



Financial Management Perspective
Journal homepage: <https://jfmp.sbu.ac.ir/>



Original Article

Dynamic Portfolio Management in Cryptocurrency Markets: A Deep Reinforcement Learning Approach with a Risk-Sensitive Adaptive Agent Framework

Saeed Eazi*

¹ Abdolrasoul Mostajeran**

Abstract

Introduction: This study aims to propose and evaluate a dynamic framework for portfolio optimization in cryptocurrency markets that ensures an effective balance between return maximization and risk control under conditions of high volatility. The primary objective is to develop a risk-sensitive adaptive agent that integrates multiple baseline deep reinforcement learning agents with a dynamic switching mechanism to adjust asset weight allocations over time. This adaptive structure enables conservative behavior during highly volatile market regimes and more aggressive positioning during trending periods. The empirical analysis is conducted on a selected set of five major cryptocurrencies-Bitcoin, Ethereum, Solana, Ripple, and Tether-to assess the effectiveness and practical applicability of the proposed framework in real-world market environments.

Method: Daily data spanning five years were collected from Yahoo Finance and, after data cleaning, were divided into training and testing sets using a 70%–30% split. The state representation comprised asset prices, trading volumes, and commonly used technical indicators. To capture long-term temporal dependencies, state encoding was implemented using a Transformer-based architecture. Two categories of methods were examined: (1) individual deep reinforcement learning agents-namely PPO, A2C, and DQN-each independently managing the portfolio; and (2) a risk-sensitive adaptive framework that dynamically switches among agents by monitoring short-term performance indicators such as rolling cumulative profit and loss and five-day moving volatility. The reward function incorporated risk-sensitive components, including a logarithmic utility of returns, a volatility penalty, and the Calmar ratio. Performance

Received: 2026. February .22, Accepted: 2026. June .21.

*PhD student , Department of Financial Management, South Tehran branch, Islamic Azad University, Tehran, Iran

**Faculty Member, Faculty of Financial Sciences, Kharazmi University, Tehran, Iran. (Corresponding Author).

E-mail: rasoul.mostajeran@khu.ac.ir

evaluation was conducted using cumulative return, annualized return, Sharpe ratio, Sortino ratio, Calmar ratio, and maximum drawdown.

Findings: On the out-of-sample test dataset, the proposed risk-sensitive adaptive framework consistently outperformed each individual deep reinforcement learning agent. During non-trending market periods, the proposed model achieved higher terminal portfolio value and superior risk-adjusted performance compared to alternative methods. For instance, it recorded a Sharpe ratio of approximately 1.13 and a Sortino ratio of about 1.81. In bullish market regimes, the framework delivered markedly stronger results, achieving a Sharpe ratio of approximately 1.73, a Calmar ratio of around 3.34, and higher annualized returns, such that the final portfolio value during the uptrend was nearly twice that of the best-performing standalone agent. The quantitative results indicate that dynamic switching among aggressive, balanced, and conservative agents enhances the exploitation of favorable market trends while mitigating losses during highly volatile periods. Although a temporary increase in maximum drawdown was observed in certain scenarios, this effect was effectively offset by a reversion mechanism that shifts the agent back toward conservative behavior.

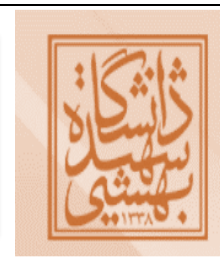
Conclusion: Integrating a Transformer architecture for temporal feature extraction with a risk-sensitive adaptive agent framework provides an effective solution for dynamic portfolio management in cryptocurrency markets, particularly when the objective is to simultaneously enhance returns and control downside risk. The findings indicate that employing a dynamic switching mechanism among multiple deep reinforcement learning agents can significantly improve portfolio performance and stability when facing shifting market regimes.

Keywords: Adaptive Portfolio Management; Deep Reinforcement Learning; Risk Sensitive Adaptive Agent; Risk Management; Volatile Cryptocurrency Market

How to Cite Eazi, , S. and Mostajeran, , A. (2026). Dynamic Portfolio Management in Cryptocurrency Markets: A Deep Reinforcement Learning Approach with a Risk Sensitive Adaptive Agent Framework. *Financial Management Perspective*, 16(1), 101-131. (in Persian)



Copyright: © 2026 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/>)



نوع مقاله: پژوهشی

مدیریت پویای پرتفوی رمز ارزها با استفاده از یادگیری تقویتی عمیق و چارچوب عامل تطبیقی حساس به ریسک

سعید ایزی *

عبدالرسول مستاجران **

چکیده

هدف: ارائه و ارزیابی چارچوبی پویا برای بهینه‌سازی پرتفوی در بازار رمزارزها که توازن میان بازده و کنترل ریسک را در شرایط نوسانی زیاد تضمین کند. در این پژوهش هدف اصلی توسعه عامل تطبیقی حساس به ریسک است که با ترکیب چند عامل پایه یادگیری تقویتی عمیق و مکانیزم جابجایی پویا، تخصیص وزن دارایی‌ها را به‌صورت زمان‌مند تنظیم کرده و رفتار محافظه‌کارانه در دوره‌های پرنوسان و رفتار تهاجمی در دوره‌های رونددار را برقرار می‌سازد. دامنه تجربی شامل پنج رمزارز منتخب شامل بیت‌کوین، اتریوم، سولانا، ریپل و تتر است تا کارایی چارچوب در محیط‌های واقعی بازار ارزیابی شود.

روش: داده‌های روزانه پنج ساله از پایگاه یاهو فایننس گردآوری و پس از پیش‌پردازش با تقسیم ۷۰٪ برای آموزش و ۳۰٪ برای آزمون آماده شدند. نمایه حالت شامل قیمت‌ها و شاخص‌های تکنیکال رایج بود و نمایش حالت با استفاده از معماری ترنسفورمر برای استخراج وابستگی‌های بلندمدت زمانی صورت گرفت. دو دسته روش مورد بررسی قرار گرفت: ۱- عوامل منفرد یادگیری تقویتی عمیق شامل PPO، A2C و DQN که هر یک مستقل پرتفوی را مدیریت می‌کنند و ۲- چارچوب تطبیقی حساس به ریسک که با پایش معیارهای کوتاه‌مدت مانند سود و زیان تجمعی پنجره‌ای و نوسان متحرک پنج روزه، بین عوامل سوئیچ می‌کند. تابع پاداش از مؤلفه‌های حساس به ریسک شامل تابع مطلوبیت لگاریتمی بازده، جریمه برای نوسان و نسبت کالمار تشکیل شد و معیارهای ارزیابی شامل بازده تجمعی، بازده سالیانه، نسبت شارپ، نسبت سورتینو، نسبت کالمار و حداکثر افت سرمایه بودند.

یافته‌ها: در مجموعه داده‌های آزمایشی، چارچوب تطبیقی حساس به ریسک عملکرد برتری نسبت به هر یک از عوامل منفرد نشان داد. در دوره‌های بدون روند بازار، مدل پیشنهادی ارزش نهایی پرتفو و نسبت‌های تعدیل شده بر ریسک بالاتری نسبت به سایر روش‌ها ثبت کرد. برای مثال نسبت شارپ ≈ ۱.۱۳ ، نسبت سورتینو ≈ ۱.۸۱ است و در دوره‌های صعودی بازار به‌طور

تاریخ پذیرش: ۱۴۰۵/۰۳/۳۱

تاریخ دریافت: ۱۴۰۴/۱۲/۰۳

* دانشجوی دکترا مهندسی مالی، دانشکده اقتصاد و حسابداری، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران.

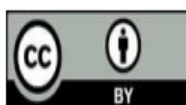
** عضو هیات علمی دانشکده علوم مالی، دانشگاه خوارزمی، تهران، ایران. (نویسنده مسئول: Email: rasoul.mostajeran@khu.ac.ir)

چشمگیری نسبت شارپ $\approx 1,73$ ، نسبت کالمار $\approx 3,34$ و بازده سالیانه بالاتری ارائه داد؛ به طوری که ارزش نهایی پرتفو در دوره صعودی تقریباً دو برابر بهترین عامل مستقل شد. نتایج کمی نشان دادند که جابجایی پویا بین عوامل تهاجمی، متعادل و محافظه کار موجب افزایش بهره برداری از روندهای مثبت و کاهش زیان در دوره های پرنوسان گردید، هرچند افت حداکثر در برخی سناریوها افزایشی مقطعی نشان داد که با مکانیسم بازگشتی به رفتار محافظه کار جبران شد.

نتیجه گیری: ترکیب معماری ترنسفورمر برای استخراج ویژگی های زمانی با چارچوب عامل تطبیقی حساس به ریسک، راهکار مؤثری برای مدیریت پویای پرتفوی در بازار رمزارزها ارائه می دهد؛ به ویژه زمانی که هدف، افزایش بازده و کنترل ریسک نزولی به طور همزمان است. یافته ها نشان می دهد استفاده از مکانیسم سوئیچ پویا میان چند عامل یادگیری تقویتی عمیق می تواند کارایی و پایداری پرتفو را در مواجهه با تغییر رژیم های بازار به طور معناداری بهبود بخشد. همچنین، نتایج آزمون رتبه علامت ویلکاکسون بر پایه ده اجرای مستقل نشان داد که برتری مدل پیشنهادی نسبت به مدل های PPO، A2C و DQN در تمامی شاخص های اصلی عملکرد از نظر آماری معنادار است ($P < 0.01$). این یافته بیانگر پایداری، استحکام و قابلیت تکرارپذیری نتایج مدل پیشنهادی بوده و نشان می دهد که عملکرد برتر آن صرفاً ناشی از نوسانات تصادفی فرآیند آموزش نیست.

کلیدواژه ها: مدیریت پرتفو تطبیقی، یادگیری تقویتی عمیق، عامل تطبیقی حساس به ریسک، مدیریت ریسک، بازار پرنوسان رمزارز

استناد دهی: ایزی، سعید، مستاجران، عبدالرسول (۱۴۰۵). مدیریت پویای پرتفوی رمزارزها با استفاده از یادگیری تقویتی عمیق و چارچوب عامل تطبیقی حساس به ریسک. چشم انداز مدیریت مالی، ۱۶(۱)، ۱۰۱-۱۳۱.



۱. مقدمه

بازارهای رمزارز طی سال‌های اخیر به یکی از مهم‌ترین طبقات دارایی نوظهور تبدیل شده‌اند و سرمایه‌گذاران خرد و نهادی به دلیل بازده‌های قابل توجه و ویژگی‌های منحصر به فرد این بازارها به آن توجه نشان داده‌اند. با این حال، نوسانات بسیار بالا، رفتارهای غیرایستا، واکنش‌پذیری لحظه‌ای به اخبار و تغییرات رژیم بازار، چالش‌هایی جدی برای مدیریت پرتفو در این طبقه دارایی ایجاد می‌کند. روش‌های کلاسیک بهینه‌سازی پرتفو مانند مدل میانگین-واریانس مارکوویتز، متکی بر فرضی نظیر نرمال بودن توزیع بازده‌ها و ثبات همبستگی‌ها هستند؛ مفروضاتی که در بازار رمزارز به‌طور سیستماتیک نقض می‌شوند و در نتیجه باعث می‌شود سرمایه‌گذاران در معرض افت شدید سرمایه و ریسک‌های غیرخطی قرار گیرند ([باهو و همکاران، ۲۰۲۴](#)). تحقیقات اخیر نشان می‌دهند که تنظیم پویا و تطبیقی استراتژی‌های پرتفو با سطوح ریسک و شرایط متغیر بازار، به ویژه در محیط‌های پرنوسان، امری حیاتی است ([جفری و همکاران، ۲۰۲۵](#)). به‌عنوان نمونه، در بازارهای صندوق‌های قابل معامله، رویکردهای جایگزین مدل‌سازی توانسته‌اند عملکرد تعدیل‌شده بر اساس ریسک را در شرایط عدم قطعیت بازار بهبود دهند و نشان دهند که مدل‌های ایستا برای محیط‌های پرنوسان کافی نیستند. با وجود گسترش کاربرد تکنیک‌های هوش مصنوعی، به‌ویژه یادگیری تقویتی عمیق^۱، در تصمیم‌گیری مالی، بخش قابل توجهی از پژوهش‌های پیشین با چند محدودیت بنیادی روبه‌رو هستند که شکاف اصلی این پژوهش را شکل می‌دهد. نخست آنکه بخش عمده مطالعات تنها از یک عامل واحد یادگیری تقویتی عمیق استفاده کرده و فرض می‌کنند یک سیاست ثابت می‌تواند در تمامی رژیم‌های بازار عملکرد مناسبی داشته باشد؛ حال آن‌که شواهد تجربی نشان می‌دهد عوامل منفرد در مواجهه با تغییرات پیچیده و ناگهانی بازار رمزارز عملکردی شکننده دارند ([کوراتامدی و همکاران، ۲۰۲۱](#)) ([چن و همکاران، ۲۰۲۱](#)). دوم آنکه بیشتر تحقیقات موجود فاقد سازوکارهای حساس به ریسک داخلی‌اند و شاخص‌هایی همچون نوسان نزولی، افت سرمایه، یا نسبت کالمار به طور صریح در طراحی تابع پاداش مدل لحاظ نشده‌اند. سوم آنکه بخش عمده مطالعات قادر به مدل‌سازی وابستگی‌های زمانی بلندمدت و ساختارهای پیچیده بین رمزارزها نیستند؛ مشکلی که ضرورت استفاده از معماری‌های مبتنی بر ترنسفورمر^۲ را توجیه می‌کند. ([واسوانی و همکاران ۲۰۱۷](#)) در پاسخ به این شکاف‌ها، پژوهش حاضر چارچوبی با عنوان "عامل تطبیقی حساس به ریسک" ارائه می‌کند که سه ویژگی کلیدی دارد:

۱- ترکیب و جابه‌جایی پویا میان چند عامل یادگیری تقویتی عمیق شامل PPO^۳، A2C^۴ و DQN^۵ بر اساس شرایط بازار؛

۲- استفاده از تابع پاداش حساس به ریسک که معیارهایی مانند نوسان پنجره‌ای، افت سرمایه و سود-زیان تجمعی کوتاه‌مدت را در تصمیم‌گیری لحاظ می‌کند؛

۳- استفاده از معماری ترنسفورمر برای نمایش حالت و استخراج وابستگی‌های زمانی بلندمدت که برای بازار رمزارز اهمیت حیاتی دارد.

بازارهای رمزارز مورد بررسی شامل بیت‌کوین، اتریوم، سولانا، ریپل و تتر از نظر نوسان، نقدشوندگی و ساختار بنیادین تفاوت قابل توجهی با سهام یا کالا دارند. وابستگی شدید به جریان‌های خبری، رفتار غیرایستا و وقوع مکرر شوک‌های قیمت باعث می‌شود مدل‌هایی که قادر به واکنش پویا به نوسانات نیستند، عملکرد پایداری نداشته باشند. در این زمینه، ادغام شاخص‌های حساس به ریسک و مکانیزم جابه‌جایی پویا بین سیاست‌ها می‌تواند در بهبود عملکرد تعدیل‌شده بر اساس ریسک نقش مهمی داشته باشد.

^۱ Deep Reinforcement Learning

^۲ Transformer

^۳ Proximal Policy Optimization

^۴ Advantage Actor-Critic

^۵ Deep Q-Network

در این مطالعه، دو الگوریتم مورد بررسی قرار گرفته‌اند: الگوریتم اول شامل عوامل منفرد یادگیری تقویتی (PPO)، (A2C و DQN) است و الگوریتم دوم چارچوب عامل تطبیقی حساس به ریسک را پیاده‌سازی می‌کند. این چارچوب به‌صورت پویا بین عوامل مختلف سوئیچ می‌کند تا در شرایط بازار متغیر بهترین عملکرد تعدیل‌شده بر اساس ریسک را ارائه دهد.

