Type of the Paper (Literature Review)

Evolution and Challenges of Reinforcement Learning in Intraday Algorithmic Trading: A Systematic Literature Review

Thomas Amorgianiotis[1], Grigorios Beligiannis[2*] and Efstratios F. Georgopoulos[3]

[1] PhD student, Dept. of Business and Organization Administration, University of Peloponnese, Greece, th.amorgianiotis@go.uop.gr
[2] Professor, Dept. of Food Science & Technology, University of Patras, Greece, gbeligia@upatras.gr
[3] Professor, Dept. of Business and Organization Administration, University of Peloponnese, Greece, e.georgopoulos@uop.gr
* Correspondence: gbeligia@upatras.gr ; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

Highlights

What are the main findings?

- *Systematic identification of 16 high-quality deep RL studies for intraday trading.*
- *Evolution from single-agent DQN to complex multi-agent and human-aligned architectures.*
- *Hybrid feature engineering and frequency decomposition effectively mitigate market volatility.*

What are the implications of the main findings?

- *Provides a roadmap for building risk-sensitive and autonomous intraday trading systems.*
- *Highlights the necessity of multimodal inputs and realistic transaction cost modeling.*

Abstract

Reinforcement learning is a type of machine learning method [1], which uses an agent for decision making. In each step, the agent makes a prediction for the future. Each move receives a reward or a punishment based on the future state of the problem [2]. Reinforcement learning is widely applied to decision-making problems, including intraday stock prediction, which is considered a highly challenging task due to market volatility and rapidly changing conditions. RL approaches can be used to construct stock portfolios and generate daily trading decisions, such as whether to buy, sell, or hold assets. The reward function is particularly suitable for this task, as excessive trading activity may reduce profitability because of transaction costs.

Searches were conducted in the ScienceDirect and Scopus databases for studies published between 2018 and 2025. Following a rigorous screening process of 276 records, 16 relevant studies were selected for inclusion in this review. The reviewed studies collectively suggest that, although RL agents can effectively capture intraday market dynamics, their success depends heavily on sophisticated feature engineering and the realistic modeling of transaction costs. This study follows the PRISMA 2020 guidelines.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: reinforcement learning; deep reinforcement learning; intraday trading;

algorithmic trading; systematic literature review; financial markets; non-stationarity; multi-agent systems; deep Q-learning; financial time series.

1. Introduction

Reinforcement learning is a type of machine learning method and it is used for various problems. We want to explore its use in intraday stock returns. It is a field that other researchers have also tried to exploit and there is significant advantage these last years. It is to be expected that intraday stock prediction is a very challenging task. Data is continuously changed and having too many transactions may eliminate your profits. Therefore, reinforcement learning is often applied in the tasks of constructing portfolios and stock prediction. One useful example of reinforcement learning are Q-learning algorithms. Q-learning was introduced in 1989 by Watkins. A new updated version is the Double Q-learning [3]. Successively, we were presented with Deep Reinforcement Learning with Double Q-learning [3] and the Clipped Double Q-learning [4]. Please note that besides these, there are also various other reinforcement learning methods. In this review we want to explore the different algorithms of reinforcement learning that are used for intraday stock prediction. As far as we investigated, we conclude that, despite the progress, there is as yet no extensive literature review about this topic.

This review considers reinforcement learning applications not only in intraday stock prediction, but also in broader algorithmic trading tasks such as portfolio optimization, trade execution, volatility forecasting, and adaptive financial decision-making systems, reflecting the increasingly overlapping nature of modern financial AI research.

This review follows the PRISMA 2020 guidelines for reporting systematic literature reviews.

2. Materials and Methods

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

2.1. Research Questions

Methology

Regarding the literature review, we used guidance from [5] and a software tool [6].

We created our review based on the following research questions:

Main questions:

1. What are the key algorithms and techniques of reinforcement learning (RL) that can be applied in intraday trading, and how have they evolved over time?
   Our aim is to identify the foundational models of intraday RL and compare them with more recent advancements or modifications.
2. How do various reinforcement learning models address the challenges of non-stationarity and high-frequency changing data that are observed in intraday trading?
   This delves into how RL algorithms handle constantly changing market conditions and the large volume of data generated within a single trading day.
3. What are the common performance metrics and evaluation methods used to assess RL-based intraday trading systems?

It's important to explore how success is measured in this context and what benchmarks or metrics are standard in evaluating RL models.

4. What role do feature selection and data preprocessing play in the performance of reinforcement learning models in intraday trading?
   Understanding the importance of data input can provide insight into optimizing RL models in the context of intraday decision-making.
5. What are the limitations and risks associated with applying reinforcement learning to intraday trading and what solutions have been proposed in the literature?
   This question addresses the practical constraints, including issues like overfitting, market impact, and the need for real-time decision-making.

