АНСАМБЛЕВЫЙ МЕТОД ГЛУБОКОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ УПРАВЛЕНИЯ ИНВЕСТИЦИОННЫМ ПОРТФЕЛЕМ НА РОССИЙСКОМ ФОНДОВОМ РЫНКЕ

А. А. Кобзев[а], **О. Н. Крахмалев**[б]

Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

[а] *ORCID: https://orcid.org/0009-0009-2369-1741,* ✍ *artem.kobzev.2001@mail.ru*

[б] *ORCID: https://orcid.org/0000-0002-9388-4137, onkrakhmalev@fa.ru*

Аннотация: в статье исследуется ансамблевый подход к управлению инвестиционным портфелем на российском фондовом рынке на основе глубокого обучения с подкреплением. Цель работы состоит в воспроизведении базовой ансамблевой архитектуры на данных российского рынка, в проверке ее переносимости и в выявлении тех модификаций, которые, действительно, улучшают качество торговли на длительный период. В качестве исходного решения рассматривается ансамбль из трех алгоритмов принятия решений, для которого последовательно анализируются расширение набора признаков, добавление макроэкономических переменных, штрафов за риск и механизма непрерывной адаптации одного из агентов в процессе торговли. Эксперименты проведены на данных 2015–2025 годов по ликвидным российским акциям, а итоговое сравнение выполнено на периоде 2023–2025 годов. Показано, что наибольший вклад в результат дает механизм непрерывного обучения, при котором агент дообучается на сделках любого активного участника ансамбля. Лучшая конфигурация обеспечивает суммарную доходность 61,7 процента и превосходит пассивные ориентиры по абсолютной доходности, однако не решает проблему защиты капитала на затяжном падающем рынке при высокой ключевой ставке.

Ключевые слова: глубокое обучение с подкреплением, управление инвестиционным портфелем, ансамблевые торговые стратегии, непрерывная адаптация, управление риском, российский фондовый рынок.

Для цитирования: Кобзев А. А., Крахмалев О. Н. Ансамблевый метод глубокого обучения с подкреплением для управления инвестиционным портфелем на российском фондовом рынке. *Успехи кибернетики.* 2026;7(2):119–125.

Поступила в редакцию: 06.04.2026. *В окончательном варианте: 14.05.2026.*

ENSEMBLE DEEP REINFORCEMENT LEARNING APPROACH FOR PORTFOLIO MANAGEMENT IN THE RUSSIAN EQUITY MARKET

A. A. Kobzev[a] , **O. N. Krakhmalev**[b]

Financial University under the Government of the Russian Federation, Moscow, Russian Federation

[a] *ORCID: https://orcid.org/0009-0009-2369-1741,* ✍ *artem.kobzev.2001@mail.ru*

[b] *ORCID: https://orcid.org/0000-0002-9388-4137, onkrakhmalev@fa.ru*

Abstract: we studied an ensemble-based deep reinforcement learning approach to portfolio management in the Russian stock market. The study aimed to reproduce a baseline ensemble architecture using Russian equity market data, evaluate its transferability, and identify modifications that improve trading performance over a long investment horizon. We implemented an initial model consisting of three decision-making agents and then extended the analysis by incorporating a broader set of features, macroeconomic indicators, risk-adjusted reward functions, and a mechanism for continuous adaptation of one agent during live trading. We trained and evaluated the models on data from 2015 to 2025 for liquid Russian equities, and we conducted the final performance comparison on the out-of-sample period from 2023 to 2025. The results show that the main driver of performance improvement is continuous off-policy training, where one agent is updated using trading data generated by any active agent in the ensemble. The best-performing configuration achieves a cumulative return of 61.7 percent and outperforms passive benchmark strategies in absolute return. However, the results also reveal a structural limitation: a long-only ensemble without an explicit allocation mechanism to low-risk assets does not provide sufficient capital protection during prolonged bear market conditions combined with high interest rates.

