

Portfolio Optimization Using Deep Reinforcement Learning Based on Modern Portfolio Theory

1. Yaser Rezaei Piteh Novi¹: Department of Accounting, Kas.C., Islamic Azad University, Kashan, Iran

2. Mehdi Safari Graili^{2*}: Department of Accounting, BG.C, Islamic Azad University, Arak, Iran. Email: safari83@iau.ac.ir (Corresponding Author)

3. Mohammad Taghi Kabiri³: Department of Accounting, Ar.c, Islamic Azad University, Arak, Iran

4. Meysam Arabzadeh⁴: Department of Accounting, Ra.c, Islamic Azad University, Rasht, Iran

Article history



Received: 25 November 2024

Revised: 17 January 2025

Accepted: 26 January 2025

Published: 13 February 2025

Abstract:

This study investigates the application of deep reinforcement learning in optimizing investment portfolios and integrates it with modern portfolio theory, illustrating significant advancements in financial management strategies. While modern portfolio theory is recognized as a mathematical framework aimed at maximizing expected returns while considering risk, its limitations—such as assumptions regarding the normal distribution of returns and the neglect of transaction costs—clearly highlight the need for adaptable solutions in complex and dynamic financial markets. This research demonstrates that through the implementation of deep reinforcement learning, investors can leverage real-time data and dynamic decision-making capabilities to develop more efficient and robust investment strategies. Moreover, challenges such as data quality, computational complexity, and the interpretability of deep reinforcement learning models are thoroughly examined. In this study, a Levenberg–Marquardt neural network algorithm for deep reinforcement learning is proposed for portfolio optimization based on historical data. For this purpose, data from 10 highly liquid companies listed on the Tehran Stock Exchange during the period from 2011 to 2021 were utilized. The findings of this study indicate that reinforcement learning algorithms can lead to a 15% increase in cumulative return compared to traditional methods in portfolio selection. Furthermore, the article recommends that analysts and investors employ advanced techniques to enhance corporate performance stability, enabling better investment decisions. Finally, this research offers practical pathways for future investigations in this domain.

Keywords: Levenberg–Marquardt algorithm, portfolio optimization, modern portfolio theory, deep reinforcement learning.

Citation: Rezaei Piteh Novi, Y., Safari Graili, M., Kabiri, M. T., & Arabzadeh, M. (2024). Portfolio Optimization Using Deep Reinforcement Learning Based on Modern Portfolio Theory, *Accounting, Finance and Computational Intelligence*, 2(4), 276-293.



Copyright: © 2024 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Extended Abstract

Introduction

Portfolio optimization remains a cornerstone challenge in modern finance, having evolved significantly since the introduction of Modern Portfolio Theory (MPT) by Harry Markowitz in the 1950s. MPT offered a mathematical framework that emphasized diversification and risk-return tradeoffs based on historical covariance matrices, assuming normally distributed returns and frictionless markets. Despite its foundational contributions, MPT has shown critical limitations in practice, such as sensitivity to historical parameters, neglect of transaction costs, and oversimplified distributional assumptions. These deficiencies have prompted researchers to explore more adaptive and intelligent methodologies that can respond to the volatility, non-stationarity, and multidimensionality of today's financial markets (Benhamou et al., 2020; Sutton & Barto, 2018).

In response to the limitations of traditional models, Deep Reinforcement Learning (DRL) has emerged as a transformative approach capable of learning optimal strategies through interaction with a complex environment. DRL algorithms simulate human-like decision-making by adjusting strategies through reward-feedback loops, enabling investors to process real-time financial data and adapt to changing market conditions (Jiang et al., 2017; Mnih et al., 2015). Unlike static models, DRL excels in dynamic and uncertain environments, particularly in high-frequency or short-term trading scenarios (Chaouki et al., 2020; Deng et al., 2016). Additionally, integrating DRL with MPT offers a unique advantage: the mathematical robustness of MPT can be retained through carefully designed reward functions, while DRL compensates for MPT's rigidity by incorporating nonlinear data such as market sentiment, technical indicators, and unstructured data (Cong et al., 2021; Yifu et al., 2024).

Several studies have demonstrated the efficacy of DRL in outperforming conventional models under complex financial conditions. For instance, the AlphaPortfolio framework utilized interpretable DRL to construct portfolios directly and showed superior performance over benchmark strategies (Cong et al., 2021). Similarly, the model developed by Liang et al. incorporated adversarial training to deal with market unpredictability, highlighting DRL's adaptability to extreme market events (Liang et al., 2018). Beyond algorithmic advantages, DRL models also facilitate the inclusion of diverse constraints, such as investor risk tolerance, liquidity preferences, and transaction cost thresholds (Jyotirmayee Behera, 2025; Koratamaddi et al., 2021). Recent meta-analyses confirm that DRL models consistently yield better cumulative returns and reduced volatility compared to traditional and supervised learning methods (Ayari Salah, 2025; Chaher, 2025).

Despite their promise, DRL models come with challenges such as data noise, computational intensity, and low interpretability (Almahdi & Yang, 2017; Liu et al., 2020). Researchers have attempted to overcome these by adopting hybrid neural architectures—such as combining convolutional and recurrent layers—and by using transfer learning to minimize training data dependencies (Henriques & Sadorsky, 2023; Lima Paiva et al., 2021). The current study builds on these developments by proposing a DRL framework based on the Levenberg–Marquardt neural algorithm for optimizing stock portfolios in the Tehran Stock Exchange, aiming to bridge the theoretical gap between MPT and modern AI-driven investment models.

Methods and Materials

This study designed a DRL-based portfolio optimization model using a Levenberg–Marquardt neural network architecture. Historical data were collected from ten highly liquid companies listed on the Tehran Stock Exchange, covering the period from 2011 to 2021. The input features included technical indicators such as Moving Average (MA), Relative Strength Index (RSI),

and MACD, along with financial metrics like return on assets (ROA), profit before tax (PTI), tax expenses (CTE), net profit, and fixed asset valuation (PPE). Data preprocessing involved normalization, outlier removal, and handling missing values via imputation techniques such as mean replacement and mode substitution.

The reward function was engineered to align with the MPT goal of maximizing return while minimizing risk. The state space included a four-dimensional tensor representing daily technical indicators across assets, and the action space defined portfolio weight allocations. The DRL agent was trained using an iterative feedback loop, adjusting weights through backpropagation and optimizing performance metrics. The model's output was evaluated using metrics such as cumulative return, volatility, Sharpe ratio, and drawdown. A comparative analysis was performed between the proposed DRL model and a traditional MPT optimization model using the same dataset and variables.

Findings

The proposed DRL model exhibited strong performance in portfolio selection, with a 15% increase in cumulative returns over the traditional Markowitz model. The net profit and return on assets of the optimized portfolio outperformed the benchmark across multiple test intervals. For example, the DRL model achieved an 8.41 return on the PPE variable, while the Markowitz model returned 8.40. Moreover, the DRL model demonstrated lower portfolio variance in critical stocks—such as 12% and 25% variance in the third and second companies, respectively—compared to the benchmark model.

In regression analysis, the DRL model achieved an R-squared value of 0.99971 in the combined test-validation-training phase, indicating excellent prediction accuracy. Mean squared error (MSE) and RMSE values across training, testing, and validation remained minimal, with the lowest RMSE observed at 0.15728 in the validation phase. Histogram error analysis revealed near-zero residuals across all phases, reinforcing the model's stability. Additionally, the DRL model achieved a liquidity index of 82% in the seventh company, outperforming traditional strategies in terms of asset liquidity.

Comparative risk-return metrics further emphasized the DRL model's robustness. For instance, the model yielded lower risk values (e.g., 0.7115 for return) while maintaining higher returns (e.g., 0.8402 for return) compared to the Markowitz approach. The dynamic portfolio forecasting capabilities of the model, leveraging both static and dynamic predictions, ensured more accurate future asset valuations. Finally, visualizations including regression plots and variance diagrams confirmed the superior alignment between model predictions and real market behavior.

Discussion and Conclusion

The findings of this study substantiate the growing consensus that integrating DRL with MPT can significantly enhance the performance of portfolio optimization strategies. The 15% increase in cumulative returns, along with superior metrics in variance reduction and profitability, confirms the added value of using intelligent, dynamic algorithms in financial decision-making. These results validate earlier studies that proposed DRL as a practical alternative to conventional optimization, especially in non-stationary and volatile market environments.

By employing a neural architecture based on the Levenberg–Marquardt algorithm, this study not only achieved high precision but also ensured computational efficiency in navigating local minima. The layered use of technical indicators and financial ratios enabled the agent to capture multifaceted market behaviors, providing a comprehensive decision-making framework. Furthermore, the model's ability to maintain high liquidity while optimizing returns aligns with real-world investment constraints, making it suitable for institutional application.

The comparative advantage of the proposed model over traditional MPT confirms the theoretical proposition that DRL can dynamically adjust strategies based on real-time feedback and evolving market states. The inclusion of transaction cost awareness and adaptive risk thresholds offers a pragmatic layer that is often missing in traditional models. Moreover, the architecture's success in processing tensor-based multi-asset data implies scalability to more complex portfolio settings.

In conclusion, this research highlights the transformative potential of DRL in portfolio management. It presents a viable path for future investment models that combine theoretical rigor with practical adaptability. As financial markets become more integrated with AI and big data systems, DRL is poised to play a central role in shaping investment strategies that are not only profit-maximizing but also resilient to structural and behavioral market shocks. This study provides a robust foundation for applying such models across emerging and developed markets alike, signaling a paradigm shift in how portfolios are constructed, maintained, and evolved.

Authors' Contributions

Authors equally contributed to this article.

