Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

Market-Adaptive Stock Trading through B-WEMA Driven Proximal Policy Optimization

Mulia Ichsan*, Amalia Zahra

Computer Science Department, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: [1,*]mulia.ichsan@binus.ac.id, [2]amalia.zahra@binus.edu
Correspondence Author Email: mulia.ichsan@binus.ac.id
Submitted: **04/02/2026**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstract-Developing automated trading strategies that achieve stable returns while controlling risk remains a central threat in quantitative finance. Many reinforcement learning-based trading systems focus on reward maximization but provide limited justification for the choice of forecasting indicators and often lack comprehensive benchmarking against alternative strategies and risk measures. This essay addresses the problem of integrating a statistically grounded price-smoothing technique with a policy optimization scheme to improve sequential trading decisions under market uncertainty. We propose a hybrid model that combines Brown's Weighted Exponential Moving Average (B-WEMA) as a trend-sensitive forecasting indicator with a Deep Reinforcement Learning agent trained using Proximal Policy Optimization (PPO). The role of B-WEMA is to provide structured price signals that reduce noise sensitivity, while PPO determines buy and sell actions through policy updates constrained for stable learning. The performance of the proposed model is evaluated over a 10-month trading horizon and compared with a buy-and-hold benchmark and an alternative reinforcement learning method, Advantage Actor-Critic (A2C), both implemented under the same experimental conditions. Empirical results show that the proposed B-WEMA-PPO framework achieved a cumulative return of 23.43% over the test period, outperforming both the benchmark and the A2C-based agent. In addition to cumulative return, risk-adjusted performance metrics, namely volatility and maximum drawdown, are reported to provide a balanced assessment of profitability and risk exposure. These findings suggest that incorporating structured exponential smoothing into policy optimization may enhance the stability and effectiveness of reinforcement learning-based trading strategies.

Keywords: Deep Reinforcement Learning; Proximal Policy Optimization; B-WEMA; Risk-Adjusted Trading Performance

1. INTRODUCTION

Shares represent ownership claims in a company and are traded on stock exchanges within the capital market system. In addition, the capital markets serve as mechanisms for capital allocation and investment, enabling investors to obtain returns commensurate with the risk they undertake [1]. Moreover, shareholder returns are motivated by various macroeconomic and firm-specific factors, including inflation, interest rates, fiscal conditions, dividend policies, supply-demand dynamics, and regulatory frameworks [2], [3]. In practice, investors generally employ two main approaches to stock testing. Technical testing relies on historical price and volume data, while fundamental analysis evaluates financial statements and broader economic indicators. Regardless of the approach used, risk management remains essential because investment errors may lead to significant financial losses when decisions are not assisted by rigorous analysis [4].

Furthermore, the development of quantitative techniques in technical investigation has led to various models for forecasting stock price movements across short-, medium-, and long-term horizons [5], [6]. Among the most widely used methods are the Simple Moving Average (SMA), Weighted Moving Average (WMA), and Exponential Moving Average (EMA). Comparative studies signal that the Weighted Exponential Moving Average (WEMA), derived from WMA and EMA, produces lower forecasting errors than other traditional moving average approaches. Further development combining WMA with Brown's Double Exponential Smoothing (BDES) resulted in the Brown Weighted Exponential Moving Average (B-WEMA) method. Empirical findings show that B-WEMA achieves lower mean squared error (MSE) and mean absolute percentage error (MAPE) compared to related smoothing techniques [7], [8]. However, minimizing statistical forecast error does not necessarily translate into improved trading performance. Measures such as MSE and MAPE assess point prediction accuracy but do not account for profitability, risk exposure, or the sequential nature of trading decisions. Stock trading involves dynamic interactions in which current actions influence future states and returns. Consequently, relying solely on prediction accuracy as an indicator of superiority is insufficient in automated trading.