Keywords: deep reinforcement learning, portfolio management, ensemble trading strategies, continuous adaptation, risk-aware optimization, Russian stock market.

Cite this article: Kobzev A. A., Krakhmalev O. N. Ensemble Deep Reinforcement Learning Approach for Portfolio Management in the Russian Equity Market. *Russian Journal of Cybernetics.* 2026;7(2):119–125.

Original article submitted: 06.04.2026. *Revision submitted: 14.05.2026.*

Введение

Задача управления инвестиционным портфелем традиционно решается методами среднедисперсионной оптимизации и ее байесовских расширений [1, 2]. Однако для российского рынка 2023–2025 годов такие подходы ограничены из-за выраженных режимных переходов: восстановительный рост 2023 года сменился длительным снижением в 2024 году на фоне ключевой ставки 21 %, после чего последовало частичное восстановление.

В качестве базового ориентира в работе используется ансамблевая архитектура FinRL [3]. Цель исследования состоит в проверке ее переносимости на данные Московской биржи и в выделении того компонента, который, действительно, повышает качество торговли вне обучающей выборки. Основная рабочая гипотеза: решающим фактором является не усложнение ансамбля, а непрерывная адаптация off-policy-агента на совокупном опыте системы.

Обзор литературы

Применение глубокого обучения с подкреплением к торговле развивалось от дискретных алгоритмов семейства DQN [4] к методам непрерывного действия, таким как DDPG [5], PPO [6] и SAC [7]. Для портфельных задач именно ансамбли и унифицированные среды FinRL-Meta [8] сделали возможным сопоставление нескольких политик в едином экспериментальном протоколе. Вместе с тем результаты на развитых рынках не гарантируют переносимости на рынок с высокой волатильностью и жесткими макроэкономическими шоками.

Для финансовых приложений особенно важны две линии работ: исследования практической применимости DRL в торговле [9] и работы по continual RL [10]. Они дополняются подходами к риск-чувствительной оптимизации, в которых используются когерентные меры риска и штрафы, связанные с рисками в хвостах распределений доходностей [11, 12]. В настоящей статье эти идеи сведены к минималистичной схеме: ансамбль из трех агентов, selection по качеству валидационного окна и непрерывное обновление SAC.

Методы

Эксперименты проведены на данных по 20 ликвидным акциям Московской биржи за 2015–2025 годы. Состояние агента включает долю денежных средств, веса текущего портфеля, 11 технических признаков на каждый актив, а также две макроэкономические переменные: ключевую ставку и курс рубля к доллару. Данные признаки чаще всего используются в торговых алгоритмах, в многочисленных работах, связанных с биржевым делом. Состав признаков приведен в таблице 1.

Формально наблюдаемое состояние системы в момент t задается формулой (1). Оно объединяет денежную позицию, текущие веса портфеля, технические признаки по каждому активу и макроэкономические переменные. Вектор действия агента интерпретируется как целевые веса портфеля и нормируется по формуле (2), что исключает отрицательные веса и обеспечивает сумму долей, равную единице.

$$s_t = \left[r_t^{\text{cash}},\, w_{1,t},\, \ldots,\, w_{N,t},\, f_{1,t},\, \ldots,\, f_{11N,t},\, k_t,\, u_t \right]^{\top}. \tag{1}$$

В формуле (1) r_t^{cash} обозначает долю капитала в денежной позиции, $w_{i,t}$ — вес i-го актива, $f_{i,t}$ — технические признаки, а k_t и u_t соответствуют ключевой ставке Банка России и курсу USD/RUB.

$$w_{i,t} = \frac{\max(0,\, a_{i,t})}{\max\left(1,\, \sum_{j=1}^{N} \max(0,\, a_{j,t})\right)}. \tag{2}$$

