

Aprendizado por Reforço com LSTM e DDPG Otimizados através de Comitê PSO para Análise Preditiva de Séries Temporais Financeiras

Adonias Junior de Albuquerque Mattos e Suzete Élide Nóbrega Correia

Resumo— A extração de sinais preditivos em séries temporais financeiras é desafiadora pela sua natureza intrinsecamente ruidosa. Este trabalho propõe uma arquitetura de Aprendizado por Reforço Profundo baseada em DDPG (*Deep Deterministic Policy Gradient*) e redes LSTM (*Long Short-Term Memory*), distribuída em um Comitê de agentes especialistas em múltiplas escalas (diário, 1h, 30m e 5m). As ações recomendadas individualmente são ponderadas via Otimização por Enxame de Partículas. Validado em *backtesting*, o modelo proposto alcançou Índice de Sharpe de 0,81 e retorno de 22,29%, superando estratégias de *benchmarks* com um decréscimo de capital significativamente reduzido.

Palavras-Chave— Aprendizado por Reforço, DDPG, LSTM, PSO, Séries Temporais, Mercado Financeiro.

Abstract— Extracting predictive signals in financial time series is challenging due to inherent noise. This paper proposes a Deep Reinforcement Learning architecture based on DDPG and LSTM (*Long Short-Term Memory*) networks, structured into a committee of specialist agents across multiple scales (daily, 1h, 30m, and 5m). The actions recommended by individual agents are weighted using Particle Swarm Optimization. Backtesting results demonstrate that the proposed framework achieved a Sharpe Ratio of 0,81 and 22,29% return, outperforming standard market benchmarks with significantly reduced maximal drawdown.

Keywords— Reinforcement Learning, DDPG, LSTM, PSO, Time Series, Financial Market.

I. INTRODUÇÃO

A extração de sinais preditivos em séries temporais financeiras lida com um ambiente altamente estocástico, onde fatores econômicos e ruído mascaram correlações latentes. Estratégias convencionais possuem limitações na capacidade de analisar simultaneamente com variações de escalas temporais dinâmicas [1].

Existem diversas arquiteturas e modelagens matemáticas aplicáveis à otimização de estratégias de negociação no mercado financeiro. Dentre esses, o Aprendizado por Reforço (*Reinforcement Learning* - RL) tem se destacado como uma abordagem promissora devido à sua capacidade de aprender políticas de decisão através da interação com o ambiente [2]. O RL por sua vez possui diversas variações, como o DDPG (*Deep Deterministic Policy Gradient*), que é um algoritmo de aprendizado por reforço *off-policy* que combina a eficiência de métodos baseados em valor com a capacidade de lidar com espaços de ação contínuos, tornando-o particularmente adequado

para aplicações financeiras onde as decisões de negociação podem envolver uma ampla gama de ações possíveis [3].

A ideia de utilizar agentes de IA para negociação financeira é relativamente recente, e existem trabalhos anteriores que exploraram essa abordagem. Thomas Fisher e Christopher Krauss [4] demonstraram que o uso de redes neurais LSTM (*Long Short-Term Memory*) pode ser eficaz na previsão de tendências de mercado.

A aplicação de RL para análises financeiras é um campo relevante, relativamente competitivo e em crescimento. Sua natureza dinâmica e virtude de possuir diversas variáveis e fatores de influência torna o mercado financeiro um ambiente ideal para a experimentação de técnicas de aprendizado por reforço como o DDPG [2]. Nesse contexto, o mercado não é visto aqui sob a ótica da economia clássica ou da análise fundamentalista de ativos, mas sim como um ambiente estocástico de alta complexidade. O objetivo primordial deste trabalho, portanto, transcende a mera busca por rentabilidade financeira; ele reside no desafio científico de modelar um agente capaz de extrair padrões em um ambiente ruidoso e não-estacionário. O mercado atua como um ambiente de validação empírica para a convergência e robustez do algoritmo DDPG proposto, onde a geração de sinais confiáveis serve como métrica de sucesso da capacidade de aprendizado do modelo.