In line with that, machine learning and deep learning methods have been widely applied to financial market analysis. Prior studies integrate fundamental data, alternative data sources, and technical indicators to generate predictive signals and investment strategies [9], [10], [11], [12]. Reinforcement learning (RL) has attracted attention because it directly models sequential decision-making problems by optimizing cumulative rewards. Among policy-gradient approaches, Proximal Policy Optimization (PPO) has demonstrated stable performance and competitive returns in trading applications [13], [14], [15]. PPO updates policies iteratively while constraining excessive policy shifts, thereby enhancing learning stability and sample efficiency [16], [17], [18], [19]. Empirical studies report that RL-based trading systems can outperform manual strategies under certain market conditions [20], [21]. Even though previous research has examined forecasting models and, separately, reinforcement learning-based trading strategies, few studies have investigated the structured integration of a smoothing-based forecasting model, such as B-WEMA, within a Deep Reinforcement Learning scheme, with evaluation focused on trading performance rather than solely on statistical error metrics. Many studies emphasize prediction accuracy or algorithmic return performance

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

independently, without explicitly examining how a smoothing-derived signal contributes to policy learning and cumulative return optimization.

This essay differs from prior research in several essential aspects. First, B-WEMA is not positioned as a standalone forecasting tool whose value is determined solely by lower MSE or MAPE; instead, it is incorporated as a feature-extraction mechanism within a sequential decision-making model. Second, the essay integrates B-WEMA with a Deep Reinforcement Learning architecture based on PPO to directly optimize cumulative returns rather than solely forecast accuracy. Third, performance evaluation emphasizes trading-oriented indicators to ensure alignment between predictive signals and economic objectives.

Thus, the objectives of this test are: (1) to develop a stock trading framework that integrates B-WEMA signals into a PPO-based Deep Reinforcement Learning model; (2) to analyze the impact of B-WEMA features on policy learning stability and cumulative return; and (3) to compare the proposed approach with benchmark strategies based on trading performance metrics. The contribution of this study lies in demonstrating how smoothing-based predictive signals can be systematically embedded within a reinforcement learning framework to support more effective automated trading decisions in the capital market.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research was conducted through a systematic sequence of stages designed to develop, train, and evaluate a stock trading model. The stages are Communication, Quick Plan, Modelling Quick Design, and Delivery & Feedback, as shown in Figure 1. Each stage is designed to ensure clarity, reproducibility, and scientific rigor in the research process, from problem identification to model evaluation. Figure 1 illustrates the sequence of the research process and the workflow of all stages.

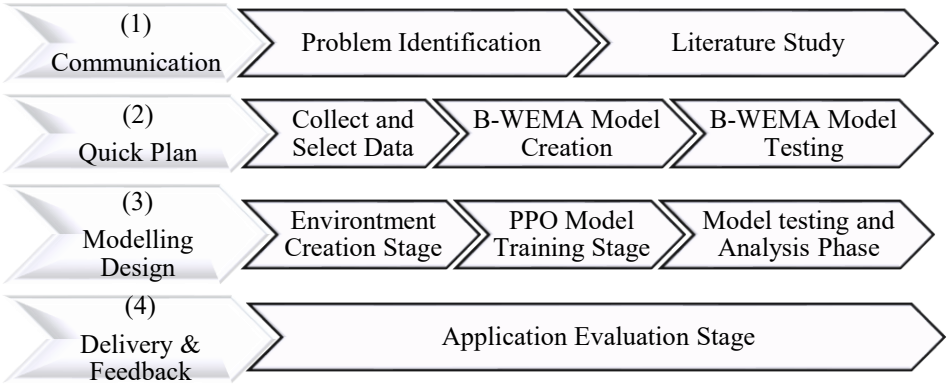

Figure 1. Shows the flowchart of the AI-based models and experimental methods applied

2.2 Communication

The Communication stage includes Problem Identification and Literature Study. The problem addressed in this research is predicting stock prices for decision-making in buying and selling shares using the B-WEMA method integrated with Proximal Policy Optimization (PPO). In the literature study, previous research on stock price prediction, reinforcement learning approaches in finance, technical analysis indicators, and deterministic forecasting methods is reviewed. It ensures the research is based on established principles and identifies gaps that the current study aims to address.