2.2. Search Strategy

We also used the electronic libraries Science Direct and Scopus for the years 2018 - 2025 with the following search criteria:

Science Direct: intraday and stock prediction and reinforcement learning

Scopus: (ALL ("intraday") AND ALL ("stock prediction") AND ALL ("reinforcement learning"))
Keywords: Intraday, Reinforcement learning, Stock Prediction

2.3. Inclusion/Exclusion Criteria

Our inclusion(I), and exclusion(E) criteria were:

- (I) Is the study a peer-reviewed publication?
- (I) Is the study relevant to the keywords above?
- (E) Does the study not focus on reinforcement learning?
- (E) Does the study not focus on stock prediction?
- (E) Does the study not include intraday data?
- (E) Is the study not a literature review?
- (E) Is the study lecture notes or a book chapter?
- (E) Is the study not in English?

This search returned 163 papers from Scopus and 113 papers from Science Direct, that is a total of 276 papers. We followed a selection phase, where we found 20 duplicated papers. We also rejected 235 papers. That left us with 21 accepted papers, 6 from Scopus and 15 from Science Direct. Lastly, there was an extraction phase where we accepted 16 out of 21 papers, 4 from Scopus and 12 from Science Direct.

The study selection process was conducted by three reviewers. Titles and abstracts were initially screened to remove irrelevant studies, followed by a full-text assessment based on predefined inclusion and exclusion criteria. Any uncertainties during the selection

process were resolved through careful re-evaluation of the study against the eligibility criteria.

2.4. Protocol Registration Statement

This systematic literature review was conducted in accordance with the methodological guidance provided by [5]. The selection and extraction processes were managed using the StArt [6] software tool, while the final PRISMA 2020 flow diagram was generated using an interactive R package and Shiny app [7] to optimize digital transparency and support Open Synthesis. While a formal protocol was not prospectively registered in a public registry, the review strictly adhered to a predefined set of five research questions and specific inclusion/exclusion criteria to ensure rigor and minimize bias. The search was limited to peer-reviewed articles from the Science Direct and Scopus databases published between 2018 and 2025.

2.5 Risk of Bias Assessment

The risk of bias in the included studies was assessed qualitatively based on study design, dataset selection, and validation methodology. Particular attention was given to issues such as overfitting, lack of out-of-sample validation, reliance on historical backtesting, and absence of real-world deployment. Studies that did not incorporate transaction costs or realistic execution constraints were considered to have a higher risk of bias.

3. Results

3.1 Study Selection (PRISMA)

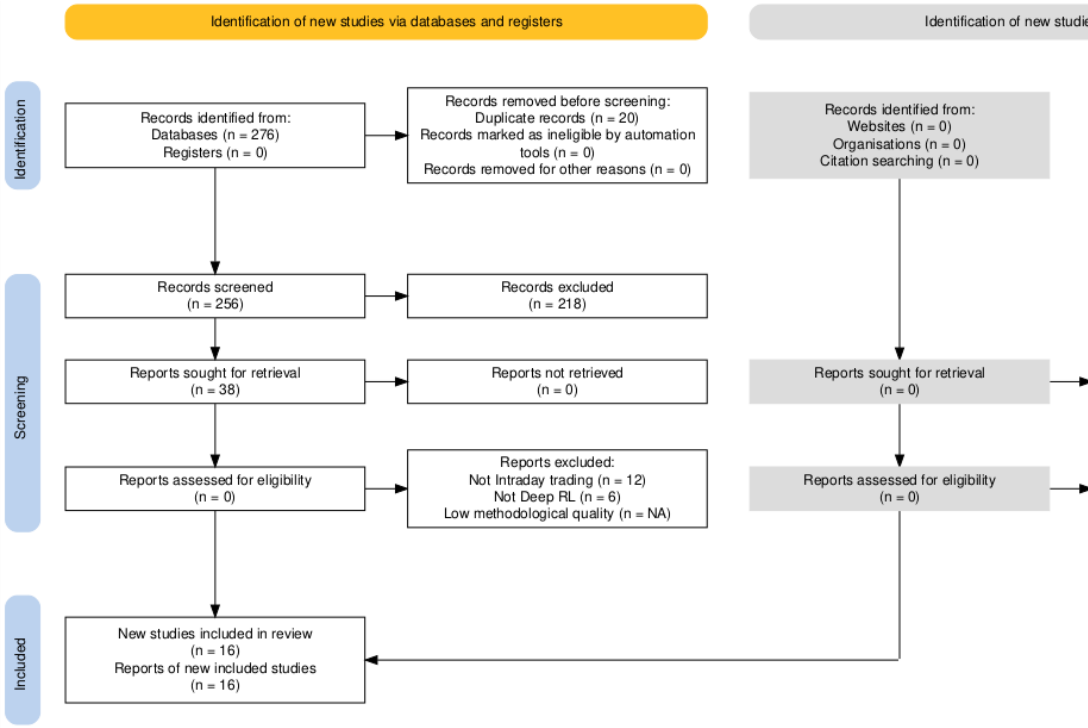

Figure 1. Flow diagram

3.2 Synthesis of Evidence

Comprehensive Synthesis Table of Reinforcement Learning and Hybrid Models in Algorithmic Trading

Table 1. Summary of Reinforcement Learning Models and Datasets in Algorithmic Trading