Este artigo propõe uma nova arquitetura que une o DDPG, redes de memória profunda (LSTM) e análise multi-escala integrados e otimizados por um agrupamento em enxame de partículas através do PSO (*Particle Swarm Optimization*). A estratégia foca na fusão de agentes treinados de forma granular em variadas janelas temporais, para convergir em um modelo transicional robusto ao ruído inerente aos ativos abordados.

II. FUNDAMENTAÇÃO TEÓRICA

A. Deep Deterministic Policy Gradient (DDPG)

O DDPG é um algoritmo do tipo ator-crítico *off-policy* construído para espaços multidimensionais de ação contínua [3]. O ator projeta estados para a escolha exata da ação enquanto o crítico calcula a função de valor ação estado (*Q-value*).

B. Long Short-Term Memory (LSTM)

As redes LSTM são uma variação de redes neurais recorrentes projetadas para lidar com o problema do *vanishing gradient*, permitindo que o modelo aprenda dependências de longo prazo em sequências de dados [5].

Matematicamente, o funcionamento de uma célula LSTM e suas portas no instante t pode ser descrito pelas seguintes equações [6][7]:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \quad (1)$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \quad (2)$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \quad (3)$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Em que:

- x_t é o vetor de entrada no instante t ;
- h_{t-1} é o estado oculto do instante anterior;
- c_{t-1} é o estado da célula do instante anterior;
- i_t, f_t, g_t, o_t são os componentes das portas de entrada, esquecimento, célula (candidato) e saída, respectivamente;
- c_t é o estado atualizado da célula e h_t é o estado oculto atualizado;
- W_i e W_h são as matrizes de pesos para as entradas e estados ocultos;
- b_i e b_h são os vetores de vies (bias);
- σ denota a função sigmoide e $\odot$ o produto de Hadamard (elemento a elemento).

C. Adaptive Moment Estimation (ADAM)

O Otimizador ADAM é utilizado para atualizar de forma eficiente os pesos das redes neurais (ator e crítico) [8]. O método combina as heurísticas do Momentum e do RMSProp, calculando médias móveis exponenciais do gradiente (primeiro momento) e do gradiente ao quadrado (segundo momento) [9].

D. Otimização por Enxame (PSO)

O PSO utiliza um algoritmo de otimização inspirado no comportamento coletivo de enxames para encontrar a melhor combinação de sinais. Cada partícula representa uma possível combinação de pesos para os agentes, e o algoritmo busca otimizar esses pesos com base em algum parâmetro de desempenho (ex: retorno acumulado, Sharpe *ratio*) em um conjunto de validação [10][11]. Nessa abordagem, será aplicado o parâmetro de retorno acumulado como função objetivo, buscando maximizar o retorno do sistema multi-agente como um todo. O modelo atualiza a velocidade v e a posição x da partícula i na iteração t segundo as equações canônicas:

$$v_i^{t+1} = wv_i^t + c_1r_1(pbest_i - x_i^t) + c_2r_2(gbest - x_i^t) \quad (7)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (8)$$

sendo w o fator de inércia, c_1 e c_2 os coeficientes cognitivo e social, r_1 e r_2 valores aleatórios entre $[0, 1]$, enquanto $pbest_i$ e $gbest$ indicam o melhor desempenho individual da partícula e o ótimo global do enxame iterativo, respectivamente. A atualização dessas memórias na iteração $t + 1$ ocorre com base em uma função objetivo (*fitness*) de retorno f . O

melhor desempenho particular $pbest_i$ retém a melhor posição alcançada pela partícula específica:

$$pbest_i^{t+1} = \begin{cases} pbest_i^t, & \text{se } f(x_i^{t+1}) \leq f(pbest_i^t) \\ x_i^{t+1}, & \text{se } f(x_i^{t+1}) > f(pbest_i^t) \end{cases} \quad (9)$$

E o ótimo global $gbest$ converge tomando a posição da partícula com maior retorno dentre todo o enxame S :

$$gbest^{t+1} = \arg \max_{pbest_i \in S} f(pbest_i^{t+1}) \quad (10)$$

E. Pré-tratamento de Dados (Normalização Z-Score)

O pré-tratamento de dados é uma etapa crucial para o sucesso de qualquer modelo de aprendizado de máquina, especialmente em séries temporais financeiras, onde os dados podem ser ruidosos, possuir escalas díspares e não serem estacionários.[12][13]

Inicialmente foi observado que dados brutos de mercado, como o preço de fechamento, possuem magnitudes absolutas que variam significativamente ao longo do tempo (ex: índice Bovespa variando de 60.000 a 130.000 pontos). Utilizar valores absolutos pode dificultar a convergência da rede neural. Portanto, características de preço são transformadas em sinais de variação relativa ou oscilação, destacando a dinâmica do movimento em vez do valor nominal.

