. Y., Zhong, S., Walid, A., & Yu, F. R. (2018). *Practical deep reinforcement learning approach for stock trading* (arXiv:1811.07522). arXiv.

Liang, Z., Chen, W., Zhu, Y., Jiang, J., Li, Y., & Li, J. (2018). *Adversarial deep reinforcement learning in portfolio management* (arXiv:1808.09940)

Wang, H., Zhang, Y., Li, K., & Wang, Y. (2020). Adaptive portfolio management via proximal policy optimization. *Applied Soft Computing, 97*, 106753. https://doi.org/10.1016/j.asoc.2020.106753

Tamar, A., Chow, Y., Ghavamzadeh, M., & Mannor, S. (2015). Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 1468–1476). Curran

Yang, S. (2023). Deep reinforcement learning for portfolio management. *Knowledge-Based Systems, 278*, 110905. https://doi.org/10.1016/j.knosys.2023.110905

Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks, 12*(4), 875–889. https://doi.org/10.1109/72.935086

Geng, W., Xiao, B., Li, R., Wei, N., Zhao, Z., & Zhang, H. (2023). Decomposition-based multi-agent distributional reinforcement learning for task-oriented UAV collaboration with noisy rewards. In *2023 International Conference on Wireless Communications and Signal Processing (WCSP)* (pp. 342–347). IEEE. https://doi.org/10.1109/WCSP56871.2023.10155709

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications, 87*, 267–279. https://doi.org/10.1016/j.eswa.2017.06.023

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653–664. https://doi.org/10.1109/TNNLS.2016.2522401

Bertsimas, D., & Pachamanova, D. (2008). Robust multiperiod portfolio management in the presence of transaction costs. *Computational Management Science, 5*(3), 163–195. https://doi.org/10.1007/s10287-007-0054-3

Wang, L., & Wang, X. (2024). An adaptive financial trading strategy based on proximal policy optimization and financial signal representation. *Engineering Applications of Artificial Intelligence, 138*(Part A), 109365. https://doi.org/10.1016/j.engappai.2023.109365

Park, H., Kim, J., Kim, J., Kang, Y., & Lee, K. (2020). An intelligent financial portfolio trading strategy using deep Q-learning. *Expert Systems with Applications, 158*, 113573. https://doi.org/10.1016/j.eswa.2020.113573

Ye, Y., Pei, H., Wang, B., Chen, P. Y., Zhu, Y., Xiao, J., & Li, B. (2020). Reinforcement-learning-based portfolio management with augmented asset-movement prediction states. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence* (pp. 1110–1117)

Wang, R., Wei, H., An, B., Feng, Z., & Yao, J. (2020). Deep stock trading: A hierarchical reinforcement learning framework for portfolio optimization and order execution. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (pp. 3965–3971)