Paper / Author(s)	Algorithm / Model	Objective / Problem Addressed	Dataset / Market
Ye & Schull (2023) [8]	Human-Aligned Trading (HAT)	Introduces a human-aligned reinforcement learning trading model that integrates supervised learning, single-step and multi-step Q-learning, and imitation learning to mimic human trading behavior.	Five blue-chip & five small-cap US stocks; S&P500 trends (upward/sideways/downward)
De Azevedo Takara et al., (2024) [9]	Extended Trading DQN (ETDQN)	Introduces a DQN variant (ETDQN) that receives feedback only upon trade liquidation, effectively addressing sparse-reward challenges in intraday trading.	Intraday stock data, Western Digital Corporation and the Cosmos cryptocurrency,
Yu et al., (2023) [10]	GVMD-Q-DBN-LSTM-GRU	The GVMD-Q-DBN-LSTM-GRU model enhances volatility forecasting by combining optimized VMD decomposition, deep neural networks, and Q-learning–based adaptive integration, achieving the most accurate and robust multi-horizon predictions across global stock indices.	Three stock price indices SSEC, SPX and FTSE.
Y. Lin et al., (2022) [11]	Multiagent-Based Deep RL (MABDRL)	Develops a multiagent deep reinforcement learning (MABDRL) framework for risk-shifting portfolio management, using nested local and global agents to balance asset-level risk and maximize overall portfolio returns.	Real market hedge fund data

Shavandi & Khedmati, (2022) [12]	Multi-Agent DQN Framework	Proposes cooperative multiagent DQN system for trading across timeframes (short, mid, long).	Nine years of EUR/USD historical data
Massahi & Mahootchi, (2024) [13]	Multi-agent DQN & DDQN for Commodity Futures	A Deep Q-learning system for futures trading that leverages limit order book and trade features, GRU-based multi-agent networks and replay memory to generate robust and profitable intraday decisions.	Commodity futures markets (gold coin futures contracts traded in the Iran mercantile exchange from 2013 until 2018.)
Kim et al., (2023) [14]	Dual-Level PPO(Proximal Policy Optimization) (Transformer + LSTM)	Uses dual-level RL for VWAP execution with PPO: Transformer for global volume shape, LSTM for local order distribution.	Intraday stock volume data. Korea Exchange (KRX). Samsung Electronics (SE), SK Hynix (SH), Kia Corporation (KC) and POSCO Holdings Inc (PH) .
Dong & Zheng, (2024) [15]	Soft Imitation Reinforcement Learning (SIRL)	Proposes imitation-enhanced RL with value decomposition for portfolio management.	Financial portfolio data. Chinese and US stock market.
Chen & Huang, (2021) [16]	Multimodal Deep RL with Sentiment Influence Model	Integrates sentiment analysis and multimodal learning in RL to include price + news impact for trading decisions.	Public stock and news datasets across sectors
Carta et al., (2021) [17]	Multi-DQN Ensemble (An ensemble of Deep Q-learning agents)	Uses ensemble of DQN agents trained independently to improve robustness and avoid overfitting in stock forecasting.	Intraday stock market data
Sun et al., (2022) [18]	Deep-Scalper	Risk-aware RL for intraday trading with dueling Q-network, action branching, hindsight bonuses, and auxiliary risk tasks.	Three years of futures data (2 stock index, 4 bond futures)
Zengeler & Handmann, (2020) [19]	LSTM-based RL for contracts for difference (CFD) Trading	Uses RL agent with LSTM to environment of partially observable Markov decision process (POMDP simulation).	Contracts for Difference (CFD) market data
Mengshetti et al., (2024) [20]	Weapon Candle Strategy	Combines four key indicators into a unified	NSE (India)

	(RSI+EMA+VWAP+MACD)	trading rule for improved accuracy and profitability.	
Kumar et al., (2025) [21]	Q-VMD-ANN-LSTM-GRU (Hybrid)	Forecasts realized volatility using decomposition and ensemble DL methods.	SSE, INFY, NIFTY (2015–2022)
Belyakov & Sizykh, (2024) [22]	Proximal Policy Optimization (PPO)-based RL for Hedge Fund Strategy Selection	Employs PPO to dynamically allocate capital between mean-reversion and momentum strategies.	Historical hedge fund market data, data of five US stocks.
Jeon et al., (2024) [23]	Adaptive Portfolio Optimization via Multi-Frequency Quantitative Analysis (FreQuant) (DFT + RL)	Portfolio optimization via frequency-domain analysis with RL, capturing both dominant and subtle patterns.	Multiple real-world trading datasets (U.S. stock market, KR stock market, Crypto market, CN stock market, JP stock market, U.K. stock market)

Table 2. Performance, Results, and Limitations of Reinforcement Learning Models in Algorithmic Trading

Key Results	Limitations / Notes
HAT outperformed baseline models, particularly in down-trending markets and for small-cap stocks due to lower trading frequency	The model's performance is limited by relying only on recent time-series data, a common issue in RL due to the partially observable nature of financial markets.
Improved trading consistency across varying market conditions vs. standard DQN.	Limited to simulated intraday data; lacks transaction cost modeling.
Outperforms comparative models in realized volatility prediction accuracy.	Model complexity; interpretability challenges; regime dependency.
Improved risk-adjusted returns and portfolio control over SOTA baselines.	Requires high coordination overhead; computationally expensive.
Outperforms independent agents and benchmarks across all risk and return metrics.	Complex training synchronization; scalability to more agents not tested.