Em seguida, aplica-se a técnica de Padronização, ou Z-Score *Normalization* [14]. Essa técnica redimensiona a distribuição dos dados para que tenham média zero ($\mu = 0$) e desvio padrão unitário ($\sigma = 1$). A transformação é pela equação 11:

$$z = \frac{x - \mu}{\sigma} \quad (11)$$

A Figura 1 ilustra o efeito do pré-tratamento, comparando o sinal bruto com o sinal processado.

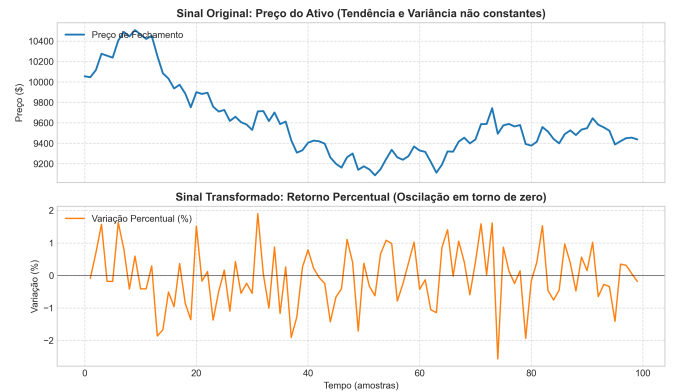


Fig. 1: Comparação entre sinal bruto e sinal pré-processado

III. METODOLOGIA

A. Base de Dados

A base de dados utilizada para o treinamento e teste dos agentes é composta por séries temporais financeiras pertencentes ao minicontrato de índice Bovespa (WIN) negociado na B3, com dados de preço de abertura, fechamento, máxima, mínima e volume, coletados em intervalos de 1 dia, 1 hora,

30 minutos e 5 minutos. Para o banco de dados de treino, o período de coleta inicial é 19 de março de 2021 e se estende até 02 de fevereiro de 2026, totalizando 1218 dias, 11661 horas, 22589 passos de meia hora. Para o agente de 5 minutos em particular, o período de coleta se inicia em 26 de maio de 2023 e se estende até 02 de fevereiro de 2026, totalizando 74704 passos de 5 minutos. Para o banco de dados de teste, o período de coleta se inicia em 03 de fevereiro de 2026 e se estende até 12 de junho de 2026. O banco de dados de teste por sua vez pode ser lido como um total de 9910 passos de 5 minutos, ou 89 dias de negociação, ou 614 horas de negociação. O banco de dados de teste é subdividido em uma sessão composta pelos primeiros 3000 passos de 5 minutos e o restante com dimensão de 6910 passos de 5 minutos. Dessa forma, tem-se 12 diferentes bancos de dados, sendo 4 para treinamento dos agentes individuais, 4 para teste dos agentes individuais e 4 para teste do Comitê otimizado pelo PSO. Alguns fatores devem ser observados:

- O período de teste utilizado para validar o Comitê otimizado pelo PSO é diferente do período utilizado para determinar os pesos otimizados.
- O período de teste dos agentes individuais é o mesmo do período utilizado para determinar os pesos otimizados pelo PSO.
- O período de teste dos agentes é posterior ao período de treinamento dos agentes.
- O período de teste do Comitê otimizado pelo PSO é posterior ao período de teste dos agentes individuais.
- Os itens com horário igual ou superior a 17:00:00 não utilizados para a validação ou teste dos agentes e Comitês, porém são considerados para o dimensionamento final dos bancos de dados como 3000 e 9910 passos de 5 minutos.
- Os dados obtidos para todos bancos de dados estão sujeitos ao *rollback* proporcional do contrato futuro, uma prática comum em ambientes de *trading*.