چارچوب عامل تطبیقی حساس به ریسک علاوه بر قابلیت تطبیق پویا، شاخص‌های حساس به ریسک را نیز در طراحی خود جای داده است. این مدل بر اساس معیارهایی مانند نسبت واریانس کوتاه‌مدت به بلندمدت، نوسانات بازار را شناسایی کرده و تخصیص دارایی را به‌صورت خودکار تعدیل می‌کند. بدین ترتیب، در دوره‌های پرنوسان، سهم دارایی نقدی (تر) افزایش یافته و ریسک کلی پرتفو کنترل می‌شود. دوره‌های بازار پرنوسان به‌صورت کمی به عنوان بازه‌هایی تعریف می‌شوند که در آن انحراف معیار متحرک بازده‌های روزانه بیش از یک انحراف معیار از میانگین بلندمدت تاریخی فراتر می‌رود و بدین ترتیب، دوره‌هایی با عدم قطعیت و فشار بالای بازار شناسایی می‌شوند. با ادغام شاخص‌های عملکرد حساس به ریسک و مکانیسم پویا برای تغییر استراتژی، این رویکرد بینش‌های قابل اجرا و کاربردی برای سرمایه‌گذاران نهادی و معامله‌گران آگاه به ریسک فراهم می‌آورد. برای ارزیابی مقایسه‌ای، عملکرد چارچوب پیشنهادی در برابر استراتژی سنتی خرید و نگهداری سنجیده شده است.

ساختار مقاله به این صورت است: بخش دوم به مبانی نظری و مرور پیشینه و کاربرد یادگیری تقویتی عمیق در بازارهای مالی و رمزارزها می‌پردازد، بخش سوم روش تحقیق و طراحی مدل‌ها و چارچوب عامل تطبیقی حساس به ریسک توضیح داده می‌شود، بخش چهارم نتایج و مقایسه عملکرد الگوریتم‌ها ارائه می‌شود و بخش پنجم شامل نتیجه‌گیری، بحث در مورد دانش افزوده و توصیه‌های مدیریتی برای سرمایه‌گذاران است.

۲. مبانی نظری و پیشینه پژوهش

مدیریت پرتفوی همواره یکی از مسائل بنیادین در مالی مدرن بوده است و نظریه پرتفوی مارکوویتز (۱۹۵۲) به عنوان نقطه آغازین رویکردهای کمی در تخصیص دارایی شناخته می‌شود. در این نظریه، سرمایه‌گذار با ایجاد تعادل میان بازده مورد انتظار و ریسک، که معمولاً با واریانس بازده اندازه‌گیری می‌شود، ترکیب بهینه‌ای از دارایی‌ها را انتخاب می‌کند. چارچوب میانگین-واریانس مارکوویتز بر چند فرض اساسی استوار است؛ از جمله نرمال بودن توزیع بازده‌ها، ثبات ساختار کوواریانس میان دارایی‌ها و ایستایی رفتار بازار در طول زمان. بر مبنای این فروض، روابط میان دارایی‌ها قابل برآورد بوده و می‌توان با استفاده از داده‌های تاریخی، ترکیب بهینه سرمایه‌گذاری را تعیین کرد.

با وجود اهمیت نظریه پرتفوی مدرن، مطالعات اخیر نشان داده‌اند که بسیاری از این فروض در بازارهای مالی نوظهور، به‌ویژه بازار رمزارزها، به‌طور سیستماتیک نقض می‌شوند (فابوزی و همکاران، ۲۰۲۲). بازار رمزارزها دارای ویژگی‌هایی همچون نوسانات شدید، تغییرات ناگهانی قیمت، وابستگی بالا به اخبار و احساسات سرمایه‌گذاران و رفتارهای غیرخطی است؛ ویژگی‌هایی که سبب می‌شوند بازده دارایی‌ها از توزیع نرمال تبعیت نکنند و ساختار ریسک بازار در طول زمان به‌صورت مداوم تغییر کند. در چنین محیطی، بازده‌ها اغلب دارای کشیدگی بالا و دنباله‌های سنگین هستند؛ به این معنا که احتمال وقوع شوک‌های قیمتی شدید بسیار بیشتر از آن چیزی است که مدل‌های کلاسیک مالی پیش‌بینی می‌کنند. در نتیجه، استفاده از معیارهای مبتنی بر واریانس به‌تنهایی نمی‌تواند تصویر دقیقی از ریسک واقعی بازار ارائه دهد. علاوه بر این، بازار رمزارزها دارای دوره‌های متوالی آرام و پرنوسان است و شدت ریسک در آن به‌صورت پیوسته تغییر می‌کند. همچنین، تغییر مکرر وضعیت بازار از صعودی به نزولی یا بدون روند، موجب ناپایداری روابط میان دارایی‌ها در

طول زمان می‌شود. در نتیجه، فرض ثبات ریسک و همبستگی دارایی‌ها که از مبانی اصلی نظریه میانگین-واریانس است، در بازار رمزارزها اعتبار محدودی دارد.

در چنین شرایطی، استفاده از مدل‌های ایستا برای مدیریت پرتفوی می‌تواند به تخصیص غیربهبوده دارایی‌ها، برآورد نادرست ریسک و افزایش احتمال افت شدید سرمایه منجر شود. پژوهش‌های اخیر تلاش کرده‌اند با به‌کارگیری یادگیری ماشین و مدل‌های خودتطبیق، اصول نظریه پرتفوی مدرن را به فضای تصمیم‌گیری پویا گسترش دهند. در این راستا، استفاده از الگوریتم‌های یادگیری تقویتی عمیق به‌عنوان رویکردی نوین در مدیریت پرتفو مورد توجه قرار گرفته است. یادگیری تقویتی عمیق با ترکیب توانایی شبکه‌های عصبی عمیق^۱ در استخراج الگوهای غیرخطی و سازوکار یادگیری مبتنی بر پاداش، امکان یادگیری سیاست‌های بهینه در محیط‌های مالی پویا را فراهم می‌سازد (ساتون و همکاران، ۲۰۱۸). برخلاف مدل‌های کلاسیک که پارامترهای ثابتی دارند، یادگیری تقویتی عمیق قادر است از داده‌های زمان‌مند، روابط بین دارایی‌ها و تغییرات لحظه‌ای بازار را بیاموزد و تصمیمات تخصیص دارایی را بر این اساس به‌روزرسانی کند. در محیط‌های مالی، بیشینه‌سازی صرف بازده نمی‌تواند معیار مناسبی برای تصمیم‌گیری باشد، زیرا سرمایه‌گذاران نسبت به زیان و افت سرمایه حساسیت نامتقارن دارند؛ از این رو، یادگیری تقویتی حساس به ریسک تلاش می‌کند تابع هدف را از بازده خام به بازده تعدیل‌شده بر اساس ریسک تغییر دهد. الگوریتم‌های PPO، A2C و DQN از جمله پرکاربردترین روش‌های یادگیری تقویتی عمیق در بهینه‌سازی پرتفو هستند. الگوریتم PPO با به‌روزرسانی پایدار، تعادلی میان کاوش و بهره‌برداری ایجاد می‌کند؛ A2C با یادگیری هم‌زمان سیاست و ارزش، سرعت همگرایی بالایی دارد؛ و DQN برای محیط‌های گسسته، سیاست‌هایی کارآمد با استفاده از تقریب تابع ارزش ارائه می‌دهد. ترکیب این الگوریتم‌ها می‌تواند طیف وسیعی از شرایط بازار را پوشش دهد و از ضعف‌های هر کدام به‌صورت منفرد بکاهد (کوراتامدی و همکاران، ۲۰۲۱). با وجود کارایی یادگیری تقویتی عمیق در مدیریت پرتفو، مطالعات نشان داده‌اند که استفاده از یک عامل منفرد در محیط‌های پرنوسان می‌تواند به شکستگی عملکرد منجر شود (لین و همکاران، ۲۰۲۳). در پاسخ به این چالش، رویکرد عامل تطبیقی حساس به ریسک به‌عنوان چارچوبی برای انتخاب پویا میان سیاست‌های مختلف بر اساس عملکرد اخیر معرفی شده است. این عامل با رصد شاخص‌های عملکردی مانند سود و زیان تجمعی و حداکثر افت سرمایه، در دوره‌های تغییر رژیم بازار، استراتژی‌های مناسب‌تری را فعال می‌کند (وانگ و لیو، ۲۰۲۵).

در پژوهش حاضر، ساختاری الهام‌گرفته از رویکرد فوق اتخاذ شده است، اما تمرکز اصلی بر عملکرد سه الگوریتم یادگیری تقویتی عمیق PPO، A2C، DQN است تا کارایی آن‌ها در شرایط پرنوسان بازار رمزارز سنجیده شود. ویژگی متمایز این مطالعه، تخصیص اولیه ۱۰۰ درصد سرمایه به دارایی نقدی است. در این چارچوب، تمام وزن پرتفو در آغاز به تتر اختصاص می‌یابد و سپس الگوریتم‌ها به‌تدریج سایر رمزارزها را به پرتفو اضافه می‌کنند. این طراحی امکان تحلیل اثر ورود تدریجی دارایی‌های پرریسک بر عملکرد کلی و کنترل نوسانات را فراهم می‌سازد و با منطبق تصمیم‌گیری تدریجی در نظریه‌های تطبیقی مالی هم‌راستا است. تابع پاداش مورد استفاده، مبتنی بر معیارهای حساس به ریسک است و به‌گونه‌ای طراحی شده که افت سرمایه و نوسان را در فرآیند یادگیری عامل جریمه می‌کند. ساختار کلی آن از تابع پاداش وانگ و لیو (۲۰۲۵) اقتباس شده است، با ضرایبی که رفتار عامل را به سمت کنترل ریسک هدایت می‌کنند. این رویکرد باعث می‌شود عامل در دوره‌های پرنوسان، ترجیح بیشتری به دارایی‌های کم‌ریسک مانند تتر نشان دهد و در دوره‌های آرام‌تر به تخصیص مجدد دارایی‌ها بپردازد.

پژوهش‌های داخلی نیز به‌تدریج به سمت به‌کارگیری هوش مصنوعی در مدیریت پرتفو حرکت کرده‌اند. زارع و همکاران (۲۰۲۰) از یک شبکه یادگیری عمیق DNN برای تشکیل پرتفو بهینه استفاده کردند. عبدی و همکاران (۲۰۲۲) یک مدل ترکیبی بهینه‌سازی سبد سهام با کمک شبکه عصبی LSTM ارائه کردند که بازدهی بیشتری نسبت

^۱ Deep Neural Network

به روش های سنتی داشت. [غلامی و همکاران \(۲۰۲۴\)](#) با ارائه مدلی برای پیش بینی قیمت سهام با استفاده از مدل LSTM-CNN توانستند در دقت پیش بینی بهبود حاصل کنند. [دیلی و همکاران \(۲۰۲۴\)](#) نشان داد که استفاده از الگوریتم یادگیری تقویتی عمیق با ساختار شبکه عصبی مبتنی بر الگوریتم لونبرگ مارد می تواند عملکرد سبد سهام را در مقایسه با روشهای کلاسیک به طور معناداری بهبود بخشد. [آیت الهی و همکاران \(۲۰۲۴\)](#) با ترکیب چارچوب یادگیری تقویتی و شبکه ترنسفورمر ها، دقت بالاتر پیش بینی و افزایش سودآوری عملکرد پرتفو را نمایش دادند. علاوه بر این، پژوهش های جدیدتر بر نقش مدل های یادگیری تقویتی در کنترل ریسک پرتفو و بهبود پایداری عملکرد تأکید دارند.

در حوزه مدل سازی داده های مالی، معماری ترنسفورمر به دلیل توانایی در درک وابستگی های بلندمدت زمانی و بین داری، بر مدل های بازگشتی سنتی مانند LSTM¹ برتری دارد [\(واسوانی و همکاران ۲۰۱۷\)](#). پژوهش [وانگ و لیو \(۲۰۲۵\)](#) نیز از این معماری برای پردازش حالت محیط بهره برد و نتایج بهتری نسبت به ساختارهای کلاسیک گزارش کرد. این یافته ها نشان می دهد که بهره گیری از ترنسفورمر در استخراج الگوهای ناپایدار بازار می تواند به بهبود عملکرد الگوریتم های یادگیری تقویتی عمیق در تصمیم گیری سرمایه گذاری منجر شود. پیشینه پژوهش در بازارهای مالی نشان می دهد که یادگیری تقویتی عمیق در محیط های پرنوسان نسبت به روش های سنتی کارایی بالاتری دارد. [دنگ و همکاران \(۲۰۱۶\)](#) با استفاده از الگوریتم PPO در بازار رمزارزها، نسبت شارپ بالاتری نسبت به روش خرید و نگهداری گزارش کردند. [زانگ و همکاران \(۲۰۲۰\)](#) نیز نشان دادند که یادگیری تقویتی عمیق قادر است در محیط های تغییر رژیم بازار، انطباق سریع تری نسبت به مدل های آماری نشان دهد.

بیشتر پژوهش های اولیه مدیریت پرتفو مبتنی بر یادگیری تقویتی، بر عامل های منفرد و سیاست های ثابت متمرکز بوده اند که در بازارهای پرنوسان و غیرایستای رمزارزها، محدودیت هایی در سازگاری و کنترل ریسک دارند. در سال های اخیر، برخی مطالعات تلاش کرده اند با استفاده از چارچوب های چندعاملی، انعطاف پذیری و توان سازگاری مدل های یادگیری تقویتی را افزایش دهند. برای نمونه، پژوهش «مدیریت پرتفو مبتنی بر یادگیری تقویتی عمیق چندعاملی با جابه جایی ریسک» ساختاری چندعاملی ارائه می کند که در آن عامل ها رفتارهای متفاوتی نسبت به ریسک دارند و به صورت مشارکتی تصمیم گیری می کنند. نتایج این پژوهش نشان می دهد که استفاده از چند عامل می تواند پایداری تصمیم گیری را نسبت به مدل های تک عاملی بهبود دهد [\(لین و همکاران، ۲۰۲۲\)](#). همچنین، مطالعه دیگری با ارائه چارچوب تطبیقی چندعاملی برای معاملات چندداری، نشان داد که عامل های چندگانه در محیط های آشوبناک و متغیر بازار عملکرد اثبات تری نسبت به عامل های منفرد دارند و می توانند نسبت های تعدیل شده بر اساس ریسک را بهبود دهند [\(چنگ و سان، ۲۰۲۴\)](#). از سوی دیگر، پژوهشگران به تدریج به اهمیت طراحی تابع پاداش حساس به ریسک توجه کرده اند. برخی مطالعات نشان داده اند که پیشینه سازی صرف بازده، در محیط های مالی پرنوسان می تواند عامل را به سمت رفتارهای پرریسک سوق دهد؛ از این رو، استفاده از معیارهای تعدیل شده بر اساس ریسک مانند نسبت شارپ، نسبت سورتینو و کنترل افت سرمایه، به یکی از محورهای مهم پژوهش های جدید تبدیل شده است [\(کورنالبو و همکاران، ۲۰۲۴\)](#).