Superior performance due to action-specific agents and double Q-learning.	Overfitting risks; performance dependent on feature quality.
Accurate VWAP tracking, outperforming prior RL execution models.	Requires large datasets; generalization to other assets not evaluated.
Improved stability and faster convergence over pure RL baselines.	Imitation data dependency; limited test cases reported.
Multimodal agents outperform price-only models by ≥13.26%.	Sensitivity to news quality; additional model complexity.
Outperforms Buy-and-Hold;	Increased computational cost; ensemble averaging reduces interpretability.
Significantly outperforms Buy & Hold (BAH), Mean Reversion (MV), Time SeriesMomentum (TSM), MLP, GRU, (LGBM) gradient boosting decision tree with gradient-based one-side sampling and exclusive feature bundling., DQN, DS-NH is a variant of DeepScalper (DS), which removes the hindsight bonus from the reward function. • DS-NA is a variant of DeepScalper (DS), which removes the risk-aware auxiliary task.across four financial metrics (Total Return (TR), Sharpe Ratio (SR) Calmar Ratio (CR), Sortino Ratio (SoR)).	High training cost; complex reward shaping; interpretability issues.
LSTM-RL agent learns policy outperforming market; scalable with larger model sizes.	Dependent on simulated data; real-world latency not fully modeled.
60.63% profitable trades; profit factor 1.882.	Parameter overfitting; ignores slippage and transaction costs.

Best performance among compared models for all indices. Higher cumulative and risk-adjusted returns vs. traditional methods. Up to 2.1× Annualized Return and 2.9× Portfolio Value vs competitors.	Complex model; generalization uncertain under regime shifts. High compute/time cost; reward sensitivity; scalability constraints. Sensitive to DFT windowing; potential loss of time-local information.

3.3 Analysis of Research Questions

3.3.1. What are the key algorithms and techniques of reinforcement learning (RL) that can be applied in intraday trading, and how have they evolved over time?

Our aim is to identify the foundational models of intraday RL and compare them with more recent advancements or modifications.

As can be seen at the previous table, most algorithms are using the Q-Learning method. These algorithms use a single or multiple levels of Q-Networks combined with other algorithms such as LSTM [19]. Some algorithms also use sentiment analysis as extra input.

In the early years, the algorithms used simple Q-Networks that gradually evolved into multiple Q- Networks that are combined with other algorithms such as LSTM. Moreover, various agent training algorithms were applied. In addition to that, the input of the algorithms was exploited using sentiment analysis and computed input data.

In 2021, Chen & Huang [16] introduced a combination of reinforcement learning, sentiment analysis, and multimodal learning. We also observe an evolution from single-agent DQN/LSTM to Double/dueling/ensembles [17].

The next year, a multiagent deep reinforcement learning framework is used for risk-shifting portfolio management [11] . Again, we find the use of a multiagent DQN system for trading across timeframes (short, mid, long) [12]. Moreover, there are various machine learning combinations like optimized Variational Mode Decomposition (VMD) with deep learning methods like Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Effort is also made to combine human knowledge (HAT)[8] and sentiment analysis. We also see the GVMD-Q-DBN-LSTM-GRU [10] model, which enhances volatility forecasting by combining optimized VMD decomposition, deep neural networks, and Q-learning–based adaptive integration.

A different approach was portfolio optimization via frequency-domain analysis with RL, capturing both dominant and subtle patterns [23]. There was a combination of four key indicators (RSI+EMA+VWAP+MACD) into a unified trading rule for improved accuracy and profitability [20] .We also found Proximal Policy Optimization (PPO)-based RL for Hedge Fund Strategy Selection [22].

Massahi & Mahootchi [13] used a Deep Q-network (DQN) algorithm and its robust double-version (DDQN). They proposed imitation-enhanced RL with adaptive weighting and value decomposition for portfolio management.

Yet another model is ETDQN which stands out by its ability to adapt its learning process to trade effectively across varying market conditions, with feedback [9]. We also observed a combination of Q-VMD-ANN-LSTM-GRUmethods [21] .

3.3.2. How do various reinforcement learning models address the challenges of non-stationarity and high-frequency changing data that are observed in intraday trading?

This delves into how RL algorithms handle constantly changing market conditions and the large volume of data generated within a single trading day.

Intraday trading involves high frequency data. The handling of this data is usually achieved with the use of multi layered Q- Networks [17] or with a combination of a lot of algorithms. Likewise, multiagent deep reinforcement learning frameworks are used for risk-shifting portfolio management [11].

To handle non-stationarity and high-frequency noise in intraday trading, algorithms preprocess stock data using Discrete Wavelet Transformation (DWT), which decomposes the time series into localized frequency sub-bands. High-frequency noise is filtered out before feeding the denoised signal into our RL model, improving learning stability and performance [8].