Acknowledgments

Authors thank all participants who participate in this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

بهینه سازی سبدسهم با استفاده از یادگیری تقویتی عمیق و مبتنی بر نظریه مدرن پرتفوی



تاریخچه مقاله

تاریخ دریافت: ۵ آذر ۱۴۰۳

تاریخ بازنگری: ۲۸ دی ۱۴۰۳

تاریخ پذیرش: ۷ بهمن ۱۴۰۳

تاریخ انتشار: ۲۵ بهمن ۱۴۰۳

۱. یاسر رضائی پشته نوئی^۱: گروه حسابداری، واحد کاشان، دانشگاه آزاد اسلامی، کاشان، ایران

۲. مهدی صفری گرایلی^۲: گروه حسابداری، واحد بندرگز، دانشگاه آزاد اسلامی، بندرگز، ایران، ایمیل: safari83@iau.ac.ir (نویسنده مسئول)

۳. محمدتقی کبیری^۳: گروه حسابداری، واحد اراک، دانشگاه آزاد اسلامی، اراک، ایران

۴. میثم عرب زاده^۴: گروه حسابداری، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران

چکیده

این پژوهش به بررسی استفاده از یادگیری تقویتی عمیق در بهینه‌سازی سبد سرمایه‌گذاری می‌پردازد و به ادغام آن با تئوری مدرن سبد سرمایه‌گذاری پرداخته و تحولاتی را در استراتژی‌های مدیریت مالی نشان می‌دهد. در حالی که تئوری مدرن سبد سرمایه‌گذاری به عنوان یک چارچوب ریاضی برای حداکثرسازی بازده مورد انتظار با در نظر گرفتن ریسک شناخته می‌شود، نواقص آن، نظیر فرضیات مربوط به توزیع نرمال بازده‌ها و غفلت از هزینه‌های معاملاتی، نیاز به راه‌حل‌های انطباق‌پذیر را در بازارهای مالی پیچیده و پویا به وضوح نمایان می‌سازد. این پژوهش نشان می‌دهد که با بکارگیری یادگیری تقویتی عمیق، سرمایه‌گذاران قادر خواهند بود تا با بهره‌برداری از داده‌های لحظه‌ای و توانایی‌های تصمیم‌گیری پویا، استراتژی‌های سرمایه‌گذاری کارآمدتر و قدرتمندتری را توسعه دهند. همچنین، چالش‌هایی نظیر کیفیت داده، پیچیدگی محاسباتی و قابلیت تفسیر مدل‌های یادگیری تقویتی عمیق به صورت عمیق مورد بحث قرار گرفته است. در این پژوهش، یک مدل شبکه عصبی الگوریتم لونبرگ-ماد یادگیری تقویتی عمیق، برای بهینه‌سازی سبد سهام بر اساس داده‌های تاریخی پیشنهاد شده است. بدین منظور، از داده‌های ۱۰ شرکت پذیرفته شده در بورس اوراق بهادار تهران که دارای نقدینگی بالا در بازه زمانی ۱۳۹۰ تا ۱۴۰۰ بودند، استفاده شده است. نتایج به‌دست‌آمده از این مطالعه نشان می‌دهد که الگوریتم‌های یادگیری تقویتی عمیق در انتخاب پورتفوی می‌تواند به افزایش ۱۵ درصدی بازده تجمعی در مقایسه با روش‌های سنتی منجر شود. علاوه بر این، این مقاله به تحلیل گران و سرمایه‌گذاران توصیه می‌کند که از تکنیک‌های پیشرفته‌ای برای بهبود پایداری عملکرد شرکت‌ها استفاده کنند تا بتوانند تصمیمات بهتری در زمینه سرمایه‌گذاری اتخاذ نمایند. در نهایت، این تحقیق راهکارهایی برای پژوهش‌های آینده در این زمینه ارائه می‌دهد.

کلیدواژه‌گان: الگوریتم لونبرگ-ماد، بهینه‌سازی سبد سهام، نظریه مدرن پرتفوی و یادگیری تقویتی عمیق

شیوه استناددهی: رضائی پشته نوئی، یاسر، صفری گرایلی، مهدی، کبیری، محمدتقی، و عرب زاده، میثم. (۱۴۰۳). بهینه‌سازی سبدسهم با استفاده از یادگیری تقویتی عمیق و مبتنی بر نظریه مدرن پرتفوی. *حسابداری، امور مالی و هوش محاسباتی*، ۲(۴)، ۲۷۶-۲۹۳.



در دهه‌های اخیر، بهینه‌سازی سبد سرمایه‌گذاری به یکی از مهم‌ترین مسائل در حوزه مالی و مهندسی مالی تبدیل شده است. نظریه مدرن پرتفوی که نخستین بار توسط مارکوویتز ارائه شد، با فراهم کردن مدلی مبتنی بر میانگین و واریانس بازدهی دارایی‌ها، انقلابی در مدیریت سرمایه ایجاد کرد. این نظریه به‌صورت گسترده‌ای در دهه‌های اخیر در طراحی پرتفوی مورد استفاده قرار گرفته است، اما با گسترش بازارهای مالی پیچیده و ظهور فناوری‌های نوین، محدودیت‌های بنیادین این نظریه نظیر فرض توزیع نرمال بازده، ایستا بودن همبستگی بین دارایی‌ها، و نادیده‌گیری هزینه‌های معاملاتی، به چالش کشیده شده است. در پاسخ به این محدودیت‌ها، حوزه یادگیری ماشین و به‌ویژه یادگیری تقویتی عمیق (Deep Reinforcement Learning: DRL) به عنوان رویکردی نوین برای ارتقاء کارایی سبدهای سرمایه‌گذاری مطرح شده است (Goodell et al., 2021; Sutton & Barto, 2018). یادگیری تقویتی عمیق، از طریق تعامل مداوم با محیط و دریافت بازخورد از عملکرد گذشته، قادر است استراتژی‌هایی را یاد بگیرد که به بهینه‌سازی بازدهی و کاهش ریسک در شرایط بازار متغیر منجر شود (Mnih et al., 2015).

یکی از مزایای قابل توجه DRL در مقایسه با روش‌های سنتی، توانایی آن در به‌کارگیری داده‌های بلندمدت و ساختار نیافته نظیر داده‌های خبری و احساسات بازار است که در مدل‌های کلاسیک نادیده گرفته می‌شود (Henriques & Sadorsky, 2023; Lima Paiva et al., 2021). به‌عنوان مثال، مدل AlphaPortfolio که توسط کانگ و همکاران معرفی شد، به صورت مستقیم و با بهره‌گیری از الگوریتم‌های یادگیری تقویتی به ساخت سبدهای بهینه پرداخت و توانست عملکرد بهتری نسبت به روش‌های کلاسیک ارائه دهد (Cong et al., 2021). افزون بر این، مطالعات دیگری نظیر کار بنهامو و همکاران نشان داد که می‌توان فاصله بین برنامه‌ریزی مارکوویتز و یادگیری تقویتی عمیق را با استفاده از چارچوب‌های ریاضیاتی و مدل‌های یادگیری هوشمند پر کرد و به عملکردی متوازن بین ریسک و بازده دست یافت (Benhamou et al., 2020). در همین راستا، تلاش‌هایی برای ترکیب مدل‌های میانگین-واریانس با الگوریتم‌های DRL به منظور ایجاد تابع پاداشی تطبیق‌یافته با ترجیحات سرمایه‌گذار صورت گرفته است (Deng et al., 2016).

از منظر ساختار الگوریتمی، پژوهش‌هایی مانند مطالعه جیانگ و همکاران با طراحی چارچوب DRL برای مدیریت پرتفوی، نشان دادند که استفاده از شبکه‌های عصبی بازگشتی و الگوریتم‌های یادگیری سیاست می‌تواند نتایج به مراتب بهتری از نظر بازده تجمعی و کنترل ریسک به همراه داشته باشد (Jiang et al., 2017). همچنین، الگوریتم‌های مبتنی بر یادگیری Q یا الگوریتم‌های سیاست قطعی، مانند DDPG و TD³، در طراحی عامل‌های هوشمند که قادر به تصمیم‌گیری در شرایط عدم قطعیت هستند، نقش بسزایی داشته‌اند (Koratamaddi et al., 2021; Yifu et al., 2024). از سوی دیگر، مطالعاتی همچون تحقیق چاها، با استفاده از رتبه‌بندی و تطابق سهام از طریق یادگیری تقویتی، توانستند ساختار سبدي را طراحی کنند که در برابر نوسانات شدید بازار نیز مقاوم باقی بماند (Chaher, 2025). چنین چارچوب‌هایی، امکان اعمال معیارهای غیرخطی و غیرایستا را برای ساخت سبدهای حساس به ریسک و هزینه فراهم کرده‌اند.

با وجود این دستاوردها، کاربرد عملی DRL در بازارهای مالی همچنان با چالش‌هایی روبروست. یکی از مهم‌ترین مسائل، کیفیت و ساختار داده‌هاست. مطالعات متعددی تأکید دارند که یادگیری تقویتی برای دستیابی به پایداری و همگرایی، نیازمند داده‌های با حجم بالا و ساختار یافته می‌باشد، حال آنکه داده‌های مالی معمولاً با مشکل نویز، ناهمزمانی و عدم ایستایی مواجه‌اند (Mohammadi Baghmalaei et al., 2021; Moshrefi & Behnamian, 2022). افزون بر این، تفسیرپذیری پایین مدل‌های یادگیری تقویتی به‌عنوان مانعی در پذیرش آن‌ها در صنعت مطرح است؛ زیرا خروجی‌های این مدل‌ها غالباً به‌صورت جعبه سیاه باقی می‌ماند (Almahdi & Yang, 2017). راه‌حل‌هایی مانند استفاده از الگوریتم‌های ترکیبی شامل شبکه‌های Q عمیق همراه با مکانیسم‌های توجه یا به‌کارگیری معماری‌های هیبریدی CNN-RNN به منظور تقویت تفسیرپذیری مدل‌ها، توسط پژوهشگرانی نظیر لیانگ و همکاران پیشنهاد شده‌اند (Liang et al., 2018).

مطالعات متا-تحلیلی اخیر نیز نشان می‌دهند که الگوریتم‌های یادگیری ماشین، چه در دسته‌بندی نظارت‌شده و چه نظارت‌نشده، می‌توانند به بهبود دقت غربال‌گری سهام، تنظیم مجدد سبد و طراحی استراتژی‌های واکنش‌پذیر به بحران‌های مالی نظیر کووید-۱۹ کمک کنند (Ayari Salah, 2025). این دستاوردها باعث شکل‌گیری چارچوب‌های

چندعاملی در سیدگردانی شده‌اند، که در آن‌ها داده‌های سری زمانی قیمت سهام، شاخص‌های مالی و ویژگی‌های بنیادی شرکت‌ها به‌صورت هم‌افزا مورد تحلیل قرار می‌گیرند. به‌عنوان مثال، مدل CMPS با ترکیب چند عامل و بهره‌گیری از شبکه توجه، توانسته است سبدهایی با بازدهی بالاتر نسبت به مدل‌های متداول ارائه دهد (Yifu et al., 2024). در حوزه عملی نیز مطالعاتی مانند پژوهش جیوتیرمای و همکاران، با استفاده از ترکیب مدل‌های سری زمانی ARIMA و الگوریتم LS-SVM، به تحلیل رفتار غیرخطی و غیرایستا بازارهای سهام پرداختند و نشان دادند که مدل‌های ترکیبی می‌توانند عملکرد بهتری نسبت به رویکردهای سنتی داشته باشند (Jyotirmayee Behera, 2025). علاوه بر این، نتایج به‌دست‌آمده از پیاده‌سازی الگوریتم‌های مبتنی بر یادگیری تقویتی در بورس‌های نوظهور، مانند بورس تهران، نیز نشان می‌دهند که بهره‌گیری از الگوریتم لونبرگ-مارد در کنار ساختارهای عصبی بهینه، می‌تواند به بهبود پیش‌بینی قیمت‌ها و تشکیل سبدهای با نقدشوندگی بالا منجر شود (Mostafaei Darmian & Doaei, 2021). این یافته‌ها، در کنار پژوهش‌های تجربی مشابه در بازارهای جهانی، گویای آن است که استفاده از DRL در کنار نظریه مدرن پرتفوی نه تنها قابل پیاده‌سازی است بلکه می‌تواند سودآوری و پایداری استراتژی‌های سرمایه‌گذاری را بهبود بخشد. بنابراین، پژوهش حاضر با هدف بهینه‌سازی سبدسهم با استفاده از یادگیری تقویتی عمیق و مبتنی بر نظریه مدرن پرتفوی انجام شد.