در بازار رمزارزها، نوسانات شدید معمولاً ناشی از تغییرات نقدینگی، احساسات جمعی و شوک های اطلاعاتی است، نه صرفاً عوامل بنیادی اقتصادی [\(باهو و همکاران، ۲۰۲۴\)](#). بنابراین، مدل هایی که به صورت تطبیقی بر اساس شاخص های نوسان عمل می کنند، مزیت نسبی دارند. در این پژوهش، شاخص نوسان کوتاه مدت به صورت انحراف

1 Long Short-Term Memory

معیار متحرک ۵ روزه برای تشخیص دوره‌های پرنوسان به کار گرفته شده است؛ مشابه رویکرد [دو و همکاران \(۲۰۲۳\)](#) که در بازار آتی از آن استفاده کرده‌اند. مطالعات خارجی دیگری مانند [کوراتامدی و همکاران \(۲۰۲۱\)](#) و [لین و همکاران \(۲۰۲۳\)](#) نیز نشان داده‌اند که مدل‌های حساس به ریسک در محیط‌های مالی ناپایدار، پایداری عملکرد بلندمدت بیشتری دارند. از این رو، چارچوب اتخاذشده در این پژوهش، با تأکید بر کنترل ریسک و انطباق پویا با شرایط بازار رمزارز، می‌تواند بینشی نوین برای سرمایه‌گذاران ارائه دهد.

با وجود پیشرفت قابل توجه پژوهش‌های مبتنی بر یادگیری تقویتی عمیق در مدیریت پرتفوی، همچنان چند خلأ اساسی در ادبیات پژوهش مشاهده می‌شود. نخست، بخش عمده مطالعات پیشین بر استفاده از یک عامل تصمیم‌گیرنده ثابت متمرکز بوده‌اند؛ در حالی که بازار رمزارزها دارای رفتار غیرایستا، تغییرات مکرر وضعیت بازار و نوسانات شدید است و یک سیاست واحد نمی‌تواند در تمامی شرایط عملکرد بهینه و پایداری ارائه دهد. دوم، در بسیاری از پژوهش‌ها، تمرکز اصلی بر بیشینه‌سازی بازده بوده و معیارهای ریسک نزولی، افت سرمایه و پایداری عملکرد، نقش محدودی در طراحی تابع پاداش داشته‌اند؛ موضوعی که می‌تواند به شکل‌گیری رفتارهای پرریسک و ناپایدار منجر شود. سوم، اگرچه برخی مطالعات از معماری‌های یادگیری عمیق برای استخراج الگوهای زمانی استفاده کرده‌اند، اما تعداد محدودی از پژوهش‌ها توانسته‌اند به صورت همزمان وابستگی‌های زمانی بلندمدت، روابط بین‌داری و سازوکار تصمیم‌گیری تطبیقی را در یک چارچوب یکپارچه مدل‌سازی کنند. افزون بر این، در حوزه مدیریت پرتفوی رمزارزها، پژوهش‌های اندکی از سازوکار تغییر پویای عامل تصمیم‌گیرنده همراه با تابع پاداش حساس به ریسک بهره گرفته‌اند. بر این اساس، پژوهش حاضر تلاش می‌کند با ارائه یک چارچوب تطبیقی حساس به ریسک مبتنی بر یادگیری تقویتی عمیق، بخشی از خلأ موجود در ادبیات پژوهش را پوشش داده و تعادلی مؤثر میان بازده، کنترل ریسک و پایداری عملکرد در شرایط متغیر بازار رمزارزها ایجاد کند.

۳. روش شناسی پژوهش

در این پژوهش چارچوبی مبتنی بر یادگیری تقویتی عمیق و معماری ترنسفورمر برای مدیریت پرتفوی رمزارزها طراحی و ارزیابی شده است. هدف، بررسی اثربخشی دو رویکرد متفاوت است: نخست، روش عامل‌های منفرد یادگیری تقویتی که هر عامل به صورت مستقل پرتفوی را مدیریت می‌کند؛ دوم، چارچوب تطبیقی عامل حساس به ریسک که با استخراج وابستگی‌های بلندمدت زمانی و روابط میان‌داری‌ها با استفاده از ترنسفورمر، به طور پویا بهترین عامل را انتخاب و استراتژی بهینه را اجرا می‌کند. استفاده از ترنسفورمر امکان پردازش همزمان داده‌های تاریخی و شناسایی الگوهای پیچیده قیمتی را فراهم می‌آورد و چارچوب پژوهش علاوه بر بیشینه‌سازی بازده، کنترل ریسک و مدیریت افت سرمایه را نیز ممکن می‌سازد. این دو رویکرد، ضمن فراهم آوردن امکان مقایسه عینی و بازتولید نتایج، قابلیت تطبیق با تغییرات بازار و شرایط نوسانی را نیز دارند.

ترنسفورمر: ترنسفورمر که نخستین بار توسط [واسوانی و همکاران \(۲۰۱۷\)](#) معرفی شد، یک معماری یادگیری عمیق برای مدل‌سازی دنباله‌ای است. مؤلفه اصلی آن، مکانیزم توجه به خود است که به صورت پویا وزن عناصر ورودی را تنظیم کرده و وابستگی‌های بلندمدت را با محاسبات موازی استخراج می‌کند. این ساختار در ابتدا برای پردازش زبان طبیعی توسعه یافت، اما بعدها در مسائل یادگیری تقویتی عمیق از جمله پیش‌بینی بازارهای مالی نیز با موفقیت به کار گرفته شد. معماری ترنسفورمر دارای ساختار رمزگذار-رمزگشا است. در این ساختار، رمزگذار ورودی‌ها را به نمایش‌های متنی تبدیل می‌کند و رمزگشا خروجی‌ها را از این تعبیه‌ها تولید می‌کند. هر لایه در بخش رمزگذار و رمزگشا شامل اجزای زیر است:

توجه چندسر^۱: محاسبه همزمان وابستگی میان عناصر ورودی.

شبکه پیش‌خور^۲: اعمال تبدیلات غیرخطی برای استخراج ویژگی‌ها.

اتصال باقی‌مانده و نرمال‌سازی لایه^۳: بهبود جریان گرادیان و پایداری مدل.

کدگذاری موقعیتی^۴: افزودن اطلاعات ترتیب دنباله، چون توجه به خود ذاتاً آگاهی مکانی ندارد.

در مکانیزم توجه به خود^۵، وابستگی میان توکن‌ها از طریق تبدیل‌های قابل یادگیری مدل می‌شود:

$$Q = XW_Q, K = XW_K, V = XW_V \quad (۱)$$

که در آن Q, K, V به ترتیب ماتریس‌های پرسش، کلید و مقدار هستند. خروجی مکانیزم توجه به صورت زیر محاسبه می‌شود:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (۲)$$

توجه چند سر: برای درک اطلاعات از چند زیرفضای نمایشی، h سر توجه به خود به طور موازی اجرا می‌شوند:

$$MultiHead(Q, K, V) = Concat(Z_1, Z_2, \dots, Z_h)W_o \quad (۳)$$

که در آن Z_i خروجی سر i ام است.

شبکه پیش‌خور: هر لایه در ترنسفورمر شامل یک شبکه پیش‌خور موضعی است که به صورت زیر تعریف می‌شود:

$$FFN(x) = ReLU(W_1x + b_1)W_2 + b_2 \quad (۴)$$

این بخش باعث افزودن غیرخطی بودن و افزایش قدرت بازنمایی ویژگی‌ها می‌شود.

کدگذاری موقعیتی: از آنجا که مکانیزم توجه به خود به صورت ذاتی به ترتیب داده‌ها حساس نیست، کدگذاری سینوسی موقعیت به ورودی‌ها افزوده می‌شود:

$$PE_{(pos,i)} = \begin{cases} \sin(pos \cdot \omega_k), & \text{اگر } i = 2k \\ \cos(pos \cdot \omega_k), & \text{اگر } i = 2k + 1 \end{cases} \quad (۵)$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

این تابع تضمین می‌کند که هر موقعیت در دنباله، کدگذاری منحصر به فردی داشته باشد و ترتیب توالی در مدل حفظ شود.

مدل‌های یادگیری تقویتی عمیق

شبکه عمیق کیو DQN: یادگیری Q سنتی را با استفاده از شبکه‌های عصبی عمیق برای تقریب تابع Q-value

بهبود می‌بخشد. این مدل فرایند آموزش را با تکنیک‌های زیر پایدار می‌کند:

- تجربه تکراری: ذخیره تجربه‌های گذشته برای آموزش دسته‌ای و افزایش کارایی یادگیری.
- شبکه هدف: نگهداری یک شبکه Q جداگانه برای کاهش نوسان در طول آموزش.

DQN با کمیته‌سازی تابع خطای زیر، Q-function را بهینه می‌کند:

1 Multi-Head Attention

2 Feed-Forward Network

3 Residual Connection and Layer Normalization

4 Positional Encoding

5 Self-Attention Mechanism

$$L = \mathbb{E}[r + \gamma \max_a Q(s', a') - Q(s, a)]^2 \quad (۶)$$

۳.۲.۲. عامل-منتقد با مزیت A2C: ترکیبی از روش‌های مبتنی بر سیاست و مبتنی بر ارزش است و شامل اجزای زیر می‌باشد:

- عامل: سیاست $\pi(a | s)$ را یاد می‌گیرد.

- منتقد: تابع ارزش حالت $V(s)$ را برآورد می‌کند.

تابع مزیت، پایداری آموزش را بهبود می‌دهد:

$$A(s, a) = Q(s, a) - V(s) \quad (۷)$$

به‌روزرسانی سیاست بر اساس رابطه زیر انجام می‌شود:

$$L_a = -\log \pi(a | s) A(s, a) \quad (۸)$$

بهینه‌سازی سیاست مجاور PPO: روش‌های گرادینان سیاست را با استفاده از یک تابع هدف بریده‌شده بهبود می‌دهد:

$$L_{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)] \quad (۹)$$

که در آن نسبت احتمال بین سیاست جدید و قدیم است. بریده‌سازی^۱ موجب به‌روزرسانی پایدار می‌شود. PPO به دلیل سادگی، کارایی و عملکرد پایدار در مسائل با ابعاد بالا به‌طور گسترده مورد استفاده قرار می‌گیرد. آماده‌سازی داده‌ها: در این پژوهش، داده‌ها برای بازه پنج سال اخیر شامل قیمت‌های روزانه شامل قیمت باز شدن، بالاترین قیمت، پایین‌ترین قیمت، قیمت پایانی معاملات پنج روزه منتخب از پایگاه داده یاهو فاینانس استخراج شده است. علاوه بر داده‌های قیمت چندین شاخص تکنیکال شامل میانگین متحرک، شاخص قدرت نسبی، میانگین متحرک همگرایی و اگرایی، شاخص میانگین محدوده واقعی، انحراف معیار متحرک نیز استخراج شده است که در معاملات کمی و مدل‌سازی مالی کاربرد دارند. برای تقسیم‌بندی داده‌ها، ۷۰٪ داده‌های اولیه به‌عنوان مجموعه آموزش و ۳۰٪ انتهایی به‌عنوان مجموعه آزمون در نظر گرفته می‌شوند.

عامل‌های منفرد یادگیری تقویتی عمیق - روش اول: در این روش، محیط معاملاتی با یک عامل یادگیری تقویتی عمیق در هر زمان ساخته می‌شود (A2C, PPO, DQN). محیط به‌عنوان یک فرایند تصمیم‌گیری مارکوف تعریف شده است که شامل حالت، عمل، تابع پاداش و ضریب تخفیف می‌باشد.

ساخت پرتفوی: وظیفه مدیریت پرتفوی به‌صورت یادگیری تقویتی عمیق فرموله شده و با تاپل (S, A, P, R, γ) نمایش داده می‌شود:

حالت s_t در هر گام زمانی شامل اطلاعات بازار و وزن‌های پرتفوی پنج دارایی است. این اطلاعات در یک بردار ویژگی نمایش داده می‌شوند. یک شبکه مبتنی بر ترنسفورمر این بردار را پردازش می‌کند تا وابستگی‌های زمانی و روابط بین دارایی‌ها که برای تصمیمات توزیع مجدد پرتفوی ضروری هستند، استخراج شود.

عمل A ، وزن‌های پیوسته پرتفوی هستند که عامل به هر دارایی اختصاص می‌دهد. وزن‌ها بین ۰ و ۱ محدود شده و جمع وزن پنج دارایی برابر ۱ است. تخصیص هر دارایی در زمان t ، a_t ، بر اساس سیاست انتخاب‌شده به‌روزرسانی می‌شود که بسته به الگوریتم مورد استفاده متفاوت است.

تابع انتقال $P(s_{t+1} | s_t, a_t)$ احتمال حرکت از حالت فعلی s_t به حالت بعدی s_{t+1} را با توجه به عمل a_t مشخص می‌کند. در این مطالعه، انتقالات به‌صورت صریح مدل‌سازی نمی‌شوند؛ بلکه توسط داده‌های تاریخی بازار هدایت می‌شوند و محیط به‌صورت داده‌محور بدون فرض فرایند تصادفی شناخته‌شده تکامل می‌یابد.

^۱ Clipping

پاداش r_t برای تعادل بین بازده و ریسک طراحی شده و شامل بازده پرتفوی و نوسان آن است. بازده پرتفوی در زمان t ، R_t به صورت مجموع وزنی بازده دارایی‌ها محاسبه می‌شود:

$$R_t = \sum_{i=1}^N w_{i,t-1} \cdot \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}} \quad (10)$$

که در آن $w_{i,t-1}$ وزن اختصاص یافته به دارایی i در زمان $t-1$ و $P_{i,t}$ قیمت دارایی i در زمان t است. این فرمول تغییرات روزانه ارزش پرتفوی ناشی از تغییرات قیمتی دارایی‌ها و تخصیص وزن‌ها را ثبت می‌کند. تابع پاداش به صورت زیر تعریف شده است:

$$r_t = \log(1 + R_t) + \alpha \cdot \text{Calmar Ratio} - \beta \cdot \sigma_t \quad (11)$$

که جمله اول، تابع مطلوبیت لگاریتمی بازده است و ترجیح عامل برای رشد نسبی را نشان می‌دهد و بازده منفی بزرگ را جریمه می‌کند.

نسبت کالمار در زمان t به صورت زیر محاسبه می‌شود:

$$\text{Calmar Ratio} = \frac{\log(1 + R_t)}{\text{Max Drawdown} + \epsilon} \quad (12)$$

که صورت کسر، تابع مطلوبیت لگاریتمی بازده را نشان می‌دهد (همانند جمله اول تابع پاداش) و مخرج، حداکثر افت سرمایه تا زمان t است که با یک عدد کوچک ϵ منظم‌سازی شده تا از تقسیم بر صفر جلوگیری شود. این نسبت استراتژی‌هایی که بازده تولید می‌کنند ولی باعث افت سرمایه زیاد می‌شوند را تنبیه می‌کند.

انحراف معیار σ_t با استفاده از نوسان متحرک بازده دارایی‌ها محاسبه می‌شود. در تابع پاداش، $\alpha = 0,005$ به عنوان وزن مؤلفه‌ای در نظر گرفته شده است که به توازن میان بازده و ریسک کمک می‌کند، در حالی که $\beta = 0,9$ شدت جریمه نوسان را تعیین می‌کند و نقش آن در محدودسازی رفتارهای پرریسک و ناپایدار عامل برجسته است. این دو پارامتر بر اساس تنظیم تجربی و با هدف دستیابی به تعادل میان کارایی بازدهی و پایداری سیاست آموخته شده انتخاب شده‌اند؛ به گونه‌ای که مقادیر کمتر از این سطح، کنترل ریسک را تضعیف می‌کرد و مقادیر بالاتر، به محافظه‌کاری بیش از حد و کاهش بهره‌برداری از فرصت‌های سودآور منجر می‌شد. همچنین، ضریب تنزیل $\gamma = 0,99$ نشان‌دهنده افق تصمیم‌گیری بلندمدت عامل است، به این معنا که پاداش‌های آینده با تخفیفی بسیار اندک اما مؤثر در فرآیند تصمیم‌گیری لحاظ می‌شوند. در مقابل، مقادیر کوچک γ بیانگر تمرکز بیشتر عامل بر پاداش‌های فوری و کوتاه‌مدت هستند و موجب می‌شوند سیاست آموخته شده آینده‌نگری کمتری داشته باشد. از این رو، انتخاب γ نزدیک به ۱ برای این مسئله مناسب‌تر است، زیرا مدل را به سمت توجه به پیامدهای تجمعی و بلندمدت تصمیم‌ها هدایت می‌کند.