Multiagent systems on the other hand address non-stationarity and high-frequency noise by using a hierarchical multi-agent structure where higher-timeframe agents guide lower-timeframe ones, enabling stable, adaptive trading across dynamic market conditions. [11 ,12].

In the combination of algorithms paradigm, the GVMD-Q-DBN-LSTM-GRU model sufficiently combines outstanding decomposition methods, reinforcement learning integration techniques and deep neural networks. [9]. This model overcomes non-stationarity and high-frequency noise in intraday trading by decomposing volatility signals into stable components via optimized VMD, modeling nonlinear temporal patterns with deep networks and adaptively integrating predictions through Q-learning for robust, regime-aware forecasting [10].

ETDQN mitigates non-stationarity and high-frequency noise in intraday trading by using event-based rewards, distributional learning, and prioritized experience replay to adaptively learn stable trading policies across shifting market regimes.

Massahi & Mahootchi [13] handle non-stationarity and high-frequency intraday data by using GRU-based sequence learning, replay memory, and stabilized double Q-learning within an action-specific multi-agent architecture that efficiently captures evolving market microstructure patterns.

In addition to data input, and its volatile nature, the researchers made efforts in the way of sentiment analysis of news. The multimodal RL model addresses non-stationarity by using recurrent networks and time-aware news-influence modeling, which help the agent adapt to shifting market regimes and understand how the impact of news decays over time. By combining price data with sentiment, it reduces the noise inherent in high-frequency intraday fluctuations and provides more stable signals, which are useful for decision-making. This multimodal fusion enables the agent to remain robust even when market patterns change rapidly within the trading day. [16].

3.3.3. What are the common performance metrics and evaluation methods used to assess RL-based intraday trading systems?

It's important to explore how success is measured in this context and what benchmarks or metrics are standard in evaluating RL models.

The literature shows that RL-based intraday trading systems are typically evaluated using a combination of return-based, risk-based and statistical prediction-based metrics. Return-oriented measures—such as cumulative return, annualized return, average return, annual profit, total return, and equity curve analysis—are widely used to quantify profitability [11,17,18]. To assess robustness and account for downside risk, studies commonly report maximum drawdown, Sharpe ratio, Sortino ratio, Calmar ratio, and related risk-adjusted return metrics, including CVaR and CDaR [20,22,23]. Several works also employ predictive accuracy measures—such as MAE, MSE, RMSE, MAPE, HMAE, HMSE—and statistical comparison tests like the modified Diebold–Mariano test to evaluate forecasting and model stability [21].

Evaluation commonly includes comparison against standard baselines, including Buy-and-Hold, Mean Reversion, Time-Series Momentum, supervised learning models (MLP, GRU, LGBM) and various RL agents such as DQN, Double DQN, PDDQN, Multi-Step DQN, DQfD, and DeepScalper variants [8, 17]. Together, these metrics and baselines form a comprehensive framework for assessing profitability, risk exposure, and learning effectiveness in RL-based intraday trading systems.

3.3.4. What role do feature selection and data preprocessing play in the performance of reinforcement learning models in intraday trading?

Understanding the importance of data input can provide insight into optimizing RL models in the context of intraday decision-making.

- Critical inputs: Limit Order Book (LOB) features and recent trade history are highly informative for intraday/futures trading and are frequently used (as explicitly noted in the commodity futures study). In technical indicators (like RSI, EMA, MACD, VWAP) ([20]), normalized price/volume features and LOB levels are also common.
    - Decomposition & frequency transforms: VMD/GVMD decomposition or DFT (FreQuant) [23] isolates meaningful frequency components, improving model adaptability and capturing abrupt events not visible in time-domain only models.
    - Multimodal fusion & sentiment features: Adding sentiment/news embeddings gives context for exogenous events; influence models that weight sentiment over time improve responsiveness.
    - Temporal preprocessing: aggregating intraday data into suitable windows, using U-shaped volume models (for VWAP), and ensuring consistent sampling/normalization reduces noise and mimics real execution constraints.

- Normalization & handling missing data: standard scaling, z-score, winsorization and careful handling of gaps are necessary for stable RL training on noisy intraday feeds (implicit across LOB and intraday studies.)

We observe that preprocessing and feature selection are as important as the RL algorithm itself— better features (LOB, decomposed signals, sentiment) make the learning problem easier and more robust. As such, feature selection and preprocessing improve RL performance by providing clean, informative inputs. Normalized and denoised time series (via DWT) combined with technical indicators like RSI, Bollinger Bands and moving averages help the model learn effective intraday trading strategies [8].

3.3.5. What are the limitations and risks associated with applying reinforcement learning to intraday trading and what solutions have been proposed in the literature?

This question addresses the practical constraints, including issues like overfitting, market impact, and the need for real-time decision-making.

- Overfitting to historical microstructure & data-snooping.

Mitigations: ensembles (Multi-DQN) ([17]), cross-regime validation, realistic backtesting (HAT) and hold-out periods; incorporate transaction costs and slippage.