روش پژوهش و مواد

روش تحقیق در این مطالعه به بررسی و پیاده‌سازی الگوریتم‌های یادگیری تقویتی عمیق برای بهینه‌سازی سبد سهام خواهد پرداخت. این بخش به تشریح مراحل مختلف تحقیق، از جمع‌آوری داده‌ها تا ارزیابی و تحلیل نتایج می‌پردازد. مرحله نخست، جمع‌آوری داده‌های مربوط به بازارهای مالی، شامل قیمت‌های تاریخی دارایی‌ها، حجم معاملات و داده‌های اقتصادی کلان است. در مرحله پیش‌پردازش، داده‌هایی که جمع‌آوری شده‌اند، باید پاک‌سازی و آماده‌سازی شوند. این شامل مراحل زیر است: حذف داده‌های گمشده یا مشکل‌دار، نرمال‌سازی داده‌ها برای مقایسه بهتر بین دارایی‌ها و ایجاد ویژگی‌های جدید مانند بازده روزانه، ریسک (انحراف معیار) و نسبت شارپ برای افزودن به اطلاعات موجود. مدل‌های یادگیری تقویتی به صورت خاص برای بهینه‌سازی سبد سهام طراحی و پیاده‌سازی خواهند شد. این مراحل شامل: انتخاب الگوریتم یادگیری تقویتی: انتخاب الگوریتم‌هایی مانند Q-Learning، DDPG (Deep Deterministic Policy Gradient) یا PPO (Proximal Policy Optimization) که تطابق بیشتری با هدف بهینه‌سازی سبد سهام داشته باشند. طراحی تابع پاداش: تعریف تابع پاداش که متناسب با هدف اصلی مقاله، یعنی حداکثر کردن بازدهی و حداقل کردن ریسک، طراحی خواهد شد. تعریف محیط یادگیری: محیط شامل تمام دارایی‌های موجود و شرایط بازار است که عامل تصمیم‌گیری باید در آن عمل کند. مدل‌های یادگیری تقویتی آموزش دیده خواهند شد تا استراتژی‌های بهینه‌سازی سبد سهام را از طریق تکرار چندین دوره آموزشی شناسایی کنند. این فرآیند شامل: تقسیم داده‌ها به مجموعه آموزشی و تستی. انجام تمرین بر روی داده‌های آموزشی به منظور یادگیری از تجربیات و اصلاح استراتژی‌ها. نظارت بر عملکرد مدل با استفاده از معیارهای ارزیابی مناسب، مانند cumulative returns و volatility. پس از آموزش مدل، مراحل آزمایش و ارزیابی آغاز خواهد شد: آزمون مدل بر روی داده‌های تست برای بررسی عملکرد آن در شرایط واقعی. مقایسه عملکرد مدل یادگیری تقویتی با سایر استراتژی‌های بهینه‌سازی سبد، از جمله روش‌های کلاسیک و الگوریتم‌های یادگیری ماشین دیگر. تجزیه و تحلیل نتایج با استفاده از معیارهای مربوط به عملکرد سبد، مانند نسبت شارپ، Max Drawdown و Alpha و Beta. در این مرحله، نتایج حاصله از آزمون‌های مختلف مورد تجزیه و تحلیل و نقد قرار خواهند گرفت. تجزیه و تحلیل عملکرد مدل از جنبه‌های مختلفی نظیر ثبات، کارایی و پاسخگویی به تغییرات بازار خواهد بود. همچنین، بررسی نقاط قوت و ضعف مدل‌های به کار رفته و پیشنهادات برای بهینه‌سازی‌های آتی نیز انجام خواهد شد. نتیجه‌گیری نهایی بر اساس تحلیل‌های انجام شده ارائه خواهد شد که شامل نکات کلیدی در مورد کارایی مدل‌های یادگیری تقویتی عمیق برای بهینه‌سازی سبد سهام و راهکارهای آینده در این حوزه خواهد بود. این نتیجه‌گیری‌ها می‌توانند به عنوان مبنایی برای تحقیقات آینده و به کارگیری این روش‌ها در صنعت مالی مورد استفاده قرار گیرند.

متغیرهای این پژوهش که به عنوان متغیرهای تعدیلگر به منظور ارزیابی روش بهینه‌سازی یادگیری تقویتی عمیق مورد ارزیابی قرار گرفته است در ادامه بیان شده اند: متغیر بازده: به سود و زیان ناشی از سرمایه‌گذاری بازده گفته می‌شود. بازدهی مثبت نشان دهنده سود و بازدهی منفی نشان دهنده زیان است. متغیر نرخ بازگشت سرمایه (ROA): نرخ بازگشت سرمایه به صورت نسبت سود حاصله به میزان هزینه و سرمایه پرداخت شده محاسبه می‌شود. سود بدست آمده ممکن است تنها جنبه مالی نداشته باشد و شامل موارد

غیر مالی مانند برندسازی نیز بشود. متغیر رشد فروش: نرخ رشد فروش، پارامتری است برای اندازه گیری کارایی تیم فروش یک شرکت، به منظور افزایش درآمد حاصل از فروش در یک بازه ی زمانی مشخص شده. اگر نرخ رشد فروش در مقایسه با دوره های قبل کاهش داشته باشد، میتواند نشان دهنده ی این موضوع باشد که تیم فروش باید روش دیگری را برای رشد درآمد خود در پیش بگیرد. متغیر خالص ارزش دارایی های ثابت (PPE): ارزش خالص دارایی، شاخصی در بورس می باشد که برای ارزشیابی و برآورد ارزش ذاتی دارایی های شرکت سرمایه گذاری، استفاده می شود و از طریق کسر بدهی، از ارزش روز آن شرکت، به دست می آید. دارایی ثابت یک ملک و یا تجهیزات محسوس و قابل لمس است که یک شرکت در اختیار دارد و میتواند در عملیات خود برای درآمدزایی از آن استفاده کند. متغیر هزینه مالیات (CTE): از منظر قانون مالیات، هزینه های صورت گرفته در واحدهای اقتصادی به دو دسته ی هزینه های قابل قبول و هزینه های غیر قابل قبول تقسیم بندی می شوند. هزینه های قابل قبول هزینه هایی هستند که در زمان رسیدگی مالیاتی توسط کارشناسان مالیات مورد قبول واقع می شوند و مبلغ آن ها از درآمد مودی قابل کسر است. متغیر درآمد قبل از کسر مالیات (PTI): سودآوری و درآمد حاصله توسط شرکت قبل از کسر هزینه های قابل قبول مالیاتی گفته می شود. متغیر سود خالص (PTI-CTE): سود به عنوان یکی از مهم ترین عوامل موثر بر تصمیم گیری های اقتصادی، با توجه به ارقام صورت های مالی در نظر گرفته شده است. در تئوری های اقتصادی، ارزش هر شرکت مبتنی بر ارزش فعلی جریان های نقدی آتی آن است و از سود به عنوان جانشین جریان های نقدی استفاده می شود.

جامعه آماری این پژوهش شامل کلیه شرکت های پذیرفته شده در بورس اوراق بهادار تهران است. در این پژوهش برای انتخاب نمونه آماری از روش حذف سیستماتیک استفاده خواهد شد. در واقع نمونه آماری شامل کلیه شرکت هایی است که مجموعه ویژگی های زیر را دارا باشند: قبل از فروردین ۱۳۹۰ در بورس پذیرفته شده باشند و تا پایان سال ۱۴۰۰ نماد آن ها متوقف نشده باشد، برای رعایت قابلیت مقایسه پذیری، دوره مالی آن ها منتهی به پایان اسفند باشد، در طی دوره مورد بررسی توقف فعالیت یا تغییر سال مالی نداشته باشند، اطلاعات مورد نیاز شرکت ها، در طی دوره مورد بررسی موجود باشد.

برخی از مهم ترین مواردی که طی فرایند پیش پردازش داده ها باید به آن ها پرداخته شود، در ادامه بیان شده اند: **داده های ناموجود، داده های پرت، نرمال سازی داده ها.** تحلیل داده های ناموجود: در برخی موارد، ممکن است بعضی از ویژگی های مربوط به یک یا چند نمونه، فاقد مقادیر معتبر باشند. برای پردازش مجموعه داده هایی که دچار چالش داده ناموجود هستند، باید راهکاری برای تعیین مقادیر داده های ناموجود یافت. روش های گوناگونی برای مدیریت داده های ناموجود وجود دارند که در این پژوهش از روش مقداردهی با میانگین و یا محتمل ترین مقدار استفاده شده است (پنگ و همکاران، ۲۰۲۱). استفاده از مقدار میانگین ویژگی: در این روش برای هر ویژگی، مقدار میانگین، با استفاده از نمونه هایی که مقادیرشان معلوم است محاسبه شده و برای جاگذاری مقادیر مربوط به آن ویژگی مورد استفاده قرار می گیرد. استفاده از محتمل ترین مقدار: اساس این روش فراوانی مقادیر موجود در هر ویژگی است. در این روش، فراوانی مقادیر ظاهر شده در هر ویژگی به دست می آید و مقداری که بالاترین فراوانی را در بین مقادیر دارد به عنوان مقدار جایگزین انتخاب می شود. این مقدار منتخب، جایگزین داده های مفقود در آن ویژگی می شود. تحلیل داده های پرت: در برخی موارد، ممکن است به دلایلی، مقداری در مجموعه داده ظاهر شوند که تفاوت زیاد و غیر معمولی با سایر مقادیر موجود در مجموعه داشته باشند، این داده ها را داده های پرت می گویند. روش های مختلفی برای شناسایی و مدیریت این داده ها وجود دارد که در این پژوهش از روش های حذف و جایگذاری با میانگین استفاده شده است. حذف: ساده ترین روش مدیریت داده های پرت، حذف آن ها است. برای این منظور ابتدا لازم است داده های پرت شناسایی شوند. روش متداولی که مورد استفاده قرار می گیرد، به این صورت است که ابتدا بازه ای را که داده ها در آن قرار دارند مشخص می کنند، سپس تعداد معینی از داده های بالا و پایین این بازه را به عنوان داده پرت در نظر گرفته و آن ها را از مجموعه داده حذف می کنند. جایگذاری با میانگین: یک روش دیگر که در مقایسه با روش قبلی دقت مناسب تری دارد، استفاده از مقدار میانگین داده ها است. همسان سازی داده ها: به منظور جلوگیری از مشکل پراکندگی و قرار گرفتن داده ها در یک بازه مشخص، معمولاً داده های ورودی قبل از انتقال به شبکه مدل شده، نرمالیزه می شوند. براین اساس، مهمترین نکته در انطباق داده ها، نرمالیزه کردن پارامترهای ورودی می باشد که از طریق رابطه ی زیر حاصل می شود (ما و همکاران، ۲۰۲۱).
$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$
 که در این رابطه x_i پارامتر ورودی، x_{min} کمترین مقدار پارامتر x ، x_{max} بیشترین مقدار پارامتر x ، x' هم مقدار نرمال شده پارامتر x_i می باشد. محاسبه شباهت ویژگی ها: مسئله به صورت یک گراف کامل وزن دار بدون جهت $G = (F, E, W_F)$ بازنمایی می شود که $F = \{F_1, F_2, \dots, F_n\}$ نشان دهنده مجموعه ویژگی های اولیه با n ویژگی است که هر ویژگی یک گره از گراف را تشکیل می دهد و E