Таблица 1

Технические признаки, используемые в наблюдаемом пространстве состояний

№	**Признак**	**Обозначение**	**Описание**
1	Цена закрытия	close_z	Нормированная цена закрытия тикера
2	Логдоходность	close_ret	$\log(P_t/P_{t-1})$
3	Объем	value_log	Логарифм денежного оборота
4	MACD	macd	Разность экспоненциальных средних
5	RSI	rsi	Индекс относительной силы
6	CCI	cci	Commodity Channel Index
7	ADX	adx	Индекс направленного движения
8	ATR	atr	Мера текущей волатильности
9	Bollinger %B	boll_b	Позиция цены в полосе Боллинджера
10	OBV	obv_z	Нормированный балансовый объем
11	Относительная сила	rel_strength	Доходность бумаги к рынку

Формула (2) переводит непрерывное действие $a_{i,t}$ в допустимый набор портфельных весов: отрицательные компоненты отсекаются, после чего оставшиеся веса нормируются на их сумму.

Функция вознаграждения задается формулой (3). Она сочетает логарифмическую доходность портфеля со штрафами за текущую просадку, хвостовой риск CVaR и избыточный оборот, поэтому обучение ориентировано не только на рост капитала, но и на устойчивость результата.

$$r_t = 100 \cdot \log \frac{V_t}{V_{t-1}} - \lambda_{dd} DD_t - \lambda_{cvar} \mathrm{CVaR}_{5\%}(t) - \lambda_{turn} TO_t. \tag{3}$$

В формуле (3) V_t — стоимость портфеля, DD_t — текущая относительная просадка, $\mathrm{CVaR}_{5\%}(t)$ — условная стоимость под риском на уровне 5 %, а TO_t — оборот портфеля.

Штрафной член CVaR в формуле (3) раскрывается формулой (4). Он оценивает средний убыток в левом хвосте распределения недавних доходностей и тем самым усиливает чувствительность агента к экстремальным неблагоприятным сценариям.

$$\mathrm{CVaR}_{\alpha}(t) = -\mathbb{E}\left[r \le q_{\alpha}(r_{[t-20,\ t]})\right]. \tag{4}$$

В формуле (4) q_α задает α-квантиль доходностей на скользящем окне, а математическое ожидание берется только по наблюдениям, лежащим ниже этого порога.

Ансамбль состоит из PPO, A2C и SAC. В конфигурации v3 межагентное непрерывное обучение реализуется через сочетание on-policy и off-policy обновлений. Базовая целевая функция PPO задается формулой (5) и отражает клиппированное обновление политики; A2C использует ту же логику, но без PPO-клиппинга. Межагентное обновление критика SAC задается формулой (6). Для каждого валидационного окна длиной 63 торговых дня качество агента дополнительно оценивается коэффициентом Sortino по формуле (7), составным показателем по формуле (8), а правило выбора активного агента на следующее окно задается формулой (9).

$$L^{PPO}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t,\ \mathrm{clip}(\rho_t(\theta),\, 1-\epsilon,\, 1+\epsilon)\hat{A}_t\right)\right]. \tag{5}$$

В формуле (5) $\rho_t(\theta)$ обозначает отношение правдоподобий новой и старой политик, $\hat{A}_t$ — оценку преимущества, а оператор clip ограничивает слишком агрессивное изменение политики между итерациями.

$$L^{SAC}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - y\right)^2\right], \qquad y = r + \gamma \bar{V}(s'). \tag{6}$$

В формуле (6) $\mathcal{D}$ — буфер повтора, содержащий переходы от любой политики ансамбля, а целевое значение y опирается на целевую функцию ценности следующего состояния. Именно этот шаг обеспечивает адаптацию SAC во время торговли.

$$\text{Sortino}_k^{(i)} = \frac{\bar{r}_k^{(i)}}{\sigma_{down,k}^{(i)} + \epsilon}. \tag{7}$$

В формуле (7) средняя дневная доходность агента на валидационном окне сопоставляется с нисходящим стандартным отклонением, а ϵ выступает малой стабилизирующей добавкой в знаменателе.