B. Pipeline de Desenvolvimento

O desenvolvimento deste trabalho foi estruturado em um *pipeline* experimental sequencial, desenhado para validar a hipótese de que a diversificação temporal em agentes de RL melhora a robustez das estratégias de negociação. A Figura 2 apresenta um diagrama de desenvolvimento:

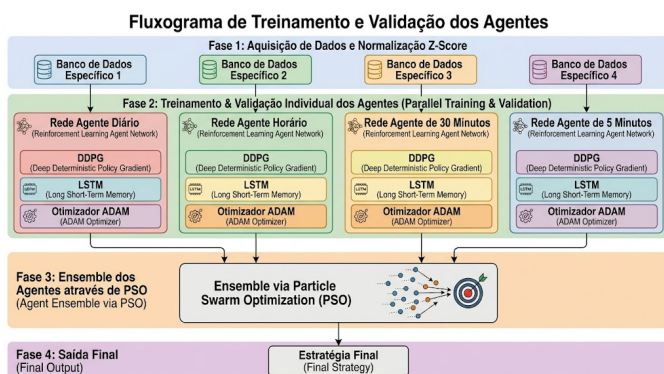


Fig. 2: Diagrama do *Pipeline* de Desenvolvimento

Aonde é possível observar o fluxo de desenvolvimento, desde a coleta e pré-processamento dos dados, passando pelo treinamento dos agentes individuais, até a fase de otimização dos pesos do Comitê através do PSO e a avaliação final do sistema multi-agente otimizado. Relevante observar que durante o treinamento o ADAM é utilizado para atualizar os pesos das redes neurais (ator e crítico), momento a partir de qual o ciclo se reinicia.

C. Processo de Decisão

Os agentes temporais (diário, 1h, 30m e 5m) propõem uma ação que varia no intervalo $[-1, 1]$, onde valores positivos acima de 0.1 indicam uma ação de compra, valores negativos abaixo de -0.1 indicam uma ação de venda e valores entre -0.1 e 0.1 indicam abstenção.

O limiar de decisão é um parâmetro que determina a sensibilidade do sistema multi-agente ao tomar uma decisão de compra, venda ou abstenção. Ele atua como um filtro que define o nível mínimo de confiança necessário para que uma ação seja executada com base nos sinais combinados dos agentes. A escolha do limiar de decisão pode impactar significativamente o desempenho do sistema, influenciando a taxa de acerto, o número de operações realizadas e o retorno acumulado. Foram analisadas 4 abordagens distintas para o limiar de decisão dos comitês de agentes:

- **Voto Majoritário:** Consiste em determinar a ação final através de um sistema de votação majoritária, de forma que cada agente individual contribui com um voto de compra, venda ou abstenção, caso todos os agentes que compõem o Comitê proponham a mesma ação, essa é executada, caso contrário, o Comitê se abstém. Uma vez que os agentes temporais durante a fase de treinamento utilizam o intervalo de -0.1 a 0.1 como seus limiares de decisão individuais, é consistente utilizar os mesmos para definir se um voto de um agente específico é considerado como um voto de compra, venda ou abstenção. Caso a ação sugerida por um dos agentes seja maior do que 0.1 será considerado voto de compra, caso seja menor do que -0.1 será considerado voto de venda e caso esteja entre -0.1 e 0.1 será considerado voto de abstenção.
- **Voto Ponderado:** No contexto de votação ponderada, o valor da ação proposta pelos agentes individuais é somado diretamente antes de determinar qualquer ação (compra, venda, abstenção), dessa forma, caso a soma seja maior do que 0.1 será considerada ação de compra, caso seja menor do que -0.1 será considerada ação de venda e caso esteja entre -0.1 e 0.1 será considerada ação de abstenção.
- **PSO-Padrão:** No Comitê PSO-Padrão, o algoritmo de otimização PSO é utilizado para de forma simultânea, definir o limiar de decisão do Comitê e atribuir pesos aos agentes individualmente, de forma que a ação proposta por cada agente seja multiplicada pelo seu peso correspondente, e em seguida somada, obtendo-se assim uma ação agregada que será comparada ao limiar de decisão do Comitê. Dessa forma, caso a ação agregada seja maior do que o limiar de decisão do Comitê, será considerada ação de compra, caso seja menor do que o limiar de

decisão do Comitê, será considerada ação de venda e caso contrário será considerada ação de abstenção.