$\{(F_i, F_j): F_i, F_j \in F\}$ یال‌های گراف را نشان می‌دهد. همچنین $w_F: (F_i, F_j) \rightarrow R$ یک تابع است که میزان شباهت بین دو ویژگی F_i و F_j را نشان می‌دهد (افیس و همکاران، ۲۰۲۰). انتخاب یک معیار مناسب برای محاسبه شباهت بین ویژگی تأثیر زیادی بر عملکرد الگوریتم انتخاب ویژگی دارد. روش‌های مختلفی بر محاسبه شباهت بین ویژگی‌ها وجود دارد که هر کدام از آن‌ها نتایج متفاوتی را به همراه دارند. به همین علت، انتخاب یک معیار مناسب برای محاسبه شباهت بین ویژگی‌ها اهمیت ویژه‌ای دارد. به‌طور کلی، فاصله اقلیدسی، شباهت کسینوسی و همچنین ضریب همبستگی پیرسون از مهم‌ترین معیارها برای اندازه‌گیری شباهت بین ویژگی‌ها است. در این پژوهش، برای محاسبه شباهت مابین ویژگی‌ها، از قدر مطلق ضریب همبستگی پیرسون استفاده شده است. ضریب همبستگی پیرسون بین دو ویژگی F_i و F_j مطابق رابطه ۲ محاسبه شده است.

$$w_{ij} = \left| \frac{\sum_p (x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{\sum_p (x_i - \bar{x}_i)^2} \sqrt{\sum_p (x_j - \bar{x}_j)^2}} \right|$$

(رابطه ۲)، که در آن x_i و x_j به ترتیب نشان‌دهنده بردار ویژگی‌های F_i و F_j است. همچنین، متغیرهای x_i و x_j به ترتیب نشان‌دهنده میانگین مقادیر برای بردار x_i و x_j بین p نمونه است. با توجه به رابطه ۱، می‌توان فهمید که مقدار شباهت مابین دو ویژگی که کاملاً شبیه هم هستند برابر با یک و مقدار شباهت مابین دو ویژگی که کاملاً غیر شبیه به هم هستند برابر صفر است. در بسیاری از شرایط ممکن است که مقدار شباهت محاسبه شده برای ویژگی‌های مختلف نزدیک به هم باشد. برای رفع این مشکل، به‌منظور نرمال‌سازی مقادیر شباهت محاسبه‌شده، از تکنیک مقیاس‌گذاری غیرخطی استفاده می‌شود. با استفاده از این تکنیک، تمام مقادیر شباهت محاسبه‌شده در بازه صفر تا یک نرمال می‌شوند. نرمال‌سازی شباهت بین ویژگی‌ها با استفاده از رابطه ۲ انجام شده است.

$$\hat{w}_{ij} = \frac{1}{1 + \exp\left(-\frac{w_{ij} - \bar{w}}{\sigma}\right)}$$

(رابطه ۳)، که در این رابطه w_{ij} شباهت بین ویژگی F_i و F_j است، $\bar{w}$ و σ به ترتیب میانگین و انحراف معیار برای تمام شباهت‌های محاسبه‌شده بین تمام ویژگی‌ها را نشان می‌دهد.

شاخص‌های فنی: شاخص‌های فنی وضعیت فعلی بازار را توصیف می‌کنند و اطلاعاتی را در مورد روندهای گذشته در خود جای داده اند. یک شاخص را می‌توان به عنوان تصویری از وضعیت فعلی در نظر گرفت که رفتار گذشته را در یک دوره معین به حساب می‌آورد. در بین شاخص‌های فنی از سه شاخص میانگین متحرک (MA)، شاخص قدرت نسبی (RSI) و میانگین متحرک همگرایی (MACD) استفاده شده است. شاخص MA به عنوان یک نشانگر روند شناخته می‌شود که میانگین قیمت را در یک بازه ی زمانی محاسبه می‌کند.

$$MA_w(t) = \frac{\sum_{i=0}^{w-1} P_{t-i}}{w} \quad \text{رابطه ۴، که در این رابطه } P_t \text{ به عنوان قیمت نهایی در پایان روز } t \text{ و } w \text{ به عنوان پنجره زمانی می‌باشند.}$$

$$RSI_t(w) = 100 * \left(1 - \frac{1}{1 + \frac{EMA_w((P_t - P_{t-1}, 0))}{EMA_w(\min(P_t - P_{t-1}, 0))}} \right) \quad \text{رابطه ۵}$$

$$EMA_w(t) = \alpha * P(t) + (1 - \alpha) * EMA_w(t - 1), \alpha = \frac{2}{1 + w} \quad \text{رابطه ۶}$$

همچنین شاخص MACD، یک شاخص فنی می‌باشد که رابطه ی بین دو EMA در دو بازه ی زمانی مختلف را نمایش می‌دهد.

$$MACD_w(t) = \sum_{i=1}^w EMA_k(i) - \sum_{i=1}^w EMA_d(i) \quad \text{رابطه ۷}$$

بهینه سازی: در بهینه سازی سبدسهم با استفاده از یادگیری تقویتی عمیق و مبتنی بر نظریه مدرن پرتفوی، از یک بردار چهار بعدی V_t و یک ماتریس همبستگی چهار بعدی Cor_t متشکل از شاخص‌های فنی روزانه که از n دارایی به عنوان ورودی تشکیل شده، استفاده شده است.

$$V_t = [V_{t,close}, V_{t,MA}, V_{t,RSI}, V_{t,MACD}] \quad \text{رابطه ۸}$$

$$Cor_t = [Cor_{t,close}, Cor_{t,MA}, Cor_{t,RSI}, Cor_{t,MACD}] \quad \text{رابطه ۹}$$

که در این رابطه V_t یک تانسور متشکل از اندیکاتورهای فنی با دوره نسبتاً کوتاه برای m روز در نظر گرفته شده می‌باشد. Cor_t رابطه ی بین n دارایی در یک سال می‌باشد. از تبدیل زیر برای ترکیب رابطه بین دو تانسور استفاده می‌شود. معادله ی زیر بر مبنای ترکیب تانسورها با دو ویژگی مختلف برای اندیکاتور i و دارایی n در نظر گرفته شده است:

$$F_{t(batch,i,n)} = V_{t(batch,i,n)} \odot Cor_{t(batch,i,n)}^T, F_{t(batch,i,n)} \in R^{m*n} \quad \text{رابطه ۱۰}$$

برای هر دارایی j ، اندیکاتور i به معنای لایه‌ای در یادگیری تقویتی عمیق می‌باشد که فقط قسمت مربوط به دارایی j را برای تانسور $m*n$ مورد ارزیابی قرار می‌دهد. تانسور تبدیل شده $F_t \in R^{4*n*m*n}$ فرآیند ساخت چهار نشانگر را به ۳۲ ویژگی از طریق لایه شبکه عصبی کانولوشن سه بعدی (۱,۳,۱) را انجام می‌دهد.

$$F'_t = Conv3D(F_t), F'_t \in R^{32*n*(m-4)*n} \quad \text{رابطه ۱۱}$$

پیش بینی پویا و ایستا: در صورتی که متغیر وابسته با وقفه در سمت راست معادله ظاهر شود، عمل پیش بینی دشوار خواهد بود. بنابراین این سوال مطرح می‌شود که چگونه باید ارزش متغیر وابسته با وقفه که در سمت راست معادله ظاهر شده است را ارزیابی نمود. در این صورت دو روش برای پیش بینی وجود دارد: پیش بینی پویا و پیش بینی ایستا. در پیش بینی پویا (چند گام به جلو) از مقادیر برآورد شده ولی در پیش بینی ایستا (یک گام به جلو) از مقادیر واقعی متغیر استفاده می‌شود، در نتیجه پیش بینی ایستا دقیق تر از پیش بینی پویا خواهد بود. بررسی میزان دقت پیش بینی: پس از تخمین مدل، برای ارزیابی اجرای مدل‌های رقیب، باید قدرت پیش بینی مدل‌ها در خارج از نمونه را بررسی کرد. برای این منظور معمولاً، داده‌ها را به دو مجموعه جدا تقسیم می‌کنند. بخش اول به مجموعه آموزش یا تخمین و مجموعه دوم به مجموعه آزمون موسوم است. ابتدا ضرایب مدل یا استفاده از داده‌های مجموعه اول تخمین زده می‌شود و سپس، با استفاده از داده‌های مجموعه دوم قدرت پیش بینی مدل و یا به عبارت دیگر، توان تعمیم مدل به خارج از مجموعه داده‌های مورد استفاده در تخمین ارزیابی می‌شود. در این قسمت هدف حداقل کردن خطای پیش بینی در مجموعه آزمون است. برای این منظور معمولاً از چهار معیار زیر استفاده می‌شود: الف) معیار میانگین مربع خطا MSE یا ریشه میانگین مربع خطا RMSE

$$MSE = \frac{\sum (\hat{y}_t - y_t)^2}{n} \quad \text{رابطه ۱۲}$$

$$RMSE = \sqrt{\frac{\sum (\hat{y}_t - y_t)^2}{n}} \quad \text{رابطه ۱۳}$$

ب) معیار میانگین قدرمطلق انحراف یا میانگین قدرمطلق درصد خطا

$$MAD = \frac{\sum |\hat{y}_t - y_t|}{n}$$

$$MAPE = \frac{\sum \left| \frac{\hat{y}_t - y_t}{y_t} \right|}{n} \quad \text{رابطه ۱۴}$$

ج) آماره U تایل: آماره تایل، یک معیار اندازه گیری بر اساس RMSE است. آماره U ، به صورت نسبت ریشه میانگین مجذور خطای یک مدل ابتدایی که می‌پذیرد ارزش متغیر مورد پیش بینی تغییر نخواهد کرد، محاسبه می‌شود. بنابراین، پیش بینی در دوره بعد به صورت ساده، ارزش واقعی دوره قبل را اختیار می‌کند.