$$\text{score}_k^{(i)} = \text{Sortino}_k^{(i)} \cdot \exp\left(-\frac{\left|\text{MaxDD}_k^{(i)}\right|}{\sigma_{dd}}\right). \tag{8}$$

Формула (8) вводит штраф за глубокую просадку $\left|\text{MaxDD}_k^{(i)}\right|$, поэтому высокое значение Sortino само по себе не гарантирует выбор агента. Параметр σ_{dd} задает чувствительность штрафа и в экспериментах принимается равным 0,2.

$$i_k^* = \arg \max_{i \in \{PPO,\ A2C,\ SAC\}} \text{score}_k^{(i)}. \tag{9}$$

Формула (9) фиксирует правило выбора политики с максимальным значением score на очередном окне и замыкает контур принятия решений ансамбля.

Сопоставление конфигураций выполняется по суммарной доходности, годовой доходности, коэффициентам Sharpe и Sortino, а также по максимальной просадке. Как видно на рисунке 1, SAC быстрее других агентов выходит на высокие валидационные значения доходности и Sortino, поэтому именно он выбран в качестве адаптивного компонента ансамбля.

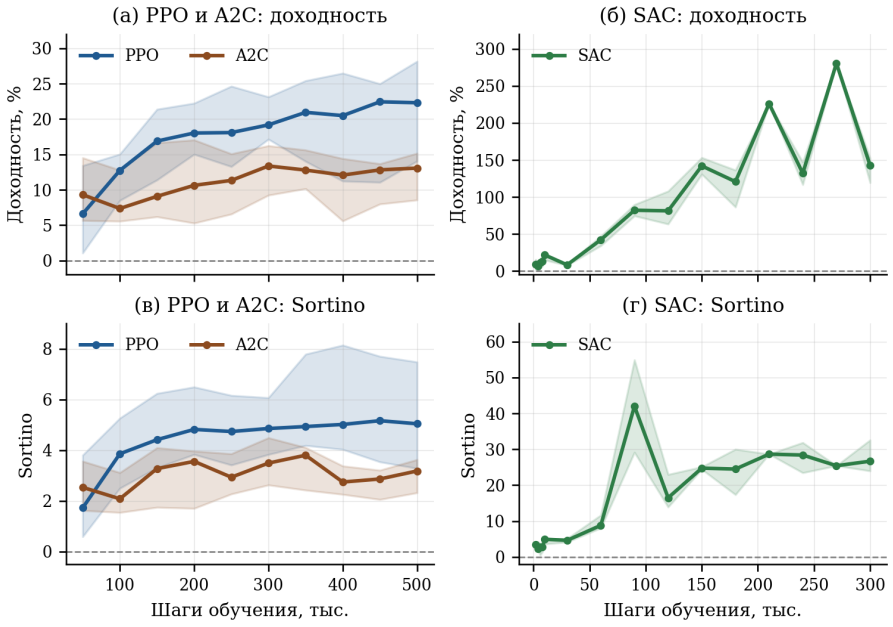

Рис. 1. *Динамика валидационных метрик PPO, A2C и SAC в ходе обучения*

Результаты

Ключевые воспроизводимые конфигурации сведены в таблицу 2. Базовая архитектура v1 практически не создает избыточной доходности на тесте, тогда как замена DDPG на SAC и расширение пространства признаков уже в v2 дают заметный прирост качества. Наилучший результат по абсолютной доходности получен в конфигурации v3: +61,7 % при Sharpe 0,590 и Sortino 0,730.