- **PSO-Semanal:** Assim como no Comitê PSO-Padrão, no Comitê PSO-Semanal, o algoritmo de otimização PSO é utilizado para de forma simultânea, definir o limiar de decisão do Comitê e atribuir pesos aos agentes individualmente, porém, a análise de otimização dos pesos do Comitê é realizada para cada dia da semana, de forma que o Comitê PSO-Semanal possui 5 limiares de decisão e 5 conjuntos de pesos, um para cada dia da semana.

D. Backtesting e Otimização de Pesos do Comitê

Em fase de *Backtesting*, os agentes foram avaliados sob métricas: Retorno Acumulado, Volatilidade Anualizada, *Draw-down* Máximo (maior distância relativa do maior pico ao pico mais baixo subsequente), Índice de Sharpe (relação retorno e risco) e Índice de Calmar (relação retorno e risco).

Uma vez que concluída a fase de inferência dos agentes individuais, é possível iniciar o processo de otimização dos pesos do Comitê através do PSO, utilizando os dados de desempenho dos agentes individuais. A Tabela I apresenta os pesos otimizados pelo PSO para cada agente especialista, bem como o limiar de voto utilizado para determinar a ação final do Comitê. Esses pesos e limiares foram determinados com base no desempenho em termos de retorno total dos agentes individuais durante o período de teste pré-PSO, e serão aplicados em um período de teste subsequente para validar a eficácia da otimização dos pesos do Comitê.

TABELA I: Pesos e Limiares de Ativação do Otimizador por Enxame de Partículas (PSO)

Comitê (PSO)	Amostra	Limiar Voto	1D	1H	30M	5M
PSO Padrão	2.868	0.008500	0.3421	0.2533	0.0476	0.3567
PSO Segunda	540	0.003601	0.0372	0.2749	0.2320	0.4556
PSO Terça	540	0.004920	0.3260	-0.1636	0.1423	-0.3680
PSO Quarta	600	0.008538	0.3790	0.1409	0.1008	0.3790
PSO Quinta	648	0.006493	0.1972	-0.1748	0.2717	-0.3560
PSO Sexta	540	0.004538	0.2322	-0.3367	-0.0062	0.4247

IV. RESULTADOS E DISCUSSÃO

A Tabela II apresenta os resultados do Comitê de agentes e *benchmarks* técnicos em um período de teste de 62 dias (428 horas, 6910 passos de 5 minutos). Os resultados incluem o retorno acumulado, volatilidade anualizada, Índice de Sharpe, *Drawdown Máximo* e Índice de Calmar para cada estratégia avaliada.

TABELA II: Resultados do Comitê de Agentes e Benchmarks Técnicos em Período de Teste de 62 dias (428 horas, 6910 passos de 5 minutos).

Estratégia	Retorno (%)	Volat. (%)	Sharpe	Max DD (%)	Calmar
<i>Limiar Teórico Máximo (Oráculo)</i>					
Oráculo 5M	361,57	1,27	18,60	0,00	-
<i>Benchmarks de Referência Base</i>					
Buy & Hold	-5,66	2,63	-0,14	-16,89	-0,34
Sell & Hold	5,66	2,63	0,14	-14,25	0,40
Random Walk	0,80	1,82	0,03	-6,12	0,13
<i>Benchmarks Técnicos Intradiários</i>					
Cross Media	-5,80	1,82	-0,21	-12,08	-0,48
RSI	14,25	1,81	0,51	-3,78	3,77
Bollinger	12,16	1,81	0,44	-4,73	2,57
VWAP	-3,68	1,82	-0,13	-8,03	-0,46
MACD	-7,07	1,82	-0,26	-15,27	-0,46
<i>Agentes DDPG Individuais</i>					
Agente 1D	19,41	2,35	0,53	-3,10	6,27
Agente 1H	19,36	1,70	0,75	-2,00	9,69
Agente 30M	9,80	2,06	0,42	-8,96	1,09
Agente 5M	10,12	1,81	0,37	-4,84	2,09
Comitê Multi-Agente					
Majoritário	5,45	0,64	0,56	-1,28	4,25
Ponderado	17,02	1,76	0,63	-4,66	3,66
PSO Padrão	22,29	1,80	0,81	-1,90	11,70
PSO Semanal	11,70	1,80	0,43	-3,74	3,13