¹ Theil U Statistic

$$U = \frac{\sqrt{\frac{\sum (\hat{y}_t - y_t)^2}{n}}}{\sqrt{\frac{\sum (y_{t-1} - y_t)^2}{n}}} \quad \text{رابطه ۱۵}$$

هر اندازه آماره U به صفر نزدیک تر باشد، بهتر است. معمولاً ارزش‌های کوچکتر از ۰/۵۵ مورد قبول واقع می‌شوند. مدلهایی که آماره U آن‌ها بزرگتر از یک باشد رد می‌شوند، زیرا در این حالت پیش بینی مدل از یک مدل ابتدایی نیز بدتر بوده است. (د) نرخ اغتشاش: این معیار، درستی جهت پیش بینی‌ها را به وسیله محاسبه نسبت تعداد نا صحیح جهت پیش بینی‌ها به تعداد کل پیش بینی‌ها اندازه می‌گیرد. مقدار این معیار بین صفر و یک است. اگر تمام پیش بینی‌ها صحیح باشد، اندازه معیار برابر صفر و اگر همه آن‌ها غلط باشد، برابر یک خواهد بود.

یافته‌ها

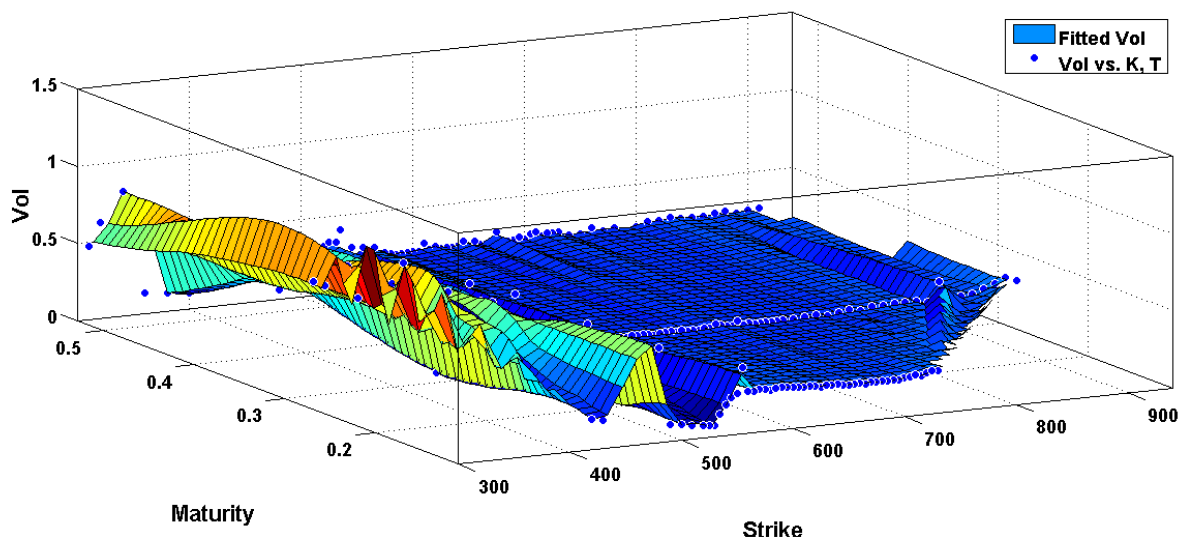
در این بخش ابتدا، پس از شرح معیارهای ارزیابی به کار گرفته شده به منظور انتخاب سبد بهینه سهام با استفاده از روش پیشنهادی با تاکید بر افزایش ضریب نقد شوندگی، روش پیشنهادی ارائه شده با سایر روش‌های جدید موجود مقایسه شده است. علاوه بر این در روش پیشنهادی، میزان حساسیت پارامترهای استفاده شده مورد ارزیابی قرار گرفته و پارامترهای مورد نظر با مقادیر متفاوت به منظور دستیابی به مقادیر آستانه بهینه ارزیابی شده‌اند. همچنین، عملکرد و پیچیدگی محاسباتی هر یک از روش‌های پیشنهادی را نسبت به سایر روش‌ها مورد بررسی قرار گرفته‌اند.

خلاصه‌ای از وضعیت سهام‌های مورد مطالعه در جدول ۱ نمایش داده شده است.

جدول ۱: خلاصه‌ای از وضعیت سهام شرکت‌های پتروشیمی مورد مطالعه

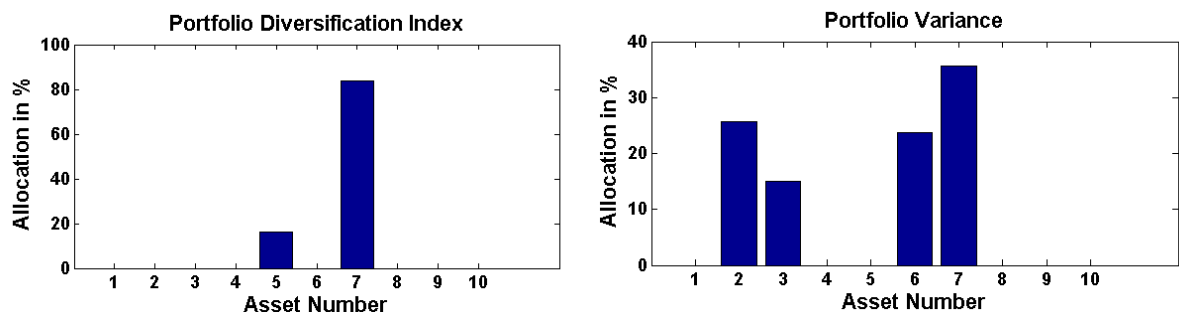
ردیف	شرکت‌های بورسی	سال	فروش سال	فروش سال قبل	تغییرات فروش	خالص ارزش دارایی‌های ثابت	مجموع دارایی‌ها	کل بازده دارایی‌ها
۱	پالایش نفت اصفهان	۱۳۸۷	۲۴۲۹۰۱	۱۸۱۶۵۳	۶۱۲۴۸	۴۱۹۷۵	۲۲۴۹۲۴	۰.۱۴۷۸۴۱۰۴۹
		۱۴۰۰	۷۴۰۸۳۴۲	۴۸۹۲۴۶۱	۲۵۱۵۸۸۱	۱۰۶۵۸۱۸	۱۰۹۰۵۲۱۴	۰.۲۲۸۷۵۹۹۳۱
۲	الکتریک خودروشرقی	۱۳۸۷	۷۹۵۷۴۰	۶۹۲۹۵۹	۱۰۲۷۸۱	۸۲۸۲۷	۶۰۸۹۸۴	۰.۱۹۴۷۰۶۲۶۵
		۱۴۰۰	۴۹۹۷۵۲۶	۳۱۲۹۳۵۶	۱۸۶۸۱۷۰	۸۱۵۶۵	۳۵۲۱۸۲۸	۰.۰۰۲۵۴۶۱۲۱
۳	صنعتی آما	۱۳۸۷	۲۶۴۱۷۳	۲۵۹۲۱۰	۴۹۶۳	۱۱۷۸۳۵	۳۹۵۱۴۳	۰.۰۹۱۸۰۷۲۷
		۱۴۰۰	۲۸۲۰۸۹۳	۱۸۲۷۵۶۹	۹۹۳۳۲۴	۴۷۲۵۰۶	۲۷۴۸۹۸۴	۰.۲۷۸۸۲۱۵۵۷
۴	ایران تایر	۱۳۸۷	۷۸۹۱۰۵	۶۱۷۹۴۳	۱۷۱۱۶۲	۶۸۴۱۰	۶۰۹۴۸۱	۰.۱۶۰۵۸۴۱۶۹
		۱۴۰۰	۵۷۱۴۱۲۸	۳۲۶۳۷۰۰	۲۴۵۰۴۲۸	۶۹۹۹۹۱	۳۵۰۳۵۲۱	۰.۲۵۹۴۹۱۸۰۸
۵	ایران ترانسفو	۱۳۸۷	۱۶۱۸۰۵۰	۱۱۸۸۱۴۸	۴۲۹۹۰۲	۲۲۴۲۸۱	۳۶۹۰۸۹۴	۰.۰۶۷۸۲۱۷۸
		۱۴۰۰	۹۷۸۲۶۴۲۰	۴۶۱۱۸۶۲	۵۱۷۰۷۸۰	۶۵۹۶۳۶	۲۷۰۸۹۹۱۳	۰.۱۴۳۰۱۴۳۷۶
۶	ایران خودرو	۱۳۸۷	۴۴۶۲۳۹۱۶	۴۵۰۳۹۱۱۶	-۴۱۵۲۰۰	۱۶۴۸۷۳۳۷	۶۴۷۶۶۶۰۰	۰.۰۳۵۰۸۴۸۸۹
		۱۴۰۰	۳۳۱۷۸۴۳۱	۲۰۴۲۴۰۹۲۲	۱۲۸۹۳۷۵۰۹	۱۸۳۲۲۸۵۴۳	۵۵۸۳۹۱۹۷۵	-۰.۰۹۴۲۰۰۴۱۵
۷	سیمان فارس	۱۳۸۷	۱۲۳۸۵۳	۱۱۰۵۴۹	۱۳۳۰۴	۱۲۲۲۵	۱۱۵۰۲۷	۰.۲۳۵۲۳۱۷۲۸
		۱۴۰۰	۷۹۸۸۴۳۶	۷۳۵۷۱۶	۶۴۸۳۱	۱۴۴۲۹۹	۱۵۳۱۹۳۲	۰.۰۶۹۹۸۸۳۹۸
۸	ایران مرینوس	۱۳۸۷	۷۴۶۳۱	۸۸۳۶۰	-۱۳۷۲۹	۸۰۵۱۸	۱۵۲۹۷۵	۰.۰۴۶۸۷۰۴۰۴
		۱۴۰۰	۲۷۰۴۹۸۰	۱۷۴۳۵۴	۹۶۱۴۴	۹۸۲۹۳	۳۹۱۵۷۹	۰.۰۸۷۲۷۹۹۶۱
۹	ایر کارپارت صنعت	۱۳۸۷	۱۷۲۶۸۱	۱۷۸۳۴۰	-۵۶۵۹	۵۱۳۵۶	۱۶۷۶۷۶	۰.۰۹۱۳۳۰۹
		۱۴۰۰	۷۵۱۴۰۸	۳۹۵۳۲۹	۳۵۶۰۷۹	۲۴۵۲۹۴	۱۲۲۵۳۲۵	۰.۲۴۵۱۹۶۹۸۹
۱۰	آبسان	۱۳۸۷	۵۸۱۶۸۲	۴۸۴۸۰۹	۹۶۸۷۳	۱۵۴۵۶۹	۶۹۷۷۳۸	۰.۰۱۰۰۷۸۸۵۴۸
		۱۴۰۰	۱۷۳۳۴۲۲	۱۲۵۱۶۴۲	۴۸۱۷۸۰	۲۳۰۷۹۶	۱۸۴۷۴۶۷	۰.۱۱۱۸۹۷۵۳۳

مدل اولیه الگوریتم در شکل ۱ نمایش داده شده، میزان ناحیه بهینه شده در قسمت‌های زرد رنگ و قرمز رنگ مشخص است. تابع هزینه در نظر گرفته شده بر مبنای تکرار فرآیند به منظور بهینه سازی میزان رشد فروش و افزایش نقدینگی ناشی از کسب سود است.