- Non-stationarity & regime shifts.

Mitigations: multi-agent/hierarchical architectures, decomposition/frequency methods (FreQuant) [23]) and online/adaptive learning schemes.

- Sparse or delayed rewards (credit assignment for trades).

Mitigations: update strategies tuned to trade liquidation (ETDQN), reward shaping (hindsight bonuses) and imitation learning to provide denser supervision.

- High action dimensionality & large action spaces.

Mitigations: action-branching, value decomposition or decomposing actions across agents to reduce dimensionality and improve learning stability.

- Risk exposure, regulatory & ethical concerns.

Mitigations: risk-aware auxiliary objectives, explicit risk constraints in reward and human-alignment [8] components to avoid extreme or unsafe strategies ([18]). DeepScalper.

- Latency and implementation / real-world execution constraints.

Mitigations: design with realistic latency assumptions in backtests; some studies show that larger model capacity can offset latency but practical deployment must test on real infra.

4. Discussion

4.1 Evolution of Reinforcement Learning Techniques in Algorithmic Trading

The reviewed studies collectively demonstrate a clear evolution in reinforcement learning (RL) methodologies applied to financial trading. Early models, such as the LSTM-based RL approach for Contracts for Difference by [19], established the feasibility of applying RL to POMDP. Subsequent works like Extended Trading DQN (ETDQN) by Takara et al. and the Multi-DQN ensemble by Carta et al. [17] introduced improvements in training stability through target network scheduling, double-Q learning, and ensemble averaging.

More advanced research expanded RL's capacity through multi-agent systems [11,12], enabling distributed decision-making across timeframes or portfolio layers. Hierarchical and dual-level designs [14] leveraged architectures such as Transformers and LSTMs for task specialization, improving execution accuracy and robustness. Another trend emphasizes risk awareness and human alignment—as seen in DeepScalper [18] and HAT. DeepScalper tries to mimic the workflow of professional intraday traders, where RL agents integrate imitation learning, multi-loss objectives, or auxiliary risk functions to emulate prudent human trading behavior and mitigate unsafe exploration.

In parallel, hybrid and frequency-domain approaches [10,23] illustrate how deep RL can be enhanced by signal decomposition and spectral feature extraction, offering resilience to market non-stationarity. Collectively, this progression shows that the RL field in trading has evolved from single-agent, static architectures to complex, multi-modal and risk-sensitive frameworks.

4.2 Addressing Market Non-Stationarity and Intraday Complexity

Handling non-stationary and high-frequency financial data remains one of the most persistent challenges in algorithmic trading. The surveyed works employed several complementary strategies:

Architectural adaptation: Multi-agent frameworks (MABDRL [11], Multi-Agent DQN) distribute decision-making across specialized agents, improving adaptability to regime changes.

Temporal and frequency decomposition: Studies like GVMD-Q-DBN-LSTM-GRU and FreQuant demonstrated that decomposing price signals into multiple frequency bands allows models to isolate stable components, enhancing predictive robustness.

Learning process adjustments: ETDQN's [9] trade-level reward updating and DeepScalper's [18] hindsight bonus represent methods of stabilizing RL training in sparse or delayed reward environments. The hindsight bonus is introduced to stabilize and improve RL training by encouraging agents to consider both short-term gains and long-term profit trends, mitigating the tendency to overfit to short-term price fluctuations in intraday trading.

Human-aligned and imitation learning: HAT [8] and SIRL incorporate supervised imitation and joint-loss objectives, grounding agent behavior in realistic human trading patterns and improving stability under changing market conditions.

These findings indicate that combining architectural diversity, reward shaping, and hybrid feature engineering is the most effective strategy for overcoming intraday data volatility and structural breaks.

4.3 Evaluation Metrics and Model Assessment Practices

Across all reviewed studies, performance evaluation consistently included return-based metrics (cumulative return, profit factor, percentage of profitable trades) and risk-adjusted measures (Sharpe ratio, maximum drawdown).

Execution-focused research, such as Kim et al.'s dual-level PPO framework, introduced VWAP deviation and execution cost as specialized metrics, while DeepScalper [18] applied four financial performance criteria for robustness, Total Return (TR), Sharpe Ratio (SR) Calmar Ratio (CR), Sortino Ratio (SoR).

Notably, HAT emphasized realistic backtesting with transaction costs and multi-trend windows [8], addressing a common limitation in earlier literature that relied solely on idealized simulations.

This consistency in performance indicators suggests a maturing methodological standard, though cross-market validation and live testing remain underexplored areas.

4.4 Feature Engineering, Inputs, and Market Representation

A recurring insight is that the quality and diversity of features substantially influence RL model performance. Indicator-based systems [20] demonstrated that fusing classical signals like RSI, EMA, VWAP, and MACD improves decision accuracy, which hybrid models later adopted as part of feature representations.

Several studies expanded beyond traditional price-based features: GVMD-Q-DBN-LSTM-GRU and FreQuant [23] integrated frequency-domain features, DeepScalper [18] employed multi-modality embeddings(encoder-decoder architecture to represent the market from the micro-level and macro-level, respectively), and Yu-Fu Chen's multimodal RL fused sentiment and news signals with market data.