الگوریتم اول: آموزش و اجرای یک عامل یادگیری تقویتی عمیق واحد

ورودی: محیط E ، عامل π_θ ، تعداد اپیزودها N

۱. برای $1 \leftarrow t$ تا N انجام شود:

۲. حالت اولیه s_0 را از محیط E مقداردهی کنید؛

۳. برای هر گام زمانی t انجام شود:

۴. انتخاب عمل $a_t = \pi_\theta(s_t)$

۵. اجرای a_t و مشاهده r_t, s_{t+1}

۶. ذخیره انتقال (s_t, a_t, r_t, s_{t+1}) :

۷. به‌روزرسانی π_θ با استفاده از تجربیات گذشته؛

۸. $s_t \leftarrow s_{t+1}$

۹. بازگرداندن عامل آموزش‌دیده π_θ

عامل تطبیقی حساس به ریسک - روش دوم: چارچوب عامل تطبیقی حساس به ریسک در این پژوهش به‌عنوان یک سازوکار پویا برای انتخاب بهترین عامل یادگیری تقویتی در هر گام زمانی طراحی شده است. در این رویکرد، سه عامل پایه شامل PPO، A2C و DQN که در روش اول به‌طور مستقل آموزش یافته‌اند، در قالب یک مجموعه تصمیم‌گیری یکپارچه به‌کار گرفته می‌شوند. هدف اصلی این چارچوب، استفاده هم‌زمان از مزایای هر عامل در شرایط مختلف بازار و کاهش نقاط ضعف روش‌های ایستا است. به‌جای تکیه بر یک سیاست ثابت، الگوی پیشنهادی عملکرد واقعی عوامل را در یک پنجره زمانی ۵ روزه پایش کرده و عاملی را انتخاب می‌کند که رفتار بهینه‌تری از نظر بازده لگاریتمی و ریسک نزولی نشان می‌دهد. معیار عملکرد اخیر هر عامل، کاملاً بر اساس تابع پاداش تعریف شده در روش اول محاسبه می‌شود. برای هر عامل π_i و در هر گام زمانی t ، پاداش محیط مطابق رابطه $r_t = \ln(1 + R_t) + \alpha \cdot \text{CalmarRatio}_t - \beta \cdot \sigma_t$ و پارامترها دقیقاً مقادیر انتخاب شده در روش اول هستند. بازده پرتفوی R_t همانند روش اول برابر است با مجموع وزنی بازده دارایی‌ها، و σ_t انحراف معیار متحرک بازده‌هاست. بنابراین، «پاداش اخیر» هر عامل در پنجره ۵ روزه به‌صورت میانگین پاداش‌های آن پنجره محاسبه می‌شود:

$$\bar{R}_t^i = \frac{1}{5} \sum_{k=t-4}^t r_k^i \quad (۱۳)$$

برای انتخاب عامل برتر، چارچوب از یک معیار تطبیقی مبتنی بر همان معیارهای ریسک‌سنجی موجود در تابع پاداش استفاده می‌کند. هر عامل بر اساس مقدار $\bar{R}_t^i$ در پنجره ۵ روزه ارزیابی می‌شود و عامل منتخب طبق رابطه $\pi_t^* = \arg \max_{\pi_i} (\bar{R}_t^i)$ انتخاب می‌گردد. از آنجا که تابع پاداش شامل سه مؤلفه بازده لگاریتمی، نسبت کالمار و جریمه نوسان است، انتخاب بر اساس $\bar{R}_t^i$ به‌صورت درونی رفتارهای پرریسک را تنبیه و سیاست‌هایی با ریسک نزولی کمتر را ترجیح می‌دهد. به بیان دیگر، سازوکار انتخاب عامل نیازی به تعریف معیار مستقل ندارد؛ زیرا تابع پاداش پایه خود شامل ارزیابی حساس به ریسک است. پس از انتخاب عامل برتر در گام t ، تصمیم نهایی تخصیص پرتفوی از سیاست عامل منتخب استخراج می‌شود:

$$a_t = \pi_t^*(s_t), \quad (۱۴)$$

و عمل انتخاب شده در محیط اجرا شده و پاداش جدید محاسبه می‌گردد. سپس عامل منتخب مطابق با تجربه تازه‌شده به‌روزرسانی شده و سایر عوامل نیز تجربه جدید را در حافظه خود ذخیره می‌کنند. این چرخه انتخاب-اجرا-به‌روزرسانی موجب می‌شود سیستم بتواند به‌صورت پویا میان عوامل جابه‌جا شده و در برابر تغییرات ناگهانی بازار سازگار باقی بماند.

یکی از اجزای مهم چارچوب که نقش حفاظتی در بازارهای پرنوسان ایفا می‌کند، شرط فعالسازی نوسان است که به صورت مستقل از عامل‌ها عمل کرده و قبل از انتخاب عامل فعال می‌شود. این شرط فعالسازی زمانی فعال می‌شود که نوسان کوتاه مدت ۱۰ روزه از ۱,۵ برابر نوسان بلندمدت ۹۰ روزه فراتر رود:

$$\text{trigger}(k) = 1\{\sigma_{10,k} > 1,5 \cdot \sigma_{90,k}\} \quad (۱۵)$$

در صورت فعال شدن شرط آستانه ای، سهم دارایی نقدی پرتفو (تتر) حداقل به ۳۰٪ افزایش یافته و از ورود پرتفوی به موقعیت‌های پرخطر جلوگیری می‌شود. در نسخه پیاده‌سازی شده، شرط فعالسازی تنها وزن نقدی را افزایش داده و در انتخاب عامل دخالت مستقیم ندارد؛ اما همین مداخله کافی است تا ریسک نزولی پرتفو در دوره‌های بحران به طور معناداری کاهش یابد.

در نهایت، ترکیب ارزیابی ۵ روزه پاداش، انتخاب عامل برتر و شرط فعالسازی نوسان، چارچوب پیشنهادی را به سیستمی هوشمند و تطبیقی تبدیل می‌کند که می‌تواند به صورت مؤثر در محیط‌های غیرایستا و پرنوسان بازار رمازار عمل کند. این سازوکار بدون اعمال محدودیت‌های پیچیده اضافی و تنها با استفاده از ساختاری ساده اما مبتنی بر اطلاعات حساس به ریسک، توانسته است عملکرد پایدارتری نسبت به عامل‌های منفرد ارائه دهد. نتایج تجربی نیز نشان می‌دهد که این رویکرد قادر است ضمن افزایش بازده تعدیل شده بر اساس ریسک، از افت سرمایه‌های شدید جلوگیری کرده و پایداری عملکرد را در بلندمدت بهبود بخشد.

الگوریتم دوم: عامل تطبیقی حساس به ریسک : ورودی: عوامل پایه $B = \{\pi_1, \pi_2, \dots, \pi_k\}$ ، پنجره متحرک w ، محیط E

۱. برای هر گام زمانی t انجام شود:

۲. برای هر عامل $\pi_i \in B$ انجام شود:

۳. ارزیابی پاداش اخیر R_t^i در طول پنجره w :

۴. انتخاب عامل برتر $\pi^* = \arg \max \pi_i(R_t^i)$:

۵. انتخاب عمل $a_t = \pi^*(s_t)$:

۶. اجرای a_t در محیط E ، مشاهده s_{t+1}, r_t :

۷. به روزرسانی عامل منتخب π^* با تجربه جدید:

۸. $s_t \leftarrow s_{t+1}$:

۹. بازگرداندن سیاست تطبیقی از طریق سوئیچ.

در مورد مدل عامل تطبیقی حساس به ریسک، مکانیزم انتخاب پویا عوامل از این معیارهای تعدیل شده بر اساس ریسک برای راهنمایی انتخاب عامل در زمان واقعی استفاده می‌کند. بدین ترتیب، مدل استراتژی‌هایی را ترجیح می‌دهد که عملکردی پایدارتر و مقاوم‌تر تحت شرایط نامطمئن ارائه می‌دهند. این ادغام دوگانه ویژگی‌های آگاه به نوسان و معیارهای ارزیابی تعدیل شده بر ریسک تضمین می‌کند که چارچوب پیشنهادی نه تنها به دنبال بازده بالا است، بلکه ریسک نزولی را نیز مدیریت می‌کند.

معیارهای ارزیابی: پس از آموزش و آزمون، مدل‌ها بر اساس بازده، ریسک و عملکرد کلی پرتفوی ارزیابی شدند. بازده تجمعی: میزان رشد کل پرتفوی در طول زمان را اندازه‌گیری می‌کند. تعریف آن به صورت زیر است:

$$R_c = \prod_{t=1}^T (1 + R_t) - 1 \quad (۱۶)$$

که در آن R_t بازده پرتفوی در زمان t و T تعداد کل روزهای معاملاتی نسبت شارپ: بازده تعدیل شده بر اساس ریسک را با تقسیم میانگین بازده روزانه بر نوسان بازده ارزیابی می‌کند.

$$\text{Sharp Ratio} = \frac{E[R - R_f]}{\sigma_R} \sqrt{T} \quad (۱۷)$$

که در آن R_f نرخ بدون ریسک و σ_R انحراف معیار بازده است. در این مطالعه، نرخ بدون ریسک ثابت صفر در نظر گرفته شده است که با روش استاندارد در مدل‌سازی مالی مبتنی بر یادگیری تقویتی عمیق همخوانی دارد.

نسبت سورتینو: مشابه نسبت شارپ است اما بر ریسک نزولی تمرکز دارد.

$$\text{Sortino Ratio} = \frac{E[R - R_f]}{\sigma_{\text{down}}} \sqrt{T} \quad (۱۸)$$

که σ_{down} انحراف معیار بازده‌های منفی است.

حداکثر افت سرمایه: بزرگترین کاهش از اوج تا کف ارزش پرتفوی را اندازه‌گیری می‌کند.

$$MDD = \max_t \frac{Peak_t - PortfolioValue_t}{Peak_t} \quad (۱۹)$$

که در آن $Peak_t = \max_{i \leq t} PortfolioValue_i$ بیشینه تجمعی تا زمان t است.

نسبت کالمار: بازده نسبت به حداکثر افت سرمایه را ارزیابی می‌کند:

$$\text{Calmar ratio} = \frac{\text{Annualized Return}}{MDD} \quad (۲۰)$$

بازده سالیانه: رشد استاندارد شده پرتفوی در طول زمان را اندازه‌گیری می‌کند تا مقایسه بین دوره‌های مختلف ممکن شود:

$$R_{\text{ann}} = \left(\frac{V_{\text{end}}}{V_{\text{start}}} \right)^{\frac{252}{N}} - 1 \quad (۲۱)$$

که در آن V ارزش پرتفوی، V_{start} و V_{end} مقادیر پرتفوی در آغاز و پایان دوره آزمایشی هستند، ۲۵۲ میانگین تعداد روزهای معاملاتی در یک سال و N تعداد روزهای معاملاتی واقعی در دوره آزمایشی است.

۴. تحلیل داده‌ها و یافته‌ها

در این پژوهش با هدف ارائه و ارزیابی یک مدل تطبیقی حساس به ریسک برای مدیریت پرتفو در بازار رمزارزها، داده‌های پنج رمزارز بیت‌کوین، اتریوم، سولانا، ریپل و تتر در دوره‌ای پنج‌ساله در بازه ۲۴ نوامبر ۲۰۲۰ تا ۲۴ نوامبر ۲۰۲۵ از پایگاه یاهو فایننس استخراج شد. در مرحله پیش‌پردازش، داده‌ها از نظر مقادیر گمشده بررسی و با استفاده از تکمیل رو به جلو و تکمیل رو به عقب اصلاح شدند تا پیوستگی سری‌های زمانی حفظ شود. سپس سری‌های داده‌ای از نظر زمانی هم‌تراز شدند، ویژگی‌های قیمتی و تکنیکال مورد نیاز محاسبه و نرمال‌سازی شدند، و در قالب نمایش ماتریسی برای ورود به مدل سازمان‌دهی شدند. در نهایت، مجموعه داده به‌صورت زمانی و با نسبت ۷۰ درصد برای آموزش و ۳۰ درصد برای آزمون تقسیم شد تا ارزیابی مدل در شرایطی بدون نشت اطلاعات انجام گیرد. با توجه به نوسانات شدید و ساختار غیرخطی بازار رمزارزها، مدل‌های یادگیری تقویتی ممکن است در دوره‌های متفاوت بازار رفتارهای متفاوت و حتی ناپایدار نشان دهند. از این رو دوره آزمون در دو بخش مجزا تحلیل شد:

بازار بدون روند: برای سنجش پایداری و کنترل ریسک در شرایط نوسان محدود در بازه ۲۰ نوامبر ۲۰۲۴ تا ۲۴ نوامبر ۲۰۲۵

بازار صعودی: برای بررسی توان مدل‌ها در بهره‌گیری از روندهای مثبت در بازه ۲۴ می ۲۰۲۴ تا ۲۴ آوریل ۲۰۲۵ تمام مدل‌ها در ژوپیتر نوت بوک پیاده‌سازی شده و در نرم افزار آناکوندا با استفاده از یک کارت گرافیکی مدل انویدیا با ۲ گیگابایت حافظه روی یک سیستم با پردازنده core-i7 اینتل نسل ۱۲ و ۱۶ گیگابایت حافظه آموزش داده شدند. در این پژوهش، دو رویکرد برای مدیریت پرتفو مورد مقایسه قرار گرفت: روش اول: مدل‌های مستقل یادگیری تقویتی شامل PPO، A2C و DQN روش دوم: مدل تطبیقی حساس به ریسک که به صورت پویا و بر اساس وضعیت ریسک بازار، بین عامل‌های مختلف تصمیم‌گیرنده جابه‌جا می‌شود و ساختار رفتاری آن بر پایه حساسیت به نوسان و زیان طراحی شده است. عملکرد مدل‌ها با شش معیار کلیدی شامل بازده سالانه‌سازی شده، بازده تجمعی، نسبت شارپ، نسبت سورتینو، نسبت کالمار و حداکثر افت سرمایه ارزیابی شد.

عملکرد مدل‌ها در بازار بدون روند (روش اول: مدل‌های مستقل یادگیری تقویتی): ارزیابی عملکرد مدل‌های یادگیری تقویتی پایه در دوره بازار بدون روند و بازار صعودی نشان‌دهنده تفاوت‌های معنادار در توانایی آن‌ها برای خلق بازده تعدیل شده بر اساس ریسک و حفظ پایداری پرتفو است. جدول ۱ نتایج دوره بازار بدون روند را نشان می‌دهد.

جدول ۱. عملکرد مدل‌های مستقل در بازار بدون روند (یافته‌های تحقیق)

Table 1. Performance of Independent Models in Sideways Markets

مدل	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتینو	نسبت کالمار	حداکثر افت سرمایه %
B&H	106,212	4,19	0,314	0,453	0,106	-39,38
PPO	147,369	30,23	0,745	1,186	0,694	-43,57
A2C	191,965	55,92	1,023	1,650	1,266	-44,16
DQN	137,570	24,26	0,693	1,064	0,616	-39,38

مدل A2C با ثبت بالاترین بازده سالیانه (۵۵,۹۲٪) عملکردی چشمگیر از نظر سودآوری خام داشته است، اما بررسی دقیق‌تر شاخص‌ها نشان می‌دهد که این بازده قیمت بسیار سنگینی در حوزه ریسک داشته است.