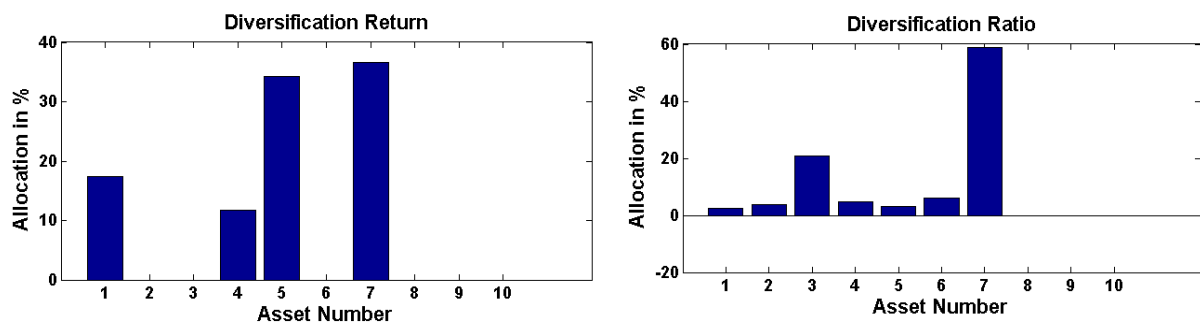


شکل ۱: مدل اولیه ی ناشی از الگوریتم پیشنهادی

میزان واریانس پرتفوی سهامداران در شرکت‌های مورد ارزیابی و ضریب شاخص پرتفوی سهامداران بر مبنای ضریب نقد شوندگی در نمودار (۴-۱۲) نمایش داده شده است. میزان واریانس سهام شرکت‌های دوم در محدود ۲۵ درصد، شرکت سوم در محدوده ی ۱۲ درصد و شرکت ششم و هفتم در محدوده ی ۲۲ درصد و ۳۵ درصد می‌باشد. بیشترین میزان ضریب نقدشوندگی شرکت هفتم در محدوده ی ۸۲ درصد قرار دارد.



شکل ۲: میزان واریانس پرتفوی سهامداران در شرکت‌های مورد مطالعه و شاخص پرتفوی سهامداران بر مبنای ضریب نقدشوندگی



شکل ۳: میزان نرخ بازگشت سرمایه بر مبنای سودآوری در حالت کلی

نتایج بیان کننده میزان انتظار بازده سرمایه‌گذاری و ریسک ناشی از واریانس آن در جدول ۲ نمایش داده شده است.

جدول ۲: مقایسه دو پرتفوی به دست آمده از یادگیری تقویتی عمیق و مبتنی بر نظریه مدرن پرتفوی

الگوریتم گرده افشانی	مدل مارکوویتز	مدل	مدل
بازده	ریسک	بازده	ریسک
۰.۸۴۰۲	۰.۷۱۱۵	۰.۸۳۳۱	۱.۱۲۱
۰.۸۱۲۱	۰.۸۰۰۱	۱.۰۱۳	۱.۱۱۸
۰.۸۴۱۲	۰.۸۳۱۴	۰.۸۳۴۱	۰.۸۳۳
۰.۸۴۱۵۱	۱.۶۱۱۵	۸.۴۰۱۵	۱.۶۲۲
۰.۸۳۲۱	۰.۰۸۱	۰.۸۱۱۲	۱.۱۰۵
۰.۸۶	۰.۰۸۳۴	۰.۸۰۲	۰.۰۹۱۵
۰.۸۴۱۶	۰.۸۳۱۲	۰.۸۳۵	۰.۸۳۱۵

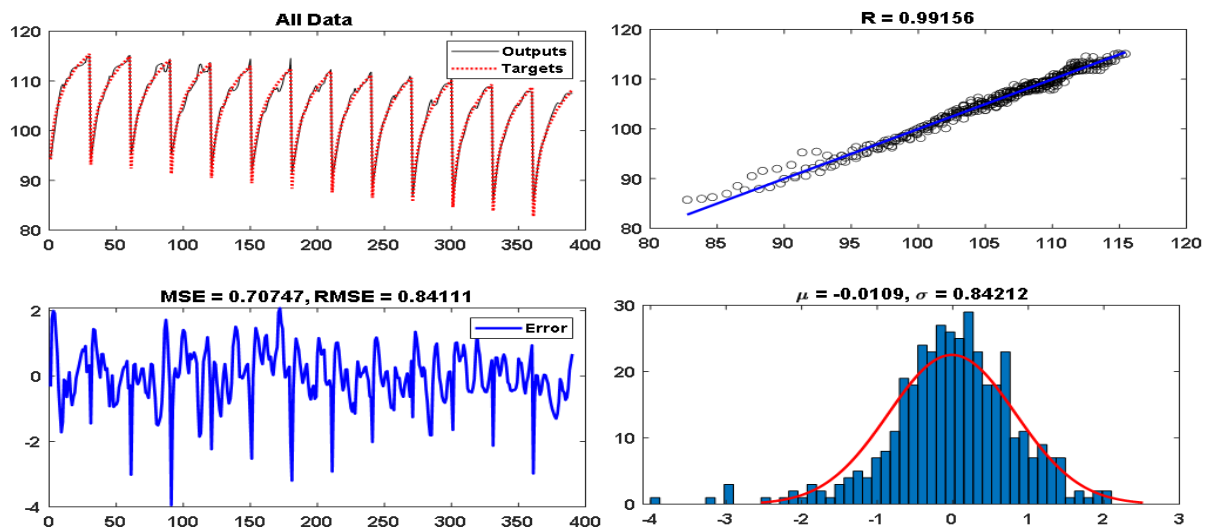
برای محاسبه بازدهی و متعاقب آن ریسک پرتفوی سهام، از شاخص‌های متعددی استفاده گردیده و مقادیر بازدهی و ریسک پرتفوی در هر ردیف جدول، مقابل آن نوشته شده است، برای الگوریتم پیشنهادی، بیشترین بازدهی در حالتیست که متغیر معرف بازدهی خالص ارزش‌داری‌های ثابت (PPE) باشد و برای پرتفوی نیز زمانی است که متغیر مورد نظر خالص ارزش‌داری‌های ثابت باشد. همچنین کمترین مقدار ریسک پرتفوی برای الگوریتم پیشنهادی زمانی است که متغیر معرف بازدهی، میزان بازگشت سرمایه (ROA) باشد و برای مدل پیشنهادی زمانی است که متغیر معرف بازدهی هزینه مالیات (CTE) باشد. بر مبنای نتایج حاصله از مدل شبکه عصبی ارائه شده در این پژوهش می‌توان در یافت که شبکه عصبی طراحی شده توانایی پیش‌بینی روند قیمت سهام با استفاده از شاخص‌های کل و صنعت را دارا می‌باشد و این امر علاوه بر تایید دیگری بر توانایی شبکه عصبی در پیش‌بینی حوزه‌های مالی، سودآوری استراتژی پیش‌بینی قیمت در بورس تهران را نیز تایید می‌کند.

معیارهای عملکرد شبکه برای ارائه نتایج عبارتند از: میانگین مربعات خطا، مجذور میانگین مربعات خطا، انحراف معیار، میزان رگرسیون، میانگین و انحراف معیار استاندارد می‌باشند. نتایج حاصل از استفاده از الگوریتم پس‌انتشار در شبکه عصبی پیشنهادی بر مبنای مراحل تست، آموزش، اعتبارسنجی و حالت کلی در جدول ۳ نمایش داده شده است.

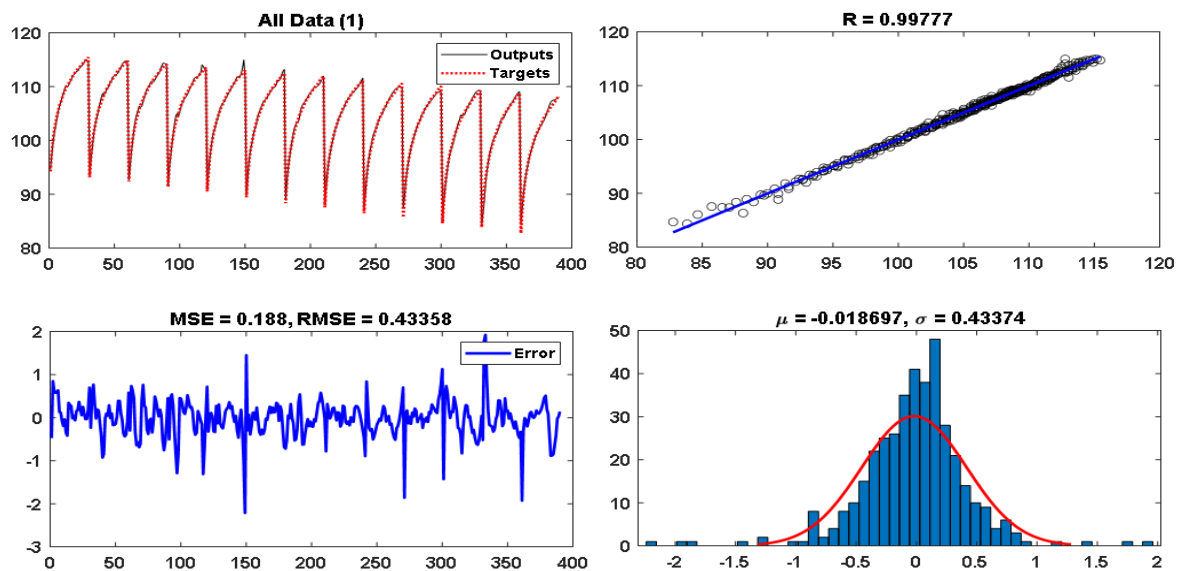
جدول ۳: نتایج الگوریتم پس‌انتشار در شبکه عصبی پیشنهادی