A Figura 3 apresenta a curva de retorno acumulado da estratégia de Voto Ponderado, Voto Majoritário, do Comitê PSO Padrão e PSO Semanal, comparando-os com os *benchmarks* de mercado e a melhor estratégia de referência (*RSI de 5 minutos*). Observa-se que o Comitê padrão otimizado pelo PSO superou significativamente os *benchmarks* de mercado e as estratégias de referência, evidenciando a eficácia da otimização dos pesos do Comitê para melhorar o desempenho geral do sistema multi-agente. O Comitê PSO Padrão apresentou um desempenho superior ao PSO Semanal, o que pode ser atribuído à maior capacidade de generalização do modelo otimizado em uma base mais ampla, enquanto o PSO Semanal pode ter sido mais suscetível ao ruído e à aleatoriedade do mercado em dias específicos da semana. Os primeiros 3000 passos da curva de retorno se referem ao período de teste dos agentes individuais, e as curvas dos dados otimizados pelo PSO, motivo pelo qual a curva é zerada a partir do passo 3000, dando início assim ao período de teste do Comitê otimizado pelo PSO.

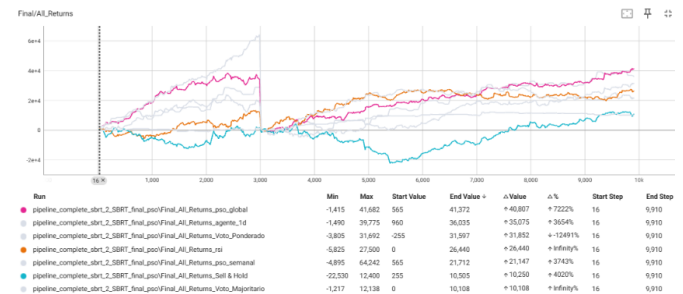


Fig. 3: Curva de Retorno Acumulado do Comitê PSO Padrão e PSO Semanal em pontos do índice.

Sobre o desempenho do Comitê PSO Padrão, observa-se que alcançou um retorno total de 22,29% e um Sharpe de 0,81, superando o *benchmark* de mercado e as estratégias

de referência. Apesar de o algoritmo ter atribuído um peso próximo a zero ao agente de 30 minutos, ação que a princípio pode parecer desfavorável, os outros parâmetros otimizados compensaram, resultando em um desempenho final positivo e superior ao *benchmark* de mercado e agentes individuais, o que evidencia a eficácia da otimização dos pesos do Comitê através do PSO para melhorar o desempenho geral do sistema multi-agente.

Sobre o desempenho do Comitê PSO Semanal, observa-se que alcançou um retorno total de 11,70% e um Sharpe de 0,43, superando 3 dos 5 *benchmarks* de mercado e as estratégias de referência, embora com um desempenho inferior ao Comitê PSO Padrão, o que evidencia que a otimização dos pesos do Comitê através do PSO é eficaz para melhorar o desempenho geral do sistema multi-agente, mesmo quando realizada em uma base semanal, porém a proposta de otimização em razão de dias específicos da semana ao perder em capacidade de generalização fica a mercê do ruído e da aleatoriedade do mercado, o que pode explicar a inferioridade do Comitê PSO Semanal em relação ao Comitê PSO Padrão.

Com relação a acurácia dos modelos, a Tabela III apresenta a taxa de acerto (acurácia) do Comitê PSO Padrão, PSO Semanal, Voto Majoritário, Voto Ponderado e dos agentes individuais, a taxa de acerto é calculada como a proporção de decisões corretas (compra ou venda) em relação ao total de decisões tomadas durante o período de teste. Este contexto não leva em consideração o retorno financeiro das decisões, mas sim a capacidade do modelo de identificar corretamente as oportunidades de compra e venda com base nos sinais gerados pelos agentes e combinados pelo Comitê.

TABELA III: Taxa de acerto (acurácia), acertos, erros e total de ações para agentes individuais e estratégias de comitê.