- افت سرمایه بسیار بالا (۴۴,۱۶٪) نشان می‌دهد که سیاست عامل A2C در مواجهه با نوسانات محدود بازار بدون روند، رفتار پرریسک و تهاجمی داشته است. همچنین نسبت شارپ ۱,۰۰۲ و نسبت کالمر ۱,۲۶۶ اگرچه قابل قبول‌اند، اما نسبت به میزان ریسکی که تحمل شده، بازده اضافی ایجاد نمی‌کنند. ساختار به‌روزرسانی A2C که به میانگین‌گیری از گرادینان‌ها وابسته است، باعث می‌شود در بازارهای کم‌روند، رفتار بیش‌برازش نسبت به نوسانات کوتاه‌مدت داشته باشد. A2C توان «خلق بازده» دارد اما فاقد هرگونه کنترل ریسک است؛ بنابراین عملکرد آن پایدار و قابل اتکا نیست و این رفتار پرنوسان، آن را برای مدیریت پرتفو در بازارهای کم‌سیگنال نامناسب می‌کند.

مدل PPO نسبت به A2C رفتار باثبات‌تری از خود نشان داده است.

- بازده سالیانه ۳۰,۲۳٪ اگرچه کمتر از A2C است، اما نسبت شارپ ۰,۷۴۵ و سورتینو ۱,۱۸۶ نشان می‌دهد که بازده با کیفیت بهتری تولید شده است. همچنین افت سرمایه ۴۳,۵۷٪ همچنان بالا است و نشان می‌دهد که PPO در بازار بدون روند نمی‌تواند از وارد شدن به موقعیت‌های پرریسک خودداری کند. به دلیل استفاده از سازوکار هدف محدود، الگوریتم PPO در تغییر سیاست محتاط عمل می‌کند و در محیط‌هایی با نوسان کم، گاهی بیش از حد منفعل رفتار می‌کند.

PPO نسبت به A2C پایدارتر است اما همچنان با ماهیت بازار بدون روند سازگار نیست. این مدل به دلیل ماهیت محافظه‌کارانه‌اش نمی‌تواند از نوسانات محدود بهره‌برداری کند و در عین حال، همچنان ریسک‌های سنگینی را متحمل می‌شود.

مدل DQN از میان سه مدل، ضعیف‌ترین عملکرد نسبی را داشته است:

- بازده سالیانه $24,26\%$ تقریباً نزدیک به خرید و نگهداری است. نسبت شارپ $0,693$ و سورتینو $10,064$ نشان می‌دهد که بازده تولیدشده کیفیت پایینی دارد. ساختار گسسته DQN با فضای تصمیم‌گیری پیوسته تخصیص وزن‌ها سازگار نیست و این باعث می‌شود در بازارهای بدون روند رفتار نوسانی، غیرسازگار و کم‌بازده داشته باشد.

DQN در محیط‌های پیوسته و کم‌نوسان عملکرد ضعیفی دارد و توانایی تطبیق با تغییرات کوچک بازار را ندارد. این مدل از نظر عملی نمی‌تواند جایگزین مناسبی برای مدیریت پرتفو باشد.

نتایج نشان می‌دهد که هر سه مدل مستقل یادگیری تقویتی عمیق دارای بازده بالا اما ریسک بسیار زیاد هستند. بازار بدون روند نیازمند مدلی است که:

- از نوسانات خرد بهینه استفاده کند،
- از ورود به موقعیت‌های پرریسک اجتناب کند،
- رفتار تطبیقی در مواجهه با تغییرات لحظه‌ای داشته باشد.

هیچ‌یک از مدل‌های روش اول به این سه ویژگی دست نیافته‌اند.

این ضعف بنیادی، دقیقاً همان چیزی است که ضرورت مدل تطبیقی حساس‌به‌ریسک را توجیه می‌کند؛ مدلی که با شرط فعالسازی حساس‌به‌ریسک و ساختار چندعاملی، نه‌تنها بازده بیشتری ثبت کرده، بلکه رفتار آن در بازار بدون روند بسیار پایدارتر و کنترل‌شده‌تر است.

۴.۱.۲. روش دوم: مدل تطبیقی حساس‌به‌ریسک

برای درک کارایی مدل تطبیقی حساس‌به‌ریسک، عملکرد آن در مقابل مدل‌های مستقل یادگیری تقویتی عمیق در بازار بدون روند بررسی شده است. جدول زیر خلاصه‌ی مقایسه را نشان می‌دهد:

جدول ۲. مقایسه عملکرد روش اول و مدل تطبیقی حساس‌به‌ریسک در بازار بدون روند (یافته‌های تحقیق)

Table 2. Performance Comparison: Method 1 vs. Risk-Sensitive Adaptive Model in Sideways Markets

مدل	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتینو	نسبت کالمار	حداکثر افت سرمایه %
B&H	106,212	4,19	0,314	0,453	0,106	-39,38
PPO	147,369	30,23	0,745	1,186	0,694	-43,57
A2C	191,965	55,92	1,023	1,650	1,266	-44,16
DQN	137,570	24,26	0,693	1,064	0,616	-39,38
پیشنهادی	213,727	67,52	1,132	1,808	1,587	-42,55

تحلیل مقایسه‌ای

بازده سالیانه و ارزش نهایی پرتفو

مدل تطبیقی حساس‌به‌ریسک بیشترین بازده سالیانه ($67,52\%$) و بالاترین ارزش نهایی پرتفو ($213,727$ دلار) را ثبت کرده است، در حالی که بهترین مدل مستقل، A2C، بازده سالیانه $55,92\%$ داشت. این افزایش بازده نشان‌دهنده‌ی توان مدل تطبیقی در استفاده بهینه از نوسانات کوچک بازار بدون روند و انتخاب استراتژی‌های مناسب به‌صورت پویا است.

کنترل ریسک و پایداری

با مقایسه نسبت‌های شارپ و سورتینو:

- مدل تطبیقی نسبت شارپ ۱،۱۳۲ و سورتینو ۱،۸۰۸ دارد، که بالاتر از تمام مدل‌های مستقل است.
- نسبت‌های بالاتر به معنای بازده تعدیل‌شده بهتر بر اساس ریسک و پایداری بیشتر در برابر نوسانات بازار است.
- در حالی که افت سرمایه حداکثری مدل تطبیقی -۴۲،۵۵٪ ثبت شده است، این مقدار کمتر یا مشابه مدل‌های مستقل است و نشان می‌دهد که مدل تطبیقی توانسته است رفتار محافظه‌کارانه در زمان نوسان شدید را با بهره‌برداری از فرصت‌های سودآور ترکیب کند.

رفتار تطبیقی و سوئیچ پویا

ویژگی برجسته مدل تطبیقی حساس به ریسک، توانایی سوئیچ بین عامل‌های تصمیم‌گیرنده براساس شرط فعالسازی حساس به ریسک است:

- در دوره‌هایی که بازار نوسان کم یا سیگنال ضعیف دارد، مدل به عامل‌های محافظه‌کار روی می‌آورد و از ورود به موقعیت‌های پرریسک اجتناب می‌کند.
- در دوره‌هایی که نوسانات فرصت‌ساز ظاهر می‌شود، به عامل‌های تهاجمی سوئیچ می‌کند تا سود بیشتری کسب شود.
- این رفتار انعطاف‌پذیری و تطبیق سریع با شرایط بازار بدون روند را فراهم می‌کند، در حالی که مدل‌های مستقل فاقد چنین مکانیسمی هستند.

تحلیل ترکیبی بازده و ریسک: اگر نمودار تجمعی بازده مدل‌ها را در بازار بدون روند بررسی کنیم.

- مدل‌های مستقل نوسانات کوتاه‌مدت شدید دارند، جهش‌های سودآور موقت و کاهش‌های تند را تجربه می‌کنند.
- مدل تطبیقی حساس به ریسک، علی‌رغم نوسانات کوتاه، خط تجمعی بازده بالای صفر را حفظ می‌کند و همواره کنترل منطقی بر افت سرمایه دارد.

این امر نشان‌دهنده‌ی ترکیب بهینه بازده و ریسک است که در محیط‌های کم‌روند و نوسانی بازارهای رمزارز بسیار حیاتی است.

- مدل تطبیقی حساس به ریسک در بازار بدون روند، هم بازدهی بالاتر و هم پایداری بیشتری نسبت به تمام مدل‌های مستقل ارائه کرده است. این مدل با استفاده از سوئیچ پویا و حساسیت به ریسک توانسته است نوسانات محدود بازار را به فرصت تبدیل کند، در حالی که مدل‌های مستقل در مواجهه با همان نوسانات رفتار غیرقابل پیش‌بینی و ریسک‌پذیری نامطلوب نشان داده‌اند. تحلیل شاخص‌های تعدیل‌شده بر اساس ریسک و حداکثر افت سرمایه تأیید می‌کند که مدل تطبیقی می‌تواند بهترین تعادل بین رشد پرتفو و کنترل ریسک را ارائه دهد، به‌ویژه در شرایطی که بازار روند مشخصی ندارد.

شکل ۱ نمایی از مقایسه عملکرد مدل های خرید و نگهداری، مدل های یادگیری تقویتی مستقل و مدل پیشنهادی تطبیقی حساس به ریسک (بازار بدون روند) را به نمایش گذاشته است.



شکل ۱. مقایسه عملکرد مدل های خرید و نگهداری، مدل های یادگیری تقویتی مستقل و مدل پیشنهادی (بازار بدون روند)

Figure 1. Performance Comparison of Buy-and-Hold Models, Independent Reinforcement Learning Models, and the Proposed Model in a Sideways Market

عملکرد مدل ها در بازار صعودی: بازار صعودی معمولاً با روند مثبت پایدار و افزایش ارزش دارایی ها مشخص می شود. در چنین محیطی، الگوریتم های یادگیری تقویتی باید توانایی شناسایی روند و بهره برداری از آن را داشته باشند، همزمان که کنترل ریسک رعایت شود. در ادامه، عملکرد مدل ها به تفکیک روش ها بررسی می شود.

۴.۲.۱. روش اول: مدل های مستقل یادگیری تقویتی

جدول ۳ نتایج دوره صعودی مدل های مستقل یادگیری تقویتی را نشان می دهد.

جدول ۳. عملکرد مدل های مستقل در بازار صعودی (یافته های تحقیق)

Table 3. Performance of Independent Models in a Bull Market

مدل	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتنو	نسبت کالمار	حداکثر افت سرمایه %
B&H	135,279	25,95	0,742	1,059	0,659	-39,38
PPO	223,805	85,00	1,553	2,577	2,358	-36,04
A2C	158,215	41,96	0,933	1,446	0,936	-44,83
DQN	136,837	27,06	0,761	1,084	0,687	-39,38

بازده و ارزش پرتفو

- PPO بیشترین ارزش نهایی پرتفو (۲۲۳,۸۰۵ دلار) و بالاترین بازده سالیانه (۸۵٪) را در میان مدل های مستقل دارد.

- A2C عملکرد متوسط با بازده ۴۱,۹۶٪ داشته و DQN تقریباً مشابه خرید و نگهداری عمل کرده است.

- مشاهده می شود که مدل های مستقل می توانند از روند صعودی سود ببرند، اما میزان بازده با پایداری و کنترل ریسک محدود همراه است.

کنترل ریسک و نسبت‌های تعدیل‌شده

- نسبت شارپ PPO (۱,۵۵۳) و سورتینو (۲,۵۷۷) بالاتر از سایر مدل‌هاست، نشان‌دهنده‌ی توان نسبی در تعدیل ریسک است.
 - با این حال، افت سرمایه حداکثر (-۳۶,۰۴٪) همچنان قابل توجه است و بیانگر ریسک‌های کوتاه‌مدت و نوسانات شدید پرتفو است.
 - A2C با افت سرمایه ۴۴,۸۳٪ و نسبت شارپ ۰,۹۳۳، بازده نسبی خوبی دارد اما کنترل ریسک ضعیف‌تر نسبت به PPO نشان می‌دهد.
- رفتار نوسانی و سازگاری با بازار
- DQN در بهره‌برداری از روند صعودی عملکرد محدود و نوسانی دارد.
 - خرید و نگهداری ساده، بازده پایدار اما کم ارائه می‌کند و هیچ مکانیسم تطبیقی برای بهره‌برداری از روند ندارد.

روش اول توانایی تولید بازده در بازار صعودی را دارد، اما مدل‌ها فاقد سازوکار تطبیقی برای کنترل ریسک و مدیریت پویای پرتفو هستند. در نتیجه، عملکرد آن‌ها در دوره‌های نوسان شدید یا تغییرات روند کوتاه‌مدت ناپایدار است.

روش دوم: مدل تطبیقی حساس به ریسک: جدول ۴ نتایج دوره صعودی مدل تطبیقی حساس به ریسک را نشان می‌دهد.

جدول ۴. عملکرد مدل تطبیقی حساس به ریسک در بازار صعودی (یافته‌های تحقیق)

Table 4. Performance of Independent Models in a Bull Market

مدل	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتینو	نسبت کالمار	حداکثر افت سرمایه٪
B&H	135,279	25,95	0,742	1,059	0,659	-39,38
PPO	223,805	85,00	1,553	2,577	2,358	-36,04
A2C	158,215	41,96	0,933	1,446	0,936	-44,83
DQN	136,837	27,06	0,761	1,084	0,687	-39,38
پیشنهادی	379,734	176,17	1,730	3,002	3,339	-52,75

بازده و ارزش نهایی پرتفو

- مدل تطبیقی بیشترین ارزش نهایی و بالاترین بازده سالیانه را در کل پژوهش ثبت کرده است.
- بازده ۱۷۶,۱۷٪ تقریباً دو برابر PPO و چهار برابر DQN است. این نشان‌دهنده توانایی شناسایی و بهره‌برداری از روند صعودی با حداکثر استفاده از فرصت‌ها است.

کنترل ریسک و نسبت‌های تعدیل‌شده

- نسبت شارپ ۱,۷۳۰ و سورتینو ۳,۰۰۲ نشان‌دهنده پایداری بالای بازده نسبت به نوسانات است.
- نسبت کالمار ۳,۳۳۹، اگرچه مدل پیشنهادی کمترین افت سرمایه را ایجاد نکرده است، اما به دلیل بازده بسیار بالاتر، نسبت کالمار آن به طور قابل توجهی بهبود یافته است. بنابراین مزیت مدل در بهینه‌سازی بازده تعدیل‌شده بر اساس ریسک است، نه لزوماً حداقل‌سازی مطلق افت سرمایه.