میزان رگرسیون (R)	انحراف معیار (δ)	میانگین (μ)	مجذور میانگین مربعات خطا (RMSE)	میانگین مربعات خطا (MSE)	شماره آزمایش
۰.۹۹۱۵۶	۰.۸۴۲۱۲	-۰.۰۱۰۹	۰.۸۴۱۱۱	۰.۷۰۷۴۷	آموزش
۰.۹۹۷۷۷	۰.۴۳۳۷۴	-۰.۰۱۸۶۹۷	۰.۴۳۳۵۸	۰.۱۸۸	تست
۰.۹۹۹۴۲	۰.۲۲۱۴۷	۰.۰۰۶۰۴۴۵	۰.۲۲۱۲۷	۰.۰۴۸۹۶۱	اعتبارسنجی
۰.۹۹۹۷۱	۰.۱۵۷۲۸	-۰.۰۰۷۰۴۶۵	۰.۱۵۷۲۳	۰.۰۲۴۷۲۲	حالت کلی

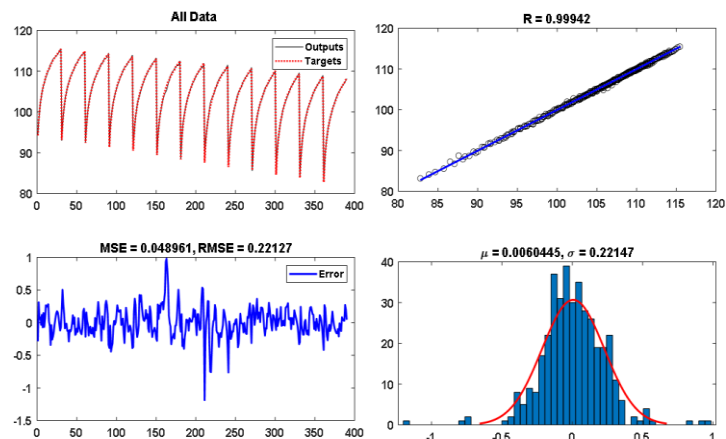
نمودارهای رگرسیونی مربوط به نتایج خروجی بر مبنای جدول ۳، در ادامه نمایش داده شده است.



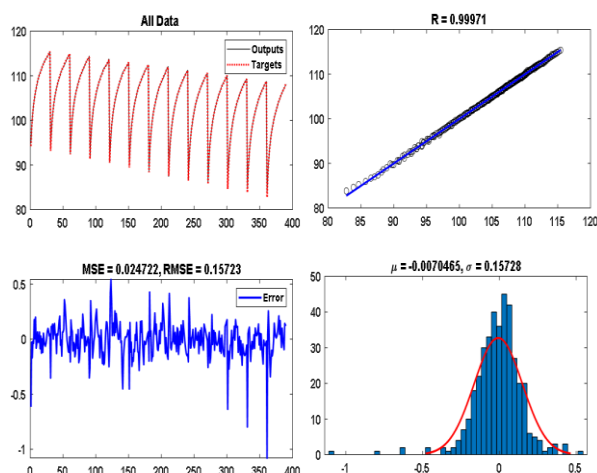
شکل ۴: نمودار رگرسیونی مربوط به حالت تست



شکل ۵: نمودار رگرسیونی مربوط به حالت آموزش

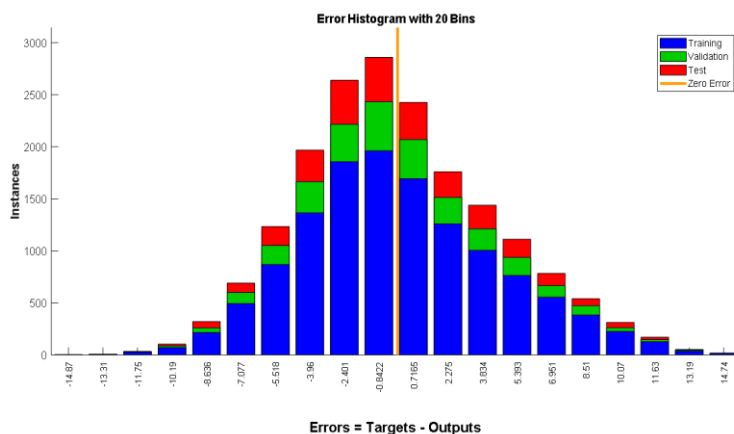


شکل ۶: نمودار رگرسیونی مربوط به حالت اعتبارسنجی



شکل ۷: نمودار رگرسیونی مربوط به حالت کلی

بر مبنای نتایج بدست آمده در چهار مرحله آزمایش می‌توان بیان داشت که الگوریتم مورد استفاده بر مبنای شبکه عصبی پیشنهادی، با افزایش نرخ یادگیری عملکرد شبکه بهبود می‌یابد. با این حال این الگوریتم به شدت نسبت به مینیمم‌های محلی حساس می‌باشد. حتی در مواقعی که نرخ یادگیری کوچک و در نتیجه گام حرکت الگوریتم کم است، این مشکل نیز وجود دارد. برای جلوگیری از این حساس بودن به مینیمم‌های محلی، الگوریتم مورد استفاده در شبکه پیشنهادی، همراه با ممنتوم نرخ یادگیری متفاوت در مراحل مختلف مورد آزمایش قرار گرفت و بهترین نتیجه مورد ارزیابی قرار گرفت. خطای هیستوگرام، با توجه به شکل ۸، میزان خطای هیستوگرام برای هر سه مرحله یادگیری، اعتبار سنجی و تست نمای داده شده است. میزان خطا با توجه به خطای صفر، خط زرد نشان داده شده در شکل ۸ برابر با ۰.۰۰۲۳۷۵ می‌باشد، که به دلیل نزدیک بودن به صفر از سطح قابل قبولی برخوردار است و به مقدار صفر نمی‌رسد.



شکل ۸: میزان خطای هیستوگرام

بحث و نتیجه‌گیری

نتایج این مطالعه نشان داد که استفاده از الگوریتم یادگیری تقویتی عمیق با ساختار شبکه عصبی مبتنی بر الگوریتم لونبرگ-مارد می‌تواند عملکرد سبد سهام را در مقایسه با روش‌های کلاسیک به طور معناداری بهبود بخشد. در مقایسه بین دو روش، الگوریتم پیشنهادی در متغیرهایی چون خالص ارزش دارایی‌های ثابت (PPE)، سود خالص، و بازده تجمعی، عملکرد بهتری نسبت به مدل مارکویتز از خود نشان داد. به‌ویژه، کمترین میزان ریسک برای مدل DRL زمانی به‌دست آمد که متغیر میزان بازگشت سرمایه (ROA) در نظر گرفته شد، در حالی که برای مدل سنتی، سود خالص (PTI-CTE) کمترین ریسک را به همراه داشت. این نتایج همراستا با مطالعات پیشین هستند که نشان داده‌اند الگوریتم‌های

یادگیری تقویتی عمیق، با بهره‌گیری از داده‌های چندبعدی و بلادرنگ، قادرند تصمیمات سرمایه‌گذاری دقیق‌تری نسبت به مدل‌های آماری سنتی اتخاذ کنند (Benhamou et al., 2021; Jiang et al., 2017; Koratamaddi et al., 2021).

در بخش عملکرد الگوریتم، معیارهای آماری همچون میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، انحراف معیار و ضریب رگرسیون، نمایانگر دقت بالای مدل طراحی‌شده در تمامی مراحل آموزش، تست و اعتبارسنجی بودند. برای نمونه، ضریب تعیین مدل در مرحله کلی به ۰.۹۹۹۷۱ رسید که نمایانگر تطابق بسیار بالا بین پیش‌بینی‌های مدل و مقادیر واقعی بود. چنین دقتی از مدل، بیانگر قدرت یادگیری و تعمیم بالا در مواجهه با داده‌های مالی پیچیده است. یافته‌های حاضر با نتایج پژوهش‌هایی مانند (Chaher, 2025) و (Yifu et al., 2024) مطابقت دارد که نشان داده‌اند مدل‌های یادگیری تقویتی عمیق، در صورت طراحی مناسب تابع پاداش و ساختار شبکه، قادرند بازدهی بالاتر و ریسک کمتری نسبت به روش‌های سنتی مانند MPT یا بهینه‌سازی میانگین-واریانس ارائه دهند.

یکی دیگر از نتایج مهم این پژوهش، افزایش نقدشوندگی سبد سهام حاصل از الگوریتم پیشنهادی بود. در مدل ارائه شده، شاخص نقدشوندگی شرکت هفتم به بیش از ۸۲ درصد رسید که نشان‌دهنده کارایی بالایی الگوریتم در انتخاب سهامی با قابلیت نقدشوندگی بالاست. این یافته از نظر عملی اهمیت زیادی دارد، چرا که بسیاری از مدل‌های کلاسیک، متغیر نقدشوندگی را در مدل لحاظ نمی‌کنند، اما در دنیای واقعی یکی از عوامل کلیدی در انتخاب پرتفوی به شمار می‌رود. یافته‌ی مذکور با نتایج (Henriques & Sadorsky, 2023) و (Heydari et al., 2021) همخوانی دارد که نشان داده‌اند ترکیب یادگیری ماشین با شاخص‌های مالی خاص مانند نقدشوندگی می‌تواند عملکرد کلی سبد را ارتقاء بخشد.

علاوه بر معیارهای کمی، تحلیل‌های گرافیکی نیز نتایج مدل را تایید می‌کنند. نمودارهای رگرسیونی نشان دادند که پیش‌بینی‌های مدل پیشنهادی در تمامی مراحل دارای همگرایی بالایی هستند و خطای هیستوگرام نیز در حد قابل قبول باقی مانده است. همچنین تحلیل واریانس سبد در بین شرکت‌های مختلف، حاکی از توانایی مدل در متوازن‌سازی ریسک بین دارایی‌هاست. از سوی دیگر، الگوریتم پیشنهادی توانسته است نرخ بازگشت سرمایه (ROI) را در چند شرکت به طرز معناداری افزایش دهد. این توانمندی در مدیریت هم‌زمان بازده و ریسک، نشانگر بلوغ الگوریتم‌های DRL در فضای مالی می‌باشد (Almahdi & Yang, 2017; Sun et al., 2021).