Such multi-dimensional input strategies provide agents with a richer state representation, essential for navigating complex market structures. The evolution of input design thus mirrors the broader trend toward multimodal, decomposed, and context-aware learning in algorithmic trading.

4.5 Limitations, Risks, and Emerging Research Directions

Despite their successes, the studies reviewed also highlight several limitations.

- Overfitting and regime dependence: Many RL agents perform well in specific datasets but degrade when market conditions shift.
- Training cost and data efficiency: PPO [22] and deep ensemble methods require extensive computation, limiting scalability.
- Reward design complexity: Sparse or delayed rewards complicate policy optimization, motivating innovations such as hindsight bonuses (DeepScalper) [18] or imitation-guided rewards [8] (HAT, SIRL).

- Practical deployment challenges: Latency, transaction costs, and regulatory constraints remain under-addressed in most experimental setups.

To mitigate these risks, researchers increasingly advocate for risk-aware objectives, online retraining mechanisms, and realistic backtesting environments. The integration of volatility forecasts [21] and meta-RL controllers [22] represents a promising move toward adaptive and dynamically constrained learning frameworks.

In addition, this review may be subject to reporting bias, as studies with positive or statistically significant results are more likely to be published, while negative or inconclusive findings may be underrepresented in the literature.

4.6 Synthesis and Implications for Future Work

The overall direction of the field points toward hierarchical, hybrid, and interpretable RL architectures. Indicator-based rule systems provide explainable low-level signals, while hybrid models (e.g., Q-VMD-ANN-LSTM-GRU) [21] and frequency-domain RL systems (FreQuant) supply robust high-level features.

Meta-RL controllers like PPO [22] for hedge funds enable adaptive strategy selection, ensuring that trading agents remain aligned with changing market dynamics. The inclusion of sentiment, volatility, and multi-frequency features fosters a multi-perspective understanding of market behavior, improving both performance and resilience.

In essence, the convergence of reinforcement learning, deep neural architectures, and hybrid financial modeling is driving algorithmic trading toward a new phase of autonomous, risk-sensitive, and human-aligned decision systems.

Future research should prioritize real-world validation, online learning adaptation, and explainability, ensuring that RL-based trading systems can transition from experimental prototypes to sustainable financial technologies.

The overall certainty of the evidence presented in this review is considered moderate. This is due to the heterogeneity of datasets, differences in evaluation methodologies, and the limited availability of real-world validation across studies. As a result, while the findings are indicative of strong trends, they should be interpreted with caution.

5. Conclusions

This literature review highlights the growing sophistication of reinforcement learning (RL) in algorithmic trading, illustrating how the field has transitioned from basic Q-learning and single-agent systems to hybrid, hierarchical, and risk-aware architectures. Across all reviewed studies, RL has proven effective in capturing complex temporal and structural market dynamics, outperforming traditional rule-based and statistical approaches when properly designed and tuned.

Several consistent findings emerge. First, hybrid and multi-agent approaches—such as those integrating PPO, DQN, and decomposition frameworks—demonstrate superior adaptability and stability under changing market conditions. Second, signal decomposition and multimodal inputs (including sentiment and frequency-domain features) improve robustness and enhance the interpretability of trading decisions.

Third, risk awareness and imitation learning are increasingly recognized as essential components for ensuring that trading agents make safer, more realistic, and human-aligned decisions.

Nevertheless, the literature also reveals critical challenges that remain unresolved. Many RL models suffer from overfitting, high computational cost and limited generalization beyond the training environment. Sparse or delayed rewards complicate the optimization process, while limited access to high-quality, real-time financial data restricts scalability and live deployment. Furthermore, most experimental setups overlook transaction costs, slippage and regulatory constraints, which are essential for realistic profitability assessments.

Overall, the synthesis suggests that the future of RL in algorithmic trading lies in developing adaptive, explainable and data-efficient agents that integrate deep reinforcement learning with hybrid signal processing, volatility forecasting and meta-learning strategies. Future research should focus on:

1. Designing real-time adaptive agents capable of continuous online learning.

2. Embedding interpretability and transparency within RL architectures to improve trust and regulatory compliance.

3. Validating models using live trading simulations that incorporate realistic execution constraints and transaction costs.

In conclusion, reinforcement learning has the potential to transform algorithmic trading from static rule-based strategies into autonomous, intelligent and risk-aware decision systems. Continued progress depends on bridging the gap between algorithmic innovation and practical deployment in live financial markets.