Таблица 2

Сводная таблица ключевых конфигураций и бенчмарков

Сводная таблица ключевых конфигураций и бенчмарков						
Конфигурация	Ключевые особенности	Total, %	Annual, %	Sharpe	Sortino	MaxDD, %
v1 — Базовая архитектура	PPO + A2C + DDPG; 7 признаков; log-return без риск-штрафов	+2,8	∼1,0	<0,30	—	−44,8
v2 — расширенная архитектура	SAC вместо DDPG; 11 признаков; макрофакторы; CVaR; адаптивная нормализация	+57,4	+15,1	0,547	0,703	−39,3
v3 — межагентное непрерывное обучение SAC	SAC обновляется на опыте любого активного агента после каждого торгового дня	+61,7	+16,1	0,590	0,730	−33,4
v4 — подтверждающая конфигурация	Обучение с нуля; online SAC; без дополнительных модулей рыночного режима	+51,4	+13,7	0,571	0,739	−32,1
v5 — резервный переход в кэш	При отрицательных score всех агентов портфель удерживается без торговли	+58,7	+15,3	0,500	1,020	−68,2
Buy&Hold IMOEX	Пассивный бенчмарк по индексу MOEX	+27,3	+8,3	0,499	0,739	−32,1
Равновзвешенный портфель	Пассивный бенчмарк по 20 бумагам	+40,8	+11,0	0,627	0,952	−29,8

Из рисунка 2 можно понять, что между суммарной доходностью, коэффициентом Sortino и глубиной просадки сохраняется выраженный компромисс. Конфигурация v4 подтверждает устойчивость вывода при обучении с нуля: она уступает v3 по абсолютной доходности, но практически совпадает с индексом IMOEX по максимальной просадке. Конфигурация v5 повышает Sortino до 1,020, однако достигает этого ценой неприемлемого роста максимальной просадки.

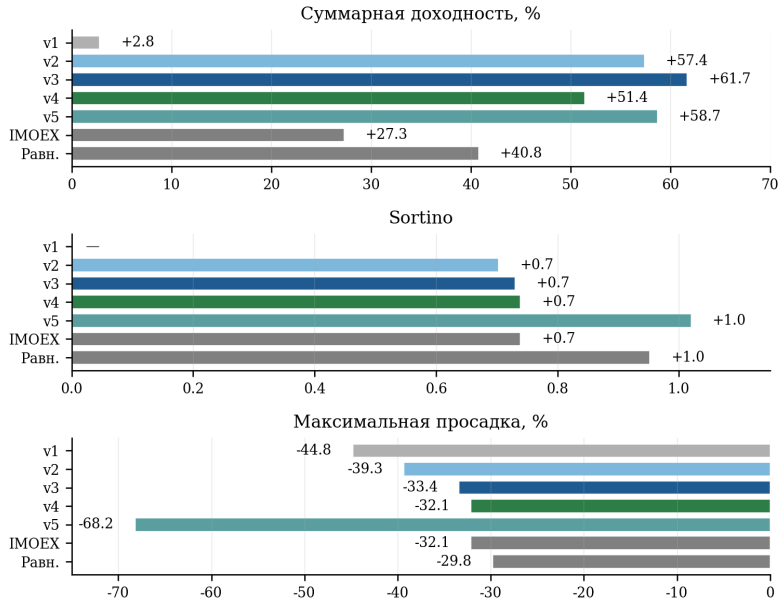

Рис. 2. *Сопоставление ключевых конфигураций по доходности, Sortino и просадке*

На рисунке 3 эффект непрерывного обучения SAC проявляется на ряде окон с отрицательным качеством у всех агентов базовой архитектуры, обновляемый SAC сохраняет положительное преимущество или быстрее восстанавливается на следующих окнах. Это подтверждает, что основной вклад в результат дает именно межагентное накопление опыта, а не увеличение числа эвристических модулей.