مزیت سوئیچ پویا و تطبیق با ریسک

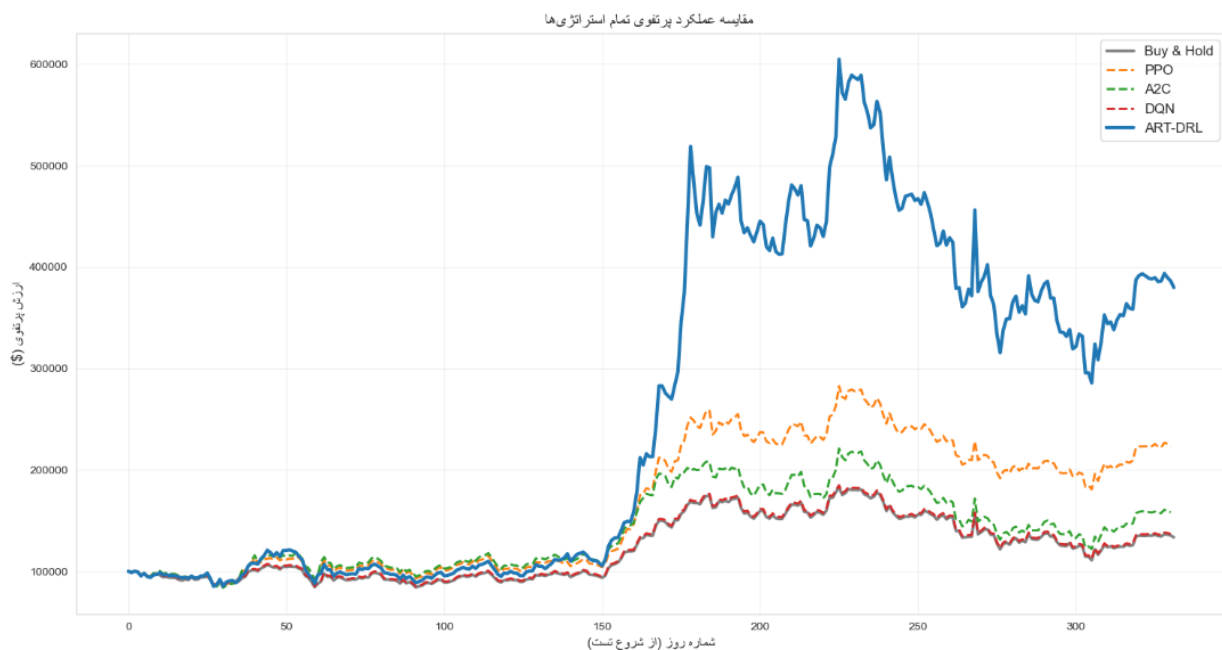
- مکانیسم سوئیچ پویا بین عامل‌های مختلف (تهاجمی، متعادل، محافظه‌کار) اجازه می‌دهد که در شروع روند صعودی، رفتار محافظه‌کار حفظ شود تا ریسک کاهش یابد و با تشخیص روند مثبت، عامل‌های تهاجمی فعال شوند و بیشترین سود کسب شود.
- این انعطاف‌پذیری، کنترل ریسک را با بهره‌برداری حداکثری از روند بازار ترکیب می‌کند.

تحلیل جامع شاخص‌ها

- مقادیر شاخص‌های شارپ، سورتینو و کالمار تأیید می‌کند که مدل تطبیقی، بهترین تعادل ممکن بین بازده و ریسک را فراهم می‌آورد.
- نمودار تجمعی بازده نشان می‌دهد که روند رشد مستمر و خط تجمعی بالای صفر حفظ شده است، در حالی که مدل‌های مستقل دچار جهش‌ها و افت‌های کوتاه‌مدت می‌شوند.

جمع‌بندی مقایسه‌ای بازار صعودی

- مدل تطبیقی حساس‌به‌ریسک در بازار صعودی برتری مطلق دارد: بیشترین ارزش پرتفو، بیشترین بازده سالیانه و بهترین نسبت‌های تعدیل‌شده بر اساس ریسک را ارائه می‌دهد.
 - مدل‌های مستقل یادگیری تقویتی عمیق توانایی شناسایی روند دارند اما فاقد سازوکار تطبیقی و کنترل دقیق ریسک هستند.
 - توانایی سوئیچ پویا در مدل تطبیقی، عامل اصلی موفقیت آن در تحقق بازده بالا و پایداری استراتژی در طول دوره صعودی است.
 - این نتایج تأکید می‌کنند که در بازارهای صعودی، استفاده از مدل‌های تطبیقی حساس‌به‌ریسک به مراتب کارآمدتر و مطمئن‌تر از مدل‌های مستقل یادگیری تقویتی عمیق است.
- شکل ۲ نمایی از مقایسه عملکرد مدل‌های خرید و نگهداری، مدل‌های یادگیری تقویتی مستقل و مدل پیشنهادی تطبیقی حساس به ریسک (بازار صعودی) را به نمایش گذاشته است.



شکل ۲. مقایسه عملکرد مدل های خرید و نگهداری، مدل های یادگیری تقویتی مستقل و مدل پیشنهادی (بازار صعودی)

Figure 2. Performance Comparison of Buy-and-Hold Models, Independent Reinforcement Learning Models, and the Proposed Model in a Bull Market

تحلیل حساسیت و پایداری پارامترهای مدل پیشنهادی: با توجه به ماهیت غیریستا، پرنوسان و وابسته به رژیم بازار رمزارزها، ارزیابی عملکرد مدل های مبتنی بر یادگیری تقویتی صرفاً در چارچوب یک تنظیم ثابت از پارامترها، قادر به ارائه تصویری جامع از پایداری، قابلیت تعمیم و منطق رفتاری آن ها نیست. از این رو، در این پژوهش تحلیل حساسیت به عنوان رویکردی تحلیلی و مکمل در کنار ارزیابی عملکرد به کار گرفته شد تا میزان وابستگی نتایج مدل به تنظیمات درونی تابع پاداش و افق تصمیم گیری عامل مورد بررسی قرار گیرد. این تحلیل در دو سطح انجام شد: نخست، مقایسه عملکرد مدل در دو رژیم متفاوت بازار، یعنی بازار بدون روند و بازار صعودی، که خود نوعی سنجش حساسیت نسبت به ساختار بازار محسوب می شود و میزان پایداری سیاست های آموخته شده در برابر تغییرات بنیادین پویایی بازار را آشکار می سازد؛ دوم، تحلیل حساسیت پارامتری نسبت به ضرایب کلیدی تابع پاداش، شامل وزن مؤلفه کالمار (α)، وزن جریمه نوسان (β) و ضریب تنزیل (γ)، که به صورت متمرکز در بازار صعودی بررسی شد. تمرکز بر بازار صعودی از آن جهت صورت گرفت که در این رژیم، سیگنال های بازدهی روشن تر و تفکیک پذیرتر هستند و در نتیجه، تغییر پارامترها اثر معنادارتری بر سیاست آموخته شده عامل برجای می گذارد؛ در حالی که در بازار بدون روند، محدود بودن فرصت های سودآور موجب می شود تفاوت های رفتاری ناشی از تغییر پارامترها کمتر نمایان شود. بر این اساس، تحلیل حساسیت در بازار صعودی نه تنها برای ارزیابی میزان پایداری مدل، بلکه برای تبیین منطق انتخاب مقادیر پایه α ، β و γ و نشان دادن این نکته انجام شد که این تنظیمات با هدف ایجاد تعادل میان بازده، ریسک، افق تصمیم گیری و پایداری سیاست برگزیده شده اند.

نتایج تحلیل حساسیت پارامتری: جدول ۵ نتایج تحلیل حساسیت چارچوب تطبیقی حساس به ریسک را در برابر تغییرات پارامترهای تابع پاداش و ضریب تنزیل در دوره بازار صعودی نشان می دهد. تمامی مقادیر گزارش شده بر پایه داده های دوره آزمون محاسبه شده اند تا پایداری و کارایی واقعی مدل در همه شرایط ارزیابی شود. در این تحلیل، α بیانگر وزن

مؤلفه کالمار در تابع پاداش است و نقش آن در تقویت تصمیم‌های مبتنی بر نسبت بازده به افت سرمایه برجسته می‌شود؛ β شدت جریمه نوسان را تعیین می‌کند و از این طریق میزان حساسیت عامل نسبت به رفتارهای پرریسک و ناپایدار را شکل می‌دهد؛ و γ نیز افق تصمیم‌گیری عامل را کنترل می‌کند و مشخص می‌سازد که پاداش‌های آینده با چه شدتی در تصمیم‌گیری لحاظ شوند. به‌طور مشخص، مقادیر بزرگ‌تر γ و به‌ویژه مقادیر نزدیک به ۱ نشان می‌دهند که عامل برای پاداش‌های آینده ارزش بالایی قائل است و تصمیم‌های خود را با نگاه بلندمدت‌تر تنظیم می‌کند؛ در مقابل، مقادیر کوچک γ بیانگر اولویت‌دادن به پاداش‌های فوری و کوتاه‌مدت و در نتیجه، افق تصمیم‌گیری محدودتر هستند. از این‌رو، این انتخاب به این معناست که مدل ضمن حفظ آینده‌نگری، پاداش‌های آتی را با تخفیفی اندک اما مؤثر در تصمیم‌گیری لحاظ می‌کند و همین امر به شکل‌گیری سیاستی پایدارتر و سازگارتر با هدف بیشینه‌سازی بازده بلندمدت کمک می‌کند. بر این اساس، بررسی هم‌زمان این سه پارامتر امکان ارزیابی این نکته را فراهم می‌کند که مدل تا چه اندازه در مواجهه با تغییر تنظیمات درونی تابع پاداش، همچنان رفتار پایدار، متوازن و سازگار با شرایط بازار صعودی از خود نشان می‌دهد.

جدول ۵. تحلیل حساسیت عملکرد مدل پیشنهادی نسبت به پارامترهای تابع پاداش در بازار صعودی (یافته‌های تحقیق)

Table 5. Sensitivity Analysis of the Proposed Model's Performance to Reward Function Parameters in a Bull Market

α	β	Γ	نسبت کالمار	نسبت سورتینو	نسبت شارپ	ارزش نهایی
۰,۰۰۱	۰,۸	۰,۹۹	2,52	2,36	1,48	282,764
۰,۰۰۱	۰,۸	۰,۸	3,61	2,89	1,70	358,804
۰,۰۰۱	۰,۹	۰,۹۹	5,26	3,23	1,83	399,953
۰,۰۰۱	۰,۹	۰,۸	3,76	2,93	1,76	386,929
۰,۰۰۵	۰,۸	۰,۹۹	3,84	2,87	1,63	346,910
۰,۰۰۵	۰,۸	۰,۸	3,57	2,84	1,68	252,225
۰,۰۰۵	۰,۹	۰,۹۹	3,34	3,00	1,73	379,734
۰,۰۰۵	۰,۹	۰,۸	5,69	3,82	2,01	399,961
۰,۰۰۱	۰,۸	۰,۹۹	5,21	3,54	1,92	363,832
۰,۰۰۱	۰,۸	۰,۸	3,61	3,23	1,76	336,437
۰,۰۰۱	۰,۹	۰,۹۹	3,32	2,74	1,60	338,962
۰,۰۰۱	۰,۹	۰,۸	4,10	2,89	1,69	355,005

در ادامه برای ارزیابی پایداری نتایج و کاهش اثر تصادفی بودن ناشی از مقداردهی اولیه در الگوریتم‌های یادگیری تقویتی، تمامی آزمایش‌ها به تعداد ۱۰ عدد seed مستقل با مقادیر (۰,۱۰۷,۱۱,۲۱,۴۲,۱۰۰,۹۹۹,۱۳۳۴,۲۰۲۴) تکرار شدند. سپس، شاخص‌های عملکرد هر روش به‌صورت میانگین و انحراف معیار گزارش شد تا علاوه بر نمایش عملکرد مرکزی مدل‌ها، میزان پراکندگی و حساسیت آن‌ها نسبت به تغییرات تصادفی نیز مشخص شود. این رویکرد امکان ارزیابی منصفانه‌تر و دقیق‌تر روش‌ها را فراهم می‌کند و از اتکا به نتایج یک اجرای منفرد جلوگیری می‌نماید. با توجه به اینکه در محیط‌های مالی پرنوسان، عملکرد الگوریتم‌ها می‌تواند به‌طور معناداری تحت تأثیر تصادفی بودن فرآیند آموزش قرار گیرد، گزارش نتایج بر مبنای چند مقداردهی اولیه مستقل نقش مهمی در افزایش قابلیت تکرارپذیری و اعتبار تجربی پژوهش دارد. بر این اساس، جدول ۶ خلاصه‌ای از عملکرد مدل‌های مورد بررسی را بر پایه ۱۰ اجرای مستقل ارائه می‌دهد و مبنایی مناسب برای مقایسه دقیق‌تر میان روش‌های پایه و مدل پیشنهادی فراهم می‌سازد.

جدول ۶. بررسی عملکرد مدل‌ها (میانگین و انحراف معیار) بر پایه ۱۰ اجرای مستقل در بازار صعودی (یافته‌های تحقیق)

Table 6. Performance of the Models (Mean and Standard Deviation) Based on 10 Independent Runs in a Bull Market

مدل	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتینو	نسبت کالمار	حداکثر افت سرمایه %
B&H	135,279	25,95	0,742	1,059	0,659	-39,38
PPO	174,834±44,815	52,41±29,6	1,05±0,37	1,61±0,67	1,42±0,98	-40,321±8,8
A2C	191,812±55,261	63,4±35,4	1,12±0,31	1,71±0,55	1,60±0,93	-40,372±8,3
DQN	137,065±431	27,2±0,4	0,76±0,05	1,08±0,08	0,69±0,08	-39,380±0,0
پیشنهادی	347,33±68,956	157,2±39,2	1,71±0,15	2,96±0,32	3,80±1,14	-42,178±6,7

در ادامه برای مقایسه عملکرد مدل پیشنهادی با مدل‌های مبنا از آزمون ناپارامتری رتبه علامت ویلکاکسون^۱ در سطح معناداری ۰.۰۵ استفاده شد. انتخاب آزمون ویلکاکسون به این دلیل صورت گرفت که توزیع شاخص‌های عملکرد در الگوریتم‌های یادگیری تقویتی لزوماً از فروض نرمال بودن تبعیت نمی‌کند و حجم نمونه نیز محدود است. این آزمون یکی از روش‌های توصیه‌شده برای مقایسه عملکرد الگوریتم‌های یادگیری ماشین و یادگیری تقویتی در مطالعات تجربی محسوب می‌شود. نتایج این آزمون در جدول ۷ نمایش داده شده است.

جدول ۷. نتایج آزمون ناپارامتری رتبه علامت ویلکاکسون در بررسی شاخص‌های عملکرد در بازار صعودی (یافته‌های تحقیق)

Table 7. Wilcoxon Signed-Rank Test Results for Performance Metrics in a Bull Market

مقایسه مدل‌ها	پارامتر آزمون	ارزش نهایی	بازده سالیانه	نسبت شارپ	نسبت سورتینو	نسبت کالمار	حداکثر افت سرمایه %
مدل vs PPO پیشنهادی	آماره w	۰,۰	۰,۰	۰,۰	۰,۰	۱,۰	۲۴,۰
	اختلاف میانه	۱۶۴,۶۵۸	۱۰۰,۵۵	۰,۷۱	۱,۴۴	۲,۱۱	۰,۵۷-
	p-value	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۴	۰,۷۷۰
مدل vs A2C پیشنهادی	آماره w	۰,۰	۰,۰	۰,۰	۰,۰	۰,۰	۲۱,۰
	اختلاف میانه	۱۳۳,۷۸۸	۸۰,۶۱	۰,۶۰	۱,۳۳	۱,۹۸	۱,۳۷-
	p-value	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۵۵۷
مدل vs DQN پیشنهادی	آماره w	۰,۰	۰,۰	۰,۰	۰,۰	۰,۰	۱۹,۰
	اختلاف میانه	۲۰۹,۵۷۳	۱۳۰,۴۰	۱,۰۰	۱,۹۷	۲,۸۳	۱,۶۸-
	p-value	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۰۰۲	۰,۴۳۲

یافته‌های حاصل از آزمون ویلکاکسون بیانگر آن است که مدل پیشنهادی نه تنها از نظر آماری، بلکه از منظر کمی نیز عملکردی به مراتب بهتر از مدل‌های پایه ارائه داده است. به جز شاخص حداکثر افت سرمایه که تفاوت معناداری را نشان نمی‌دهد، مدل پیشنهادی در تمامی معیارهای سنجش عملکرد (از جمله ارزش نهایی، بازده سالیانه و نسبت‌های شارپ، سورتینو و کالمار) با اختلاف میانه قابل توجه و $p - value$ بسیار پایین، عملکردی برتر و به لحاظ آماری معنادار کسب کرده است.