References

1. Nayeri, Z.M.; Ghafarian, T.; Javadi, B. Application Placement in Fog Computing with AI Approach: Taxonomy and a State of the Art Survey. *Journal of Network and Computer Applications* **2021**, *185*, 103078, doi:10.1016/j.jnca.2021.103078.
2. Lin, B. Reinforcement Learning and Bandits for Speech and Language Processing: Tutorial, Review and Outlook. *Expert Systems with Applications* **2024**, *238*, 122254, doi:10.1016/j.eswa.2023.122254.
3. van Hasselt, H.; Guez, A.; Silver, D. Deep Reinforcement Learning with Double Q-Learning 2015.
4. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods 2018.
5. Kitchenham, Barbara, C., Stuart *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Keele University and Durham University, 2007;
6. LaPES (Laboratório de Pesquisa em Engenharia de Software) StArt (State of the Art).
7. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. *PRISMA2020* : An R Package and Shiny App for Producing PRISMA 2020-compliant Flow Diagrams, with Interactivity for Optimised Digital Transparency and Open Synthesis. *Campbell Systematic Reviews* **2022**, *18*, e1230, doi:10.1002/cl2.1230.
8. Ye, Z.J.; Schuller, B.W. Human-Aligned Trading by Imitative Multi-Loss Reinforcement Learning. *Expert Systems with Applications* **2023**, *234*, 120939, doi:10.1016/j.eswa.2023.120939.
9. Takara, L.D.A.; Santos, A.A.P.; Mariani, V.C.; Coelho, L.D.S. Deep Reinforcement Learning Applied to a Sparse-Reward Trading Environment with Intraday Data. *Expert Systems with Applications* **2024**, *238*, 121897, doi:10.1016/j.eswa.2023.121897.
10. Yu, Y.; Lin, Y.; Hou, X.; Zhang, X. Novel Optimization Approach for Realized Volatility Forecast of Stock Price Index Based on Deep Reinforcement Learning Model. *Expert Systems with Applications* **2023**, *233*, 120880, doi:10.1016/j.eswa.2023.120880.
11. Lin, Y.-C.; Chen, C.-T.; Sang, C.-Y.; Huang, S.-H. Multiagent-Based Deep Reinforcement Learning for Risk-Shifting Portfolio Management. *Applied Soft Computing* **2022**, *123*, 108894, doi:10.1016/j.asoc.2022.108894.

12. Shavandi, A.; Khedmati, M. A Multi-Agent Deep Reinforcement Learning Framework for Algorithmic Trading in Financial Markets. *Expert Systems with Applications* **2022**, *208*, 118124, doi:10.1016/j.eswa.2022.118124.
13. Massahi, M.; Mahootchi, M. A Deep Q-Learning Based Algorithmic Trading System for Commodity Futures Markets. *Expert Systems with Applications* **2024**, *237*, 121711, doi:10.1016/j.eswa.2023.121711.
14. Kim, S.; Kim, J.; Sul, H.K.; Hong, Y. An Adaptive Dual-Level Reinforcement Learning Approach for Optimal Trade Execution 2023.
15. Dong, L.; Zheng, H. Soft Imitation Reinforcement Learning with Value Decomposition for Portfolio Management. *Applied Soft Computing* **2024**, *151*, 111108, doi:10.1016/j.asoc.2023.111108.
16. Chen, Y.-F.; Huang, S.-H. Sentiment-Influenced Trading System Based on Multimodal Deep Reinforcement Learning. *Applied Soft Computing* **2021**, *112*, 107788, doi:10.1016/j.asoc.2021.107788.
17. Carta, S.; Ferreira, A.; Podda, A.S.; Reforgiato Recupero, D.; Sanna, A. Multi-DQN: An Ensemble of Deep Q-Learning Agents for Stock Market Forecasting. *Expert Systems with Applications* **2021**, *164*, 113820, doi:10.1016/j.eswa.2020.113820.
18. Sun, S.; Xue, W.; Wang, R.; He, X.; Zhu, J.; Li, J.; An, B. DeepScalper: A Risk-Aware Reinforcement Learning Framework to Capture Fleeting Intraday Trading Opportunities 2022.
19. Zengeler, N.; Handmann, U. Contracts for Difference: A Reinforcement Learning Approach. *JRFM* **2020**, *13*, 78, doi:10.3390/jrfm13040078.
20. Mengshetti, O.; Gupta, K.; Zade, N.; Kotecha, K.; Mutha, S.; Joshi, G. Synergizing Quantitative Finance Models and Market Microstructure Analysis for Enhanced Algorithmic Trading Strategies. *Journal of Open Innovation: Technology, Market, and Complexity* **2024**, *10*, 100334, doi:10.1016/j.joitmc.2024.100334.
21. Kumar, S.; Rao, A.; Dhochak, M. Hybrid ML Models for Volatility Prediction in Financial Risk Management. *International Review of Economics & Finance* **2025**, *98*, 103915, doi:10.1016/j.iref.2025.103915.
22. Belyakov, B.; Sizykh, D. Adaptive Algorithm for Selecting the Optimal Trading Strategy Based on Reinforcement Learning for Managing a Hedge Fund. *IEEE Access* **2024**, 1–1, doi:10.1109/ACCESS.2024.3515039.
23. Jeon, J.; Park, J.; Park, C.; Kang, U. FreQuant: A Reinforcement-Learning Based Adaptive Portfolio Optimization with Multi-Frequency Decomposition. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; ACM: Barcelona Spain, August 25 2024; pp. 1211–1221.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.