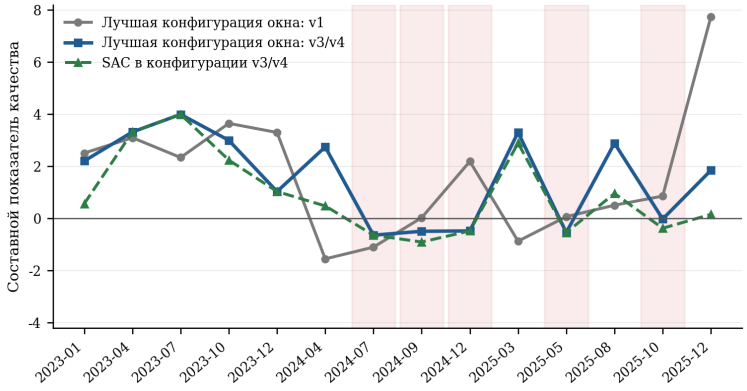

Рис. 3. *Влияние непрерывного межагентного обучения SAC на качество агентов*

При этом структурная проблема остается нерешенной. Все исследованные конфигурации являются long-only, а денежная позиция в среде не несет явной безрисковой доходности. Поэтому в режиме одновременного падения рынка и высокой ключевой ставки ансамбль не получает корректного сигнала в пользу защитного поведения. Это объясняет, почему задача защиты капитала остается открытой даже при заметном росте абсолютной доходности.

Заключение

Исследование показывает, что ансамбль PPO, A2C и SAC может быть перенесен на российский фондовый рынок, если критический компонент системы выполнен как непрерывно обновляемый off-policy агент. На данных 2023–2025 годов конфигурация v3 устойчиво превосходит пассивные ориентиры по абсолютной доходности, а подтверждающая конфигурация v4 воспроизводит основной вывод без дополнительных модулей. Следующий этап развития метода должен быть связан не с дальнейшим усложнением ансамбля, а с явным моделированием безрискового актива и режимов защиты капитала.

ЛИТЕРАТУРА

1. Markowitz H. M. Portfolio Selection. *The Journal of Finance*. 1952;7(1):77–91. DOI: 10.1111/j.1540-6261.1952.tb01525.x.
2. Black F., Litterman R. Global Portfolio Optimization. *Financial Analysts Journal*. 1992;48(5):28–43. DOI: 10.2469/faj.v48.n5.28.
3. Yang Z., Zhu Y., Guo J., Liu X.-Y., Zhong S., Walid A. Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. *ICAIF '20*. 2020:1–8. DOI: 10.1145/3383455.3422540.
4. Mnih V., Kavukcuoglu K., Silver D. et al. Human-Level Control through Deep Reinforcement Learning. *Nature*. 2015;518:529–533. DOI: 10.1038/nature14236.
5. Lillicrap T. P., Hunt J. J., Pritzel A. et al. Continuous Control with Deep Reinforcement Learning. arXiv:1509.02971. 2016. DOI: 10.48550/arXiv.1509.02971.
6. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal Policy Optimization Algorithms. arXiv:1707.06347. 2017. DOI: 10.48550/arXiv.1707.06347.
7. Haarnoja T., Zhou A., Abbeel P., Levine S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *Proceedings of ICML*. 2018:1861–1870.
8. Liu S., Rui J., Gao J. et al. FinRL-Meta: Market Environments and Benchmarks for Data-Driven Financial Reinforcement Learning. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

9. Theate T., Ernst D. An Application of Deep Reinforcement Learning to Algorithmic Trading. *Expert Systems with Applications*. 2021;173:114632. DOI: 10.1016/j.eswa.2021.114632.
10. Khetarpal K., Riemer M., Rish I., Precup D. Towards Continual Reinforcement Learning: A Review and Perspectives. *Journal of Artificial Intelligence Research*. 2022;75:1401–1476. DOI: 10.1613/jair.1.13673.
11. Artzner P., Delbaen F., Eber J.-M., Heath D. Coherent Measures of Risk. *Mathematical Finance*. 1999;9(3):203–228. DOI: 10.1111/1467-9965.00068.
12. Rockafellar R. T., Uryasev S. Conditional Value-at-Risk for General Loss Distributions. *Journal of Banking & Finance*. 2002;26(7):1443–1471. DOI: 10.1016/S0378-4266(02)00271-6.