تفسیر و تحلیل نتایج: نتایج جدول تحلیل حساسیت نشان می‌دهد که عملکرد چارچوب تطبیقی حساس به ریسک در بازار صعودی، نسبت به تغییرات پارامترهای تابع پاداش و ضریب تنزیل، رفتاری منظم و قابل تفسیر دارد و از الگوی تصادفی یا ناپایدار پیروی نمی‌کند. بررسی جدول و نمودار حساسیت آشکار می‌سازد که تغییر هم‌زمان α ، β و γ بر ارزش نهایی پرتفوی و شاخص‌های ریسک تعدیل‌شده، از جمله نسبت شارپ، نسبت سورتینو و نسبت کالمار، اثر

¹ Wilcoxon Signed-Rank test

معناداری دارد؛ با این حال، این اثر در بازه‌ای منطقی باقی می‌ماند و نشان می‌دهد که ساختار مدل از انسجام درونی برخوردار است. از این رو، تحلیل حساسیت صرفاً با هدف شناسایی بهترین ترکیب پارامترها انجام نشده است، بلکه هدف آن نشان دادن این موضوع است که عملکرد مطلوب مدل ناشی از طراحی متعادل تابع پاداش و سازوکار تصمیم‌گیری آن بوده و به یک تنظیم خاص وابسته نیست.

بر اساس نتایج جدول، بهترین عملکرد در تنظیمی مشاهده شده است که در آن $\alpha = 0,001$ ، $\beta = 0,9$ و $\gamma = 0,8$ یا مقادیر نزدیک به آن در نظر گرفته شده‌اند. در این حالت، ارزش نهایی پرتفوی به ۳۹۹,۹۵۳ و در یک تنظیم بسیار نزدیک به ۳۹۹,۹۶۱ رسیده است و هم‌زمان نسبت شارپ، نسبت سورتینو و نسبت کالمار نیز به مقادیر بالاتری مانند ۱,۸۳، ۳,۲۳ و ۵,۲۶ و در تنظیم دیگر ۲,۰۱، ۳,۸۲ و ۵,۶۹ افزایش یافته‌اند. این نتایج نشان می‌دهد که در بازار صعودی، زمانی عملکرد مدل تقویت شده که وزن مؤلفه کالمار در سطح کنترل شده باقی مانده، جریمه نوسان در سطحی مؤثر تنظیم شده و ضریب تنزیل نیز به گونه‌ای انتخاب شده که میان توجه به بازده‌های آتی و واکنش به فرصت‌های جاری بازار تعادل برقرار شود. به بیان دیگر، برتری عملکرد در این مسئله نه از آینده‌نگری افراطی، بلکه از توازن میان تمرکز بر سودهای جاری، کنترل نوسان و حفظ کیفیت تصمیم‌گیری حاصل شده است.

در مقابل، زمانی که γ در سطح بالاتری مانند ۰,۹۹ قرار گرفته، حتی با ثابت بودن سایر پارامترها، در برخی حالات افت محسوسی در عملکرد مشاهده می‌شود. برای نمونه، در تنظیم $\alpha = 0,001$ ، $\beta = 0,80$ و $\gamma = 0,99$ ، ارزش نهایی به ۲۸۲,۷۶۴ کاهش یافته و نسبت شارپ، سورتینو و کالمار نیز به ترتیب ۱,۴۸، ۲,۳۶ و ۲,۵۲ ثبت شده‌اند. این تفاوت نشان می‌دهد که در این مسئله، صرف نزدیک بودن γ به ۱ به معنای عملکرد بهتر نیست، بلکه مقدار این پارامتر باید با ویژگی‌های رژیم بازار و ماهیت فرصت‌های سودآور آن هم‌خوانی داشته باشد. به بیان دقیق‌تر، مقدار مناسب γ باید نه آن قدر بالا باشد که عامل بیش از حد به پاداش‌های دوردست متکی شود، و نه آن قدر پایین که افق تصمیم‌گیری بیش از اندازه کوتاه شود؛ بلکه باید تعادلی میان آینده‌نگری و بهره‌برداری از سیگنال‌های فوری برقرار کند.

از سوی دیگر، مقایسه نتایج ردیف‌های مختلف جدول بیانگر آن است که تغییر α از ۰,۰۰۱ به ۰,۰۰۵ و ۰,۰۱۰، در صورت ثابت بودن سایر پارامترها، الزاماً به بهبود عملکرد منجر نمی‌شود و در برخی موارد افت کارایی را به همراه دارد. برای مثال، در حالت $\alpha = 0,005$ ، $\beta = 0,80$ و $\gamma = 0,99$ ، ارزش نهایی برابر ۳۴۶,۹۱۰ و نسبت شارپ ۱,۶۳ ثبت شده است، در حالی که در ترکیب $\alpha = 0,005$ ، $\beta = 0,90$ و $\gamma = 0,8$ ، ارزش نهایی به ۳۹۹,۹۶۱ و نسبت شارپ به ۲,۰۱ رسیده است. این تفاوت نشان می‌دهد که عملکرد مدل در محدوده‌های متفاوت α و β تغییر می‌کند، اما در حوالی مقادیر آزموده شده، امکان دستیابی به نتایج بهتر و پایدارتر وجود دارد. همچنین، β نقش مهمی در مهار نوسان و تقویت رفتار پایدار دارد؛ به گونه‌ای که افزایش آن تا سطح ۰,۹ در اغلب موارد به بهبود شاخص‌های ریسک‌تعدیل شده منجر شده است. از این رو، عملکرد مدل نه تنها به مقدار هر پارامتر، بلکه به تعامل میان آن‌ها وابسته است.

علاوه بر تحلیل حساسیت پارامترها، نتایج حاصل از ۱۰ مقداردهی اولیه مستقل حاکی از آن است که مدل پیشنهادی از نظر آماری نیز دارای عملکردی پایدار، منسجم و قابل‌اتکا است. گزارش نتایج به صورت میانگین و انحراف معیار این امکان را فراهم می‌سازد که علاوه بر برآورد عملکرد مرکزی، میزان پراکندگی و حساسیت مدل نسبت به تغییرات ناشی از فرآیند آموزش نیز مورد ارزیابی قرار گیرد. بر اساس جدول ارائه شده، مدل پیشنهادی در شاخص‌های کلیدی عملکرد، از جمله ارزش نهایی پرتفوی، بازده سالیانه و نسبت‌های تعدیل شده بر اساس ریسک، نه تنها به مقادیر میانگین بالاتری دست یافته، بلکه از نوسان کمتری نسبت به روش‌های مقایسه‌ای نیز برخوردار بوده است. این امر نشان می‌دهد که برتری مشاهده شده صرفاً ناشی از یک اجرای خاص یا مقداردهی اولیه تصادفی مطلوب نیست، بلکه ریشه در ساختار یادگیری و سازوکار تصمیم‌گیری مدل دارد. در مجموع، هم‌راستایی نتایج تحلیل حساسیت با یافته‌های حاصل از ۱۰ مقداردهی اولیه مستقل نشان می‌دهد که چارچوب تطبیقی حساس به ریسک در بازار صعودی، در برابر تغییرات پارامتری

و عدم قطعیت‌های ناشی از آموزش، از پایداری مناسبی برخوردار است. این هم‌خوانی میان کارایی عملکردی و ثبات آماری، پشتوانه تجربی معتبری برای برتری مدل پیشنهادی فراهم می‌آورد و بیانگر آن است که عملکرد مطلوب آن حاصل طراحی متوازن تابع پاداش، سازوکار انتخاب تطبیقی و تنظیم مناسب افق تصمیم‌گیری است. نتایج حاصل از آزمون رتبه علامت ویلکاکسون نشان می‌دهد که مدل پیشنهادی در اغلب شاخص‌های عملکردی مورد بررسی، برتری آماری معناداری نسبت به روش‌های پایه دارد. در دو شاخص ارزش نهایی سرمایه و بازده سالیانه، مدل پیشنهادی نسبت به هر سه الگوریتم پایه عملکردی به مراتب بهتر و از نظر آماری معنادار نشان داد $p \approx 0.002$. اختلاف میانه‌ی ارزش نهایی سرمایه در مقایسه با PPO، A2C و DQN به ترتیب برابر با حدود ۱۶۴'۶۵۸، ۱۳۳'۷۸۸ و ۲۰۹'۵۷۳ دلار بود. همچنین میزان بهبود بازده سالیانه نسبت به این سه روش به ترتیب ۱۰۰،۵۵٪، ۸۰،۶۱٪ و ۱۳۰،۴۰٪ به دست آمد. مقدار آماره‌ی $W = 0$ در این مقایسه‌ها نشان می‌دهد که مدل پیشنهادی در تمامی تکرارهای آزمایش، بدون استثنا، عملکردی برتر از روش‌های پایه داشته است.

در سه شاخص نسبت شارپ، نسبت سورتینو و نسبت کالمار نیز مدل پیشنهادی برتری آماری معناداری نسبت به هر سه الگوریتم پایه نشان داد ($p\text{-value} \in [0, 0.02, 0.004]$). بهبود مشاهده‌شده در نسبت شارپ تا ۱،۰ در نسبت سورتینو تا ۱،۹۷ و در نسبت کالمار تا ۲،۸۳ بیانگر آن است که عملکرد بهتر مدل پیشنهادی صرفاً ناشی از پذیرش ریسک بیشتر نبوده، بلکه این مدل توانسته است بازده تعدیل‌شده بر حسب ریسک را نیز به‌طور معناداری ارتقا دهد. در شاخص حداکثر افت سرمایه، اگرچه مدل پیشنهادی در برخی موارد مقادیر اندکی کمتر و مطلوب‌تری نسبت به الگوریتم‌های پایه داشت، اما این تفاوت‌ها از نظر آماری معنادار نبودند (همگی بزرگ‌تر ۰،۰۵). مقادیر نسبتاً بالاتر آماره‌ی W در این مقایسه‌ها ۱۹ تا ۲۴ نیز این نتیجه را تأیید می‌کند. بنابراین، می‌توان گفت که مدل پیشنهادی از نظر کنترل بیشینه افت سرمایه عملکردی هم‌سطح با روش‌های پایه ارائه کرده است، هرچند برتری آماری معناداری در این شاخص مشاهده نشد.

به‌طور کلی، نتایج آزمون ویلکاکسون نشان می‌دهد که مدل پیشنهادی در شاخص‌های مرتبط با بازدهی و کارایی ریسک‌تعدیل‌شده، برتری آماری معناداری نسبت به الگوریتم‌های PPO، A2C و DQN دارد. با این حال، در شاخص حداکثر افت سرمایه تفاوت معناداری میان مدل پیشنهادی و روش‌های پایه مشاهده نشد و عملکرد مدل در این زمینه قابل مقایسه با سایر الگوریتم‌ها بود. این یافته‌ها بیانگر آن است که بهبود عملکرد مدل پیشنهادی پایدار، قابل اتکا و مستقل از نوسانات تصادفی میان تکرارهای آزمایش است.

۵. بحث و نتیجه‌گیری

این پژوهش با هدف ارائه یک چارچوب تطبیقی حساس به ریسک برای مدیریت پرتفو در بازار رمزارزها انجام شد؛ بازاری که به دلیل نوسان‌های شدید، تغییرات سریع رژیم، و نبود الگوهای پایدار، یکی از دشوارترین محیط‌ها برای تصمیم‌گیری مبتنی بر یادگیری تقویتی به شمار می‌آید. مسئله اصلی این بود که هیچ عامل منفردی، هرچند قدرتمند، در همه شرایط بازار عملکرد یکنواخت و پایداری ندارد. از همین رو، در این پژوهش به جای اتکا به یک سیاست ثابت، چارچوبی طراحی شد که از ترکیب سه عامل PPO، A2C و DQN بهره می‌گیرد و از طریق یک انتخاب‌گر مبتنی بر ترنسفورمر، وزن هر عامل را متناسب با وضعیت بازار تنظیم می‌کند. منطق اصلی این طراحی آن است که برتری هر الگوریتم، وابسته به رژیم بازار است؛ بنابراین، به جای تلاش برای یافتن یک عامل برتر جهانی، مدل پیشنهادی می‌کوشد عامل مناسب را در زمان مناسب فعال کند. نتایج تجربی نشان داد که این منطق طراحی به‌طور مستقیم در خروجی‌های

عددی مدل بازتاب یافته است. در بازار بدون روند، A2C در میان عامل‌های منفرد عملکرد بهتری نسبت به PPO و DQN داشت، زیرا این الگوریتم به تغییرات کوتاه‌مدت حساس‌تر است و در محیط‌های فاقد جهت‌گیری مشخص قیمت، توانایی بیشتری برای واکنش سریع دارد. با این حال، چارچوب پیشنهادی تنها به بهره‌گیری از این مزیت اکتفا نکرد، بلکه با کنترل پویا میان عامل‌ها، مانع از تبدیل این حساسیت به نوسان یا افت سرمایه شد. به همین دلیل، ارزش نهایی پرتفو به ۲۱۳,۷۲۷ دلار و بازده سالانه به ۶۷,۵۲ درصد رسید؛ رقمی که نه صرفاً بیانگر سود بالاتر، بلکه نشان‌دهنده توان مدل در تبدیل یک مزیت محلی A2C به یک عملکرد پایدار در سطح سامانه است. این نتیجه نشان می‌دهد که طراحی چندعاملی، زمانی ارزشمند می‌شود که انتخاب‌گر بتواند مزیت رفتاری هر عامل را در بافت مناسب خود فعال کند و اثرات منفی آن را در همان لحظه مهار نماید.

در بازار صعودی، الگوی عملکرد متفاوت بود. در این رژیم، PPO در میان مدل‌های منفرد بهترین نتیجه را به دست آورد، زیرا سیاست‌های پایدارتر آن برای بهره‌برداری از روندهای بلندمدت مناسب‌تر است. اما نکته مهم آن است که چارچوب پیشنهادی صرفاً به تقلید از PPO نپرداخت، بلکه با تشخیص رژیم صعودی، وزن این عامل را افزایش داد و در عین حال، از سایر عامل‌ها برای حفظ تعادل تصمیم‌گیری استفاده کرد. حاصل این سازوکار، دستیابی به ارزش نهایی پرتفو ۳۷۹,۷۳۴ دلار و بازده سالانه ۱۷۶,۱۷ درصد بود. این اعداد تنها افزایش کمی بازده را نشان نمی‌دهند؛ بلکه بیانگر آن‌اند که معماری مدل توانسته بین بهره‌برداری از روند صعودی و جلوگیری از تمرکز بیش‌ازحد بر یک سیاست خاص، توازن برقرار کند. بنابراین، برتری مدل پیشنهادی نه از یک الگوریتم منفرد، بلکه از هماهنگی میان طراحی ساختار، انتخاب‌گر پویا و رفتار رژیم‌محور عامل‌ها ناشی شده است.

شاخص‌های تعدیل‌شده بر اساس ریسک نیز این تحلیل را تأیید می‌کنند. نسبت شارپ ۱,۷۳۰، نسبت سورتینو ۳,۰۰۲ و نسبت کالمار ۳,۳۳۹ نشان می‌دهند که بهبود عملکرد مدل فقط نتیجه افزایش ریسک‌پذیری نبوده است. اگر مدل صرفاً به دنبال بیشینه‌سازی بازده بود، احتمالاً باید افزایش شدیدتری در نوسان و افت سرمایه مشاهده می‌شد، اما بهبود هم‌زمان این شاخص‌ها نشان می‌دهد که چارچوب پیشنهادی موفق شده است میان بازده و کنترل ریسک تعادل برقرار کند. این نکته به‌ویژه از منظر طراحی مدل اهمیت دارد، زیرا تابع پاداش و سازوکار انتخاب‌گر به‌گونه‌ای تنظیم شده‌اند که عامل نه فقط به سود خام، بلکه به کیفیت این سود نیز توجه کند. در نتیجه، خروجی‌های عددی مستقیماً با منطق طراحی همسو هستند و از یک طراحی تصادفی یا آزمون‌وخطای صرف حکایت نمی‌کنند.