Estratégia	Acurácia (%)	Acertos	Erros	Total
Agente 1D	53,23	33	29	62
Agente 1H	57,03	219	165	428
Agente 30M	49,23	414	427	855
Agente 5M	49,38	2470	2532	6910
Voto Majoritário	49,84	303	305	608
Voto Ponderado	49,16	2343	2423	4766
Voto PSO	50,24	2490	2466	4956
Voto PSO Persistente	53,80	503	432	935
Voto PSO Semanal	50,89	2526	2438	4964
Voto PSO Semanal Persistente	57,02	751	566	1317

Um Voto Persistente se refere a uma decisão de compra ou venda que é mantida por um período subsequente, mesmo que os sinais dos agentes possam variar, contanto que a posição atual não mude, a ação é considerada como um único acerto ou erro, definido pelo resultado final da posição (compra ou venda) ao ser encerrada. Essa métrica existe pois é a forma real como o modelo opera em um ambiente de mercado real. Ou seja, o modelo não inicia uma nova posição a cada passo de 5 minutos, mas sim mantém uma posição aberta até que um sinal de decisão diferente seja gerado. Dessa forma é possível compreender melhor o retorno total do modelo, uma vez que obtendo 53,80% de acerto sob um contexto de ação persistente, o retorno total de 22,29% do Comitê PSO Padrão é mais compreensível.

V. CONCLUSÕES

A abordagem de aprendizado por reforço baseada em um Comitê de agentes DDPG e LSTM provou superar substancialmente o uso isolado desses modelos para análises em ambientes não-estacionários intensos. O agrupamento via Comitê ponderado pela otimização PSO gerou tomadas de decisões eficientes, proporcionando rendimento lucrativo expressivo e estabilidade comprovada pela baixa redução de portfólio (*drawdown*). Tais resultados evidenciam que tratar cenários financeiros através de inteligência artificial fracionada em escalas e unificada estrategicamente estabelece um padrão promissor. No obstante à natureza estocástica do mercado e a complexidade intrínseca dos dados financeiros, a metodologia proposta demonstrou resiliência e capacidade de adaptação, sugerindo que a combinação de técnicas de aprendizado por reforço, redes neurais de memória e otimização por enxame pode ser uma abordagem eficaz para a extração de sinais preditivos em séries com alto fator de ruído e que o contexto pode ser aplicado a outras áreas.

AGRADECIMENTOS

Os autores agradecem o apoio do Instituto Federal da Paraíba (IFPB).

REFERÊNCIAS

- [1] D. K. Morgan, Y. Zhou, L. M. Bennett, and A. Patel, "A multi-scale prediction model for stock volatility based on a hybrid attention mechanism," *Preprints*, November 2025. [Online]. Available: <https://doi.org/10.20944/preprints202511.0849.v1>
- [2] Y. Bai, Y. Gao, R. Wan, S. Zhang, and R. Song, "A review of reinforcement learning in financial applications," 2024. [Online]. Available: <https://arxiv.org/abs/2411.12746>
- [3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019. [Online]. Available: <https://arxiv.org/abs/1509.02971>
- [4] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0377221717310652>
- [5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.
- [6] PyTorch Contributors, "Pytorch documentation: Lstm," 2024, acessado em: 10 de Março de 2026. [Online]. Available: <https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html>
- [7] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," 2014. [Online]. Available: <https://arxiv.org/abs/1402.1128>
- [8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [9] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSE: Neural networks for machine learning, pp. 26–31, 2012.
- [10] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948 vol.4.
- [11] Y. Shi and R. C. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the IEEE Conference on Evolutionary Computation, ICEC*, vol. 6, 06 1998, pp. 69 – 73.
- [12] B. Mandelbrot, "The variation of certain speculative prices," *The Journal of Business*, vol. 36, no. 4, pp. 394–419, 1963. [Online]. Available: <http://www.jstor.org/stable/2350970>
- [13] J.-P. Bouchaud, "Mandelbrot, financial markets and the origins of "econophysics"," 2026. [Online]. Available: <https://arxiv.org/abs/2602.02078>
- [14] M. Al-Faiz, A. Ibrahim, and S. Hadi, "The effect of z-score standardization (normalization) on binary input due the speed of learning in back-propagation neural network," *Iraqi Journal of Information & Communications Technology*, vol. 1, pp. 42–48, 02 2019.