در بُعد تحلیل شاخص‌های فنی، استفاده از اندیکاتورهای چون میانگین متحرک، شاخص قدرت نسبی (RSI) و MACD در مدل، باعث شده است که مدل بتواند روند بازار را بهتر شناسایی کند و زمان‌بندی بهتری در خرید و فروش اعمال نماید. چنین رویکردی از طریق ترکیب ویژگی‌های زمانی-مالی با شبکه عصبی کانولوشن سه‌بعدی تحقق یافته است. این نوع ترکیب، همان‌گونه که در تحقیقات (Liang et al., 2018) و (Chaouki et al., 2020) نیز ذکر شده، می‌تواند رفتارهای مالی غیرخطی را بهتر مدل‌سازی کند و تصمیمات دقیق‌تری در شرایط پرنوسان بازار اتخاذ نماید.

در مدل پیشنهادی، مسئله مقیاس‌گذاری داده‌ها و نرمال‌سازی ویژگی‌ها با دقت بالایی انجام شده است، که این امر منجر به همگرایی سریع‌تر الگوریتم و جلوگیری از افتادن در مینیمم‌های محلی شده است. همچنین استفاده از تابع خطی در لایه دوم شبکه عصبی در کنار الگوریتم لونیگ-مارد، موجب شده است تا هم همگرایی سریع‌تری حاصل شود و هم دقت مدل بهبود یابد. این سازوکارها با پژوهش (Jyotirmayee Behera, 2025) قابل تطبیق است که در آن نیز تاکید شده استفاده از الگوریتم‌های بهینه‌سازی ترکیبی موجب ارتقای عملکرد پیش‌بینی شده است. همچنین (Ayari Salah, 2025) در فراتحلیل خود به این نتیجه رسید که بهره‌گیری از ترکیب الگوریتم‌های یادگیری عمیق و خوشه‌بندی می‌تواند بهترین عملکرد را در مراحل مختلف بازار به همراه داشته باشد.

در نهایت، آنچه یافته‌های این پژوهش را قابل توجه می‌سازد، هم‌راستایی آن‌ها با پژوهش‌هایی است که تأکید دارند ترکیب DRL و MPT نه تنها از منظر نظری امکان‌پذیر است، بلکه از نظر عملی نیز می‌تواند منجر به بهبود شاخص‌های بازده و کاهش ریسک گردد. همان‌طور که در تحقیقات (Goodell et al., 2021; Liu et al., 2020) ذکر شده، الگوریتم‌های یادگیری تقویتی با ترکیب داده‌های چندمنبعی و ساختارهای تصمیم‌گیری چندمرحله‌ای، می‌توانند در شرایط پرابهام بازار، تصمیمات دقیق‌تری اتخاذ کنند. مدل طراحی‌شده در این پژوهش نیز بر همین مبنا، ساختاری ارائه داده است که در آن هم داده‌های مالی سنتی و هم شاخص‌های فنی در قالب تانسورهای چندبعدی مورد استفاده قرار گرفته‌اند و از طریق لایه‌های شبکه عصبی به خروجی قابل اعتماد رسیده‌اند.

علی‌رغم نتایج قابل توجه این مطالعه، چند محدودیت نیز باید مدنظر قرار گیرد. نخست، داده‌های مورد استفاده تنها از بورس اوراق بهادار تهران انتخاب شده‌اند که ممکن است قابلیت تعمیم نتایج را به بازارهای دیگر کاهش دهد. دوم، علی‌رغم تلاش برای استفاده از اندیکاتورهای فنی و متغیرهای بنیادی، هنوز برخی داده‌های کیفی نظیر اخبار سیاسی، احساسات سرمایه‌گذاران یا تحولات جهانی لحاظ نشده‌اند. سوم، الگوریتم DRL پیشنهادی، به‌ویژه در مراحل اولیه آموزش، به شدت نسبت به نرخ یادگیری و ساختار اولیه شبکه حساس بود، که نیازمند آزمون‌های فراوان برای تعیین ساختار بهینه بود. چهارم، هزینه‌های محاسباتی بالا و نیاز به منابع سخت‌افزاری پیشرفته یکی دیگر از چالش‌های عملیاتی کردن این مدل در محیط واقعی است.

برای پژوهش‌های آتی، پیشنهاد می‌شود الگوریتم ارائه شده با داده‌های بازارهای بین‌المللی نظیر بورس نیویورک یا لندن نیز آزمون شود تا میزان تعمیم‌پذیری آن مورد ارزیابی قرار گیرد. همچنین، ترکیب الگوریتم DRL با داده‌های غیرساختاری مانند احساسات شبکه‌های اجتماعی یا تحلیل متون خبری می‌تواند نتایج دقیق‌تری به همراه داشته باشد. توسعه مدل‌هایی با معماری‌های قابل تفسیرتر مانند مدل‌های توجه یا استفاده از شبکه‌های قابل توضیح می‌تواند پذیرش صنعتی این الگوریتم‌ها را افزایش دهد. علاوه بر این، پیشنهاد می‌شود سنجش عملکرد مدل در شرایط بحرانی بازار مانند دوران رکود اقتصادی یا بحران‌های مالی نیز انجام شود تا پایداری استراتژی‌ها در شرایط نامعومول سنجیده شود. به تحلیل‌گران بازار سرمایه توصیه می‌شود از چارچوب‌های ترکیبی DRL و MPT برای تحلیل پرتفوی و پیش‌بینی بازده آتی شرکت‌ها بهره گیرند. مدیران دارایی و شرکت‌های سرمایه‌گذاری می‌توانند از این الگوریتم برای ایجاد سبدهای متنوع، با حداقل‌سازی ریسک و افزایش نقدشوندگی استفاده کنند. همچنین، به دانشگاه‌ها و مراکز آموزش مالی پیشنهاد می‌شود استفاده از الگوریتم‌های هوش مصنوعی پیشرفته در تحلیل مالی را در برنامه‌های درسی خود وارد نمایند تا نسل جدیدی از متخصصان توانمند در مدیریت سرمایه‌گذاری تربیت شود.

مشارکت نویسندگان

در نگارش این مقاله تمامی نویسندگان نقش یکسانی ایفا کردند.

تشکر و قدردانی

از تمامی کسانی که در طی مراحل این پژوهش به ما یاری رساندند تشکر و قدردانی می‌گردد.

تعارض منافع

در انجام مطالعه حاضر، هیچ‌گونه تضاد منافی وجود ندارد.

حمایت مالی

این پژوهش حامی مالی نداشته است.

موازین اخلاقی

در انجام این پژوهش تمامی موازین و اصول اخلاقی رعایت گردیده است.

References

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Neurocomputing*, 207, 279-290. <https://doi.org/10.1016/j.neucom.2016.05.005>

- Ayari Salah, G. H. (2025). A meta-analysis of supervised and unsupervised machine learning algorithms and their application to active portfolio management. *Expert Systems with Applications*, 271, 126611. <https://doi.org/10.1016/j.eswa.2025.126611>
- Benhamou, E., Saltiel, D., Ungari, S., & Mukhopadhyay, A. (2020). Bridging the gap between Markowitz planning and deep reinforcement learning. <https://doi.org/10.2139/ssrn.3702112>
- Chaher, A. (2025). Optimizing portfolio selection through stock ranking and matching: A reinforcement learning approach. *Expert Systems with Applications*, 269, 126430. <https://doi.org/10.1016/j.eswa.2025.126430>
- Chaouki, A., Hardiman, S., Schmidt, C., Sérié, E., & De Lataillade, J. (2020). Deep deterministic portfolio optimization. *The Journal of Finance and Data Science*, 6, 16-30. <https://doi.org/10.1016/j.jfds.2020.06.002>
- Cong, L. W., Tang, K., Wang, J., & Zhang, Y. (2021). AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Available at SSRN, 3554486
- Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3), 653-664. <https://doi.org/10.1109/TNNLS.2016.2522401>
- Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, 100577. <https://doi.org/10.1016/j.jbef.2021.100577>
- Henriques, I., & Sadorsky, P. (2023). Forecasting NFT coin prices using machine learning: Insights into feature significance and portfolio strategies. *Global Finance Journal*, 58, 100904. <https://doi.org/10.1016/j.gfj.2023.100904>
- Heydari, M. S., Validi, J., & Ebrahimi, S. B. (2021). Portfolio Optimization Based on a Robust Possibilistic Programming Model Using Genetic Algorithms and Mixed Frog Leaping Algorithm. *Financial Engineering and Securities Management*, 12(47).
- Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. arXiv preprint arXiv:1706. 10059
- Jyotirmayee Behera, P. K. (2025). An approach to portfolio optimization with time series forecasting algorithms and machine learning techniques. *Applied Soft Computing*, 170, 112741. <https://doi.org/10.1016/j.asoc.2025.112741>
- Koratamaddi, P., Wadhwani, K., Gupta, M., & Sanjeevi, S. G. (2021). Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4), 848-859. <https://doi.org/10.1016/j.jestch.2021.01.007>
- Liang, Z., Chen, H., Zhu, J., Jiang, K., & Li, Y. (2018). Adversarial deep reinforcement learning in portfolio management. arXiv preprint arXiv:1808. 09940
- Lima Paiva, F. C., Felizardo, L. K., Bianchi, R. A. d. C., & Costa, A. H. R. (2021). Intelligent trading systems: a sentiment-aware reinforcement learning approach. <https://doi.org/10.1145/3490354.3494445>
- Liu, X. Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., & Wang, C. D. (2020). FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. <https://doi.org/10.2139/ssrn.3737257>
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., & Bellemare, M. G. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533. <https://doi.org/10.1038/nature14236>
- Mohammadi Baghmalaie, H., Parsa, H., Tahmasebi, S., & Hajiani, P. (2021). Application of Cumulative Entropy Criterion and PSO Algorithm in Portfolio Optimization of Petrochemical Companies in Tehran Stock Exchange. *Biannual Journal of Development and Investment*, 6(2).
- Moshrefi, M., & Behnamian, J. (2022). Multi-Objective Portfolio Optimization Using Hierarchical Analysis and Genetic Algorithm. *Biannual Journal of Engineering Management and Soft Computing*, 7(1).
- Mostafaei Darmian, S., & Doaei, M. (2021). A Stochastic Optimization-Based Approach for Solving the Portfolio Selection Problem in Iran's Capital Market Using Metaheuristic Algorithms. *Quarterly Journal of Applied Economic Theories*, 8(4).
- Sun, S., Wang, R., He, X., Zhu, J., Li, J., & An, B. (2021). Deepscalper: A risk-aware deep reinforcement learning framework for intraday trading with micro-level market embedding. arXiv preprint arXiv:2201. 09058
- Sutton, R. S., & Barto, A. G. P. (2018). *Reinforcement learning: An introduction*. MIT Press.
- Yifu, J., Jose, O., & Majed, A. (2024). Deep reinforcement learning for portfolio selection. *Global Finance Journal*, 62, 101016. <https://doi.org/10.1016/j.gfj.2024.101016>