تحلیل حساسیت نیز این رابطه را به‌صورت دقیق‌تر آشکار کرد. نتایج این تحلیل نشان داد که عملکرد مدل در بازار صعودی نسبت به تغییرات α ، β و γ الگوی منسجم و قابل تفسیر دارد. در بهترین ترکیب‌های مشاهده‌شده، مقدار α در سطحی قرار داشته که مؤلفه کالمار را به اندازه کافی در تابع پاداش نگه می‌دارد، β به‌گونه‌ای تنظیم شده که نوسان‌های نامطلوب را مهار کند، و γ در سطحی باشد که میان توجه به پاداش‌های آتی و واکنش به فرصت‌های فعلی بازار تعادل برقرار شود. نتیجه این تنظیمات، رسیدن به مقادیر بالای ارزش نهایی و شاخص‌های ریسک تعدیل‌شده بود. این موضوع نشان می‌دهد که ضرایب پاداش مستقیماً جهت‌گیری رفتاری عامل را شکل می‌دهند. از طرف دیگر، تحلیل حساسیت نشان داد که مقادیر γ نزدیک به ۱ لزوماً بهترین عملکرد را تضمین نمی‌کنند و باید در ارتباط با رژیم بازار تفسیر شوند. در بازار صعودی، مقدار $\gamma = 0,8$ در برخی تنظیم‌ها عملکرد بهتری نسبت به $\gamma = 0,99$ داشته است، زیرا مدل را از وابستگی بیش‌ازحد به پاداش‌های دوردست دور کرده و امکان استفاده بهتر از فرصت‌های جاری را فراهم کرده است. این یافته از نظر علمی مهم است، چون نشان می‌دهد ضریب تنزیل نباید به‌صورت انتزاعی و جدا از داده تفسیر شود؛ بلکه باید با ساختار بازار، افق فرصت‌ها و رفتار عامل هماهنگ باشد. همین نکته در مورد α و β نیز صادق است: تغییر این پارامترها زمانی سودمند است که در چارچوبی متعادل و در تعامل با یکدیگر صورت گیرد، نه به‌صورت مستقل و مکانیکی.

افزون بر نتایج عملکردی گزارش شده، تحلیل مبتنی بر ۱۰ مقداردهی اولیه مستقل نشان می‌دهد که برتری مدل پیشنهادی به یک اجرای منفرد یا شرایط تصادفی خاص در فرآیند آموزش محدود نبوده، بلکه در تکرارهای مختلف نیز با ثبات قابل توجهی تداوم یافته است. گزارش شاخص‌های عملکرد به صورت میانگین و انحراف معیار بیانگر آن است که مدل پیشنهادی علاوه بر دستیابی به عملکرد مرکزی مطلوب‌تر، در مقایسه با روش‌های مبنا از پراکندگی کمتر و پایداری آماری بیشتری برخوردار است. این موضوع نشان می‌دهد که مزیت مشاهده شده در مدل پیشنهادی صرفاً ناشی از یک اجرای خوش‌بینانه یا نوسانات تصادفی مقداردهی اولیه نیست، بلکه می‌توان آن را به ساختار یادگیری و سازوکار تطبیقی انتخاب عامل نسبت داد. برای تقویت این نتیجه، آزمون ناپارامتری رتبه علامت ویلکاکسون نیز در سطح معناداری ۰.۰۵ به کار گرفته شد. نتایج این آزمون نشان داد که مدل پیشنهادی در شاخص‌های ارزش نهایی سرمایه، بازده سالیانه، نسبت شارپ، نسبت سورتینو و نسبت کالمار نسبت به الگوریتم‌های پایه A2C، PPO و DQN دارای برتری آماری معنادار است. بنابراین، شواهد حاصل از تکرارهای مستقل و آزمون آماری ویلکاکسون به صورت هم‌زمان تأیید می‌کنند که بهبود عملکرد مدل پیشنهادی پایدار، معنادار و قابل اتکا است. با این حال، در شاخص حداکثر افت سرمایه تفاوت معناداری میان مدل پیشنهادی و روش‌های مبنا مشاهده نشد که بیانگر عملکرد قابل مقایسه آن‌ها در کنترل بیشینه افت سرمایه است. در مجموع، این یافته‌ها نشان می‌دهد که چارچوب پیشنهادی نه تنها از نظر بازدهی و کارایی ریسک‌تعدیل شده عملکرد برتری دارد، بلکه این برتری از پشتوانه آماری کافی برخوردار بوده و به اجرای تصادفی یا شرایط خاص آموزش وابسته نیست.

از منظر روش‌شناختی، مهم‌ترین دستاورد این پژوهش آن است که نشان می‌دهد طراحی معماری چندعاملی، تابع پاداش حساس به ریسک، و انتخاب گر مبتنی بر ترنسفورمر، هر سه در کنار هم به ایجاد یک سامانه تصمیم‌گیری مقاوم‌تر منجر شده‌اند. به بیان دیگر، برتری مدل در نتایج عددی، پیامد مستقیم طراحی نظری آن است. اگر مدل تنها بر اساس سود خام آموزش می‌دید، احتمالاً در برخی رژیم‌ها عملکرد بالایی نشان می‌داد اما پایداری آن در بازارهای دیگر کاهش می‌یافت. در مقابل، اگر فقط بر ریسک تمرکز می‌کرد، بازده نهایی افت می‌کرد. طراحی حاضر توانسته این دو هدف متعارض را به صورت هم‌زمان مدیریت کند و همین امر علت اصلی تفاوت معنادار آن با مدل‌های منفرد است.

از منظر مقایسه با ادبیات پیشین، نتایج این پژوهش یافته‌های [وانگ و لیو \(۲۰۲۵\)](#) را که بر اهمیت طراحی توابع پاداش حساس به ریسک تأکید کرده‌اند، تأیید می‌کند. با این حال، نتایج حاضر نشان می‌دهد که حتی با تابع پاداش یکسان، عملکرد عوامل منفرد در تمامی شرایط بازار پایدار نیست. نوآوری اصلی این پژوهش نسبت به مطالعه وانگ و لیو، افزودن یک لایه انتخابگر پویای عامل‌هاست که امکان تطبیق ساختاری با تغییر رژیم‌های بازار را فراهم می‌کند. همچنین، در حالی که [چن و همکاران \(۲۰۲۲\)](#) روش PPO را به عنوان برترین الگوریتم معرفی کرده‌اند، یافته‌های این پژوهش نشان می‌دهد که برتری این الگوریتم به بازارهای دارای روند محدود می‌شود و این الگوریتم در بازارهای بدون روند، در مقایسه با A2C، بخشی از فرصت‌های نوسان‌گیری را از دست می‌دهد. چارچوب تطبیقی پیشنهادی با بهره‌گیری هم‌زمان از مزایای A2C در بازار بدون روند و PPO در بازار صعودی، به صورت تجربی نتایج [باهو و همکاران \(۲۰۲۴\)](#) مبنی بر ناکارایی مدل‌های ایستا در بازارهای با تغییر رژیم سریع را تأیید می‌کند.

در نهایت، نتایج این پژوهش نشان می‌دهد که در بازار رمزارزها، مسئله صرفاً ساخت یک عامل قدرتمند نیست، بلکه طراحی سازوکاری است که بتواند قابلیت‌های هر عامل را در زمان و شرایط مقتضی به کار گیرد. یافته‌های تجربی، شاخص‌های تعدیل شده بر اساس ریسک و تحلیل حساسیت همگی در یک راستا قرار دارند و بیانگر آن هستند که چارچوب تطبیقی حساس به ریسک، به واسطه طراحی منسجم خود، توانسته است هم بازده و هم پایداری را بهبود دهد.

از این منظر، اهمیت پژوهش حاضر نه تنها در ارائه یک مدل جدید، بلکه در تبیین این نکته است که عملکرد برتر در بازار رمزارزها در گرو هم‌سویی معماری مدل، تابع پاداش و منطق انتخاب‌گری با ماهیت رژیم‌محور بازار است.

محدودیت‌ها: وجود نتایج امیدوارکننده و عملکرد برتر مدل پیشنهادی، این پژوهش با محدودیت‌هایی نیز همراه است که باید در تفسیر یافته‌ها مدنظر قرار گیرند. نخست، ارزیابی تجربی مدل صرفاً در بازار رمزارزها انجام شده است؛ از این رو تعمیم مستقیم نتایج به سایر بازارهای مالی نظیر سهام، کالا، ارز یا اوراق بدهی نیازمند بررسی‌های تکمیلی است. دوم، آزمایش‌های پژوهش بر داده‌های دارایی‌های منتخب بازار رمزارز متمرکز بوده و ممکن است ویژگی‌های ساختاری سایر دارایی‌ها یا بازارهای مالی متفاوت باشد. سوم، اگرچه برای افزایش قابلیت اطمینان نتایج، آزمایش‌ها با ده مقداردهی اولیه مستقل تکرار و از آزمون آماری ویلکاکسون برای بررسی معناداری تفاوت عملکرد مدل‌ها استفاده شد، اما ارزیابی مدل بر پایه یک تقسیم ثابت آموزش و آزمون انجام گرفته است و از چارچوب‌های Walk-Forward Validation یا Rolling Window برای بررسی پایداری زمانی در دوره‌های متوالی استفاده نشده است. در نهایت، تغییرات ساختاری بازارهای مالی، وقوع شوک‌های کلان اقتصادی و تحول رژیم‌های معاملاتی می‌تواند بر عملکرد بلندمدت مدل‌های یادگیری تقویتی اثرگذار باشد. از این رو، پیشنهاد می‌شود در پژوهش‌های آتی، چارچوب پیشنهادی در بازارهای مالی متنوع‌تر، دوره‌های زمانی طولانی‌تر و تحت روش‌های اعتبارسنجی پویاتر مورد ارزیابی قرار گیرد تا میزان تعمیم‌پذیری و پایداری آن با دقت بیشتری بررسی شود.

پیشنهادهای برای پژوهش‌های آتی: بر اساس یافته‌ها و محدودیت‌های شناسایی شده، موارد زیر برای توسعه آتی پیشنهاد می‌شود:

۱- ادغام تحلیل احساسات: از آنجا که مدل فعلی صرفاً بر داده‌های قیمت و حجم متکی است، پیشنهاد می‌شود در پژوهش‌های آینده، شاخص‌های تحلیل احساسات اخبار و شبکه‌های اجتماعی به عنوان یک ورودی اضافی به لایه ترنسفورمر افزوده شود تا واکنش مدل به شوک‌های خبری سریع‌تر گردد.

۲- مدل‌سازی شرایط واقعی بازار: برای نزدیک‌تر شدن نتایج به واقعیت معاملاتی، ضروری است هزینه‌های معاملاتی متغیر و اثرات لغزش قیمت در تابع پاداش لحاظ شود.

۳- توسعه برای پوشش ریسک: توسعه مدل به گونه‌ای که بتواند علاوه بر تتر، از ابزارهای مشتقه یا موقعیت‌های فروش تعهدی برای مدیریت ریسک در بازارهای نزولی استفاده کند.

این پژوهش در نهایت مسیر را برای گذار از سیستم‌های معاملاتی پیش‌بینی‌محور به سیستم‌های تطبیق‌محور هموار می‌سازد؛ جایی که بقا و سودآوری نه در پیش‌گویی قیمت، بلکه در مدیریت منعطف ریسک نهفته است.

تعارض منافع. برای ارائه مطالب و نگارش این مقاله هیچ‌گونه کمک مالی از هیچ فرد، نهاد و سازمانی دریافت نشده است و نتایج و دستاوردهای این مقاله به نفع یا ضرر سازمان یا فردی خاص نخواهد بود. حضور نویسندگان در این پژوهش به‌عنوان شاهدی بی‌طرف ولی متخصص بوده است و نویسندگان هیچ‌گونه تعارض منافی ندارند.

منابع

- Abdi, N., Moradzadehfard, M., Ahmadzadeh, H., & Khoddam, M. (2022). A hybrid model for portfolio optimization based on stock price forecasting with LSTM recurrent neural network using cardinality constraints and multi-criteria decision-making methods. *Journal of Financial Management Perspective*, 36, 119–143. (in Persian)
- Ayatollahi, M.R., & Jafari, M.B. (2024). Deep reinforcement learning and transformers: A novel approach to intelligent trading in Tehran Stock Exchange, *The Scientific-*

- Specialized Quarterly Journal of New Technologies in Electrical Engineering and Computer*, 4(2), 94–106. (in Persian)
- Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis. *SN Business & Economics*, 4, 23.
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 34, 15084–15097.
- Chen, L., et al. (2022). PPO for cryptocurrency portfolio optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8), 4012–4025.
- Cheng, L. C., & Sun, J. S. (2024). Multiagent-based deep reinforcement learning framework for multi-asset adaptive trading and portfolio management. *Neurocomputing*, 594, 127800.
- Cornalba, F., Disselkamp, C., Scassola, D., & Helf, C. (2024). Multi-objective reward generalization: improving performance of Deep Reinforcement Learning for applications in single-asset trading. *Neural Computing and Applications*, 36, 619–637.
- Dalili, S., Rezaei Piteh Novi, Y., Kabiri, M. T., Safari Graili, M., & Arabzadeh, M. (2024). Portfolio Optimization Using Deep Reinforcement Learning Based on Modern Portfolio Theory, *Accounting, Finance and Computational Intelligence*, 2(4), 276–293. (in Persian)
- 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.
- Du, Y., Tang, K., & Chen, K. (2023). A novel crude oil futures trading strategy based on ensemble deep reinforcement learning. *Energy*, 285, 128474.
- Fabozzi, F. J., et al. (2022). *Portfolio Theory and Risk Management*. Wiley.
- Gholami, N., & Shams Gharneh, N. (2024). Presenting an optimized CNN-LSTM model for stock price forecasting in the Tehran Stock Exchange. *Journal of Financial Management Perspective*, 14(45), 123–147. (in Persian)
- Jaffri, A., Shirvani, A., Jha, A., Rachev, S., & Fabozzi, F. (2025). Optimizing portfolios with Pakistan-exposed ETFs: Risk and performance insight. *Journal of Risk and Financial Management*, 18(3), 158. (in Persian)
- Koratomaddi, P., Wadhwani, K., Gupta, M., & Sanjeevi, S. (2021). Market sentiment-aware deep reinforcement learning for stock portfolio allocation. *Engineering Science and Technology*, 24, 848–859.
- Lin, K., et al. (2023). Risk-sensitive reinforcement learning for portfolio optimization. *Finance Research Letters*, 52, 103512.
- Lin, Y. C., Chen, C. T., Sang, C. Y., & Huang, S. H. (2022). Multiagent-based deep reinforcement learning for risk-shifting portfolio management. *Applied Soft Computing*, 123, 108894.

- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
- Wang, X., & Liu, L. (2025). Risk-sensitive deep reinforcement learning for portfolio optimization. *Journal of Risk and Financial Management*, 15, 347.
- Zare, M. H., Nilchi, M., & Fareed, D. (2020). Comparative evaluation of Markowitz approach with a new hybrid method to create an optimal portfolio using deep DNN learning method and gravitational search algorithm. *Journal of Financial Management Perspective*, 9(28), 165–188. (in Persian)
- Zhang, X., et al. (2021). Deep reinforcement learning for regime-switching markets. *Journal of Financial Econometrics*, 19(4), 789–810.
- Zhang, Y., Zohren, S., & Roberts, S. (2020). Deep reinforcement learning for trading strategies. *Quantitative Finance*, 20(9), 1459–1473.