2.3 Quick Plan

The Quick Plan stage consists of Data Collection and Selection, B-WEMA Model Creation, and B-WEMA Model Testing. Data were obtained from Yahoo Finance and include five Indonesian companies: BUMI.JK, ANTM.JK, BMRI.JK, PGAS.JK, and BBRI.JK was selected to represent diverse sectors in the Indonesian market. Although the selection is limited, the companies were chosen to provide variation in stock behavior and market capitalization, allowing initial evaluation of the methodology. The dataset covers 2009-01-01 to 2022-11-14, with 17,169 data points including date, open, high, low, close, volume, ticker, and day. Table 1 summarizes the dataset.

Table 1. Shows the dataset collected

	Date	Open	High	Low	Close	Volume	Tic	Day
0	2009-01-05	957.57	1,049.97	949.17	820.69	158,999,024	ANTM.JK	0
1	2009-01-05	485	500	472.5	341.41	163,680,000	BBRI.JK	0
2	2009-01-05	1,081.63	1,093.93	1,020.18	715.49	76,117,902	BMRI.JK	0
3	2009-01-05	940	1,000	890	879.53	326,670,000	BUMI.JK	0
4	2009-01-05	1,960	1,980	1,900	1,129.86	4	PGAS.JK	0

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

	Date	Open	High	Low	Close	Volume	Tic	Day
...	....	....	...	....	....	...	....	...

Stock price movements are predicted using the B-WEMA model, computed as:

$$\text{l}WMAt + 1 = kXt + (k-1)Xt - 1 + \cdots + Xt - (n-1)$$

$$k + (k-1) + \cdots + 1 \quad (1)$$

$$St' = \alpha Xt + (1-\alpha)St' - 1. \quad (2)$$

$$St" = \alpha St' + (1 - \alpha)\, St" - 1 \quad (3)$$

$$\alpha t = \alpha t + St' + (St' - St") = 2St' - ST \quad (4)$$

$$b = \alpha/1 - \alpha * (St' - St") \quad (5)$$

$$Ft + m = \alpha t + btm \quad (6)$$

The B-WEMA predictions are later incorporated into the reinforcement learning environment's observation space as a state feature, guiding the agent's policy while still allowing learning from market dynamics.

2.4 Modelling Design

The Modelling Quick Design stage establishes the environment and reinforcement learning agent. It consists of Environment Creation, PPO Model Training, and Model Testing and Analysis. This stage ensures that the environment accurately simulates stock trading, integrates deterministic forecasts (B-WEMA) as additional state features, and allows the reinforcement learning agent to make decisions based on both learned patterns and external predictions.

2.5 Environment Creation Stage

A custom environment extending the Gym Env class was created. It includes action and observation spaces, and methods for step(), reset(), render(), buy_stock(), sell_stock(), and memory tracking. Action Space: Actions are represented as discrete values corresponding to share amounts for each stock. For five assets, the action space is (2k+1)5(2k+1)5, where kk is the maximum number of shares tradeable per asset. This formulation is consistent with prior work [22] and captures the possibilities of multi-asset trading. Observation Space: Observations include historical prices, volume, B-WEMA predictions, turbulence, log volume, daily variance, change, portfolio balance, and asset value. Turbulence measures extreme market conditions using:

$$Turbulence_t = (y_t - \mu)\, \Sigma^{-1} (y_t - \mu)' \in R, \quad (7)$$

Where yt ∈ RD denotes the stock returns for current period t, μ ∈ RD denotes the average of historical returns, and Σ ∈ RD×D denotes the covariance of historical returns. When turbulence t exceeds a threshold, indicating extreme market conditions, it is decided that all shares are sold and buying is ceased [23]. Additional features are computed as (8):

$$Log\ volume = log\ volume_t * close_t \quad (8)$$

Calculation of the daily variance is calculated by formula (9):

$$Daily\ variance = (high_t / low) / close_t \quad (9)$$

Calculation of change is calculated by the formula (10):

$$Change = (close_t - open_t) / close_t \quad (10)$$

The step() function returns the next observation, reward, done, and info. Rewards are calculated as portfolio value change:

$$End\ total\ asset = shares\ hold * current\ price\ (close\ price)$$

$$Reward = end\ total\ assets - end\ total\ asset_{earlier} \quad (11)$$

This design balances deterministic guidance from B-WEMA with reinforcement learning exploration:

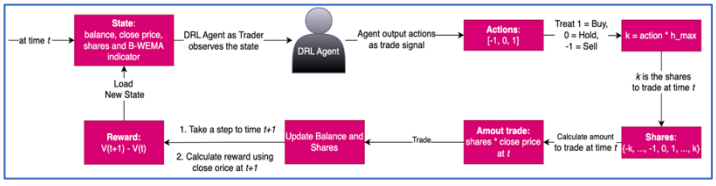

Figure 2. Flow Stock Market Environment of Deep Reinforcement Learning

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

2.6 PPO Model Training Stage

The PPO algorithm is applied using an actor-critic architecture that combines state information with B-WEMA predictions [24]. Hyperparameters (n_steps = 2048, ent_coef = 0.001, learning_rate = 0.00025, batch_size = 128) were selected based on prior studies [25] and initial tuning experiments. PPO uses clipped policy updates to maintain training stability:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{12}$$

The clipped surrogate goal function of PPO is:

$$J^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min\begin{pmatrix} r_t(\theta)\hat{A}(s_t,a_t), \\ clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}(s_t,a_t \end{pmatrix})] \tag{13}$$

This approach balances exploitation and exploration while preventing large policy swings.

2.7 Model Testing and Analysis Phase

The trained model is evaluated on data from 2022-01-01 to 2022-11-14 (1055 points). Testing follows the same environment as training. Portfolio values, trading actions, and cumulative returns are recorded. Multiple training runs are performed to ensure robustness, with early-stopping criteria applied to prevent overfitting rather than selecting solely on positive rewards.

2.8 Phase Phase Delivery and Feedback

Results are evaluated quantitatively against baseline models and qualitatively via investor feedback collected through questionnaires and online forms. This combined evaluation provides insights into practical effectiveness and identifies opportunities for improvement.

3. RESULT AND DISCUSSION

3.1 Result of B-WEMA Testing

In the B-WEMA testing phase, PGAS stock price movements.JK was predicted over 365 days from March 3, 2021, to August 31, 2022, with the last 21 days serving as the test set. Various alpha values ranging from 0.1 to 0.9 were iterated to identify the value that minimized the Mean Absolute Percentage Error (MAPE), with an optimal alpha of 0.5 yielding a MAPE of 0.021 (Figure 3).

Figure 3. Result of Testing Prediction B-WEMA on 21 Days

While this result demonstrates the model's ability to fit historical data closely, achieving a 2.1% error over a 21-day horizon is extraordinarily low for financial time series, which are typically highly volatile and influenced by numerous external factors. Such performance raises concerns about overfitting and underscores the need for validation against simple benchmarks, such as naive forecasts or moving averages, to assess whether the predictive performance is genuinely informative or merely a result of tailoring to the historical dataset. Nevertheless, the B-WEMA output was incorporated into the DRL environment to enhance the observation space for the PPO agent, potentially improving decision-making. However, the propagation of prediction errors into trading actions remains a concern.

3.2 Result of Proximal Policy Optimization Training

The PPO agent was trained on historical data spanning from January 1, 2009, to January 1, 2022, using a total of 14,875 timesteps, with a reported training duration of only 23.81 minutes (Figure 4). During training, the returns' standard deviation increased from 1.02 to 1.2 over the timesteps. Although this was interpreted as a balance between exploration and exploitation, the increase in variability could also indicate instability or insufficient convergence of the agent. The remarkably short training duration for PPO, which is typically computationally intensive, raises

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

additional doubts about whether the agent adequately explored the state-action space and learned robust trading strategies. These concerns are important because limited exploration and shallow training could compromise the agent's ability to generalize to unseen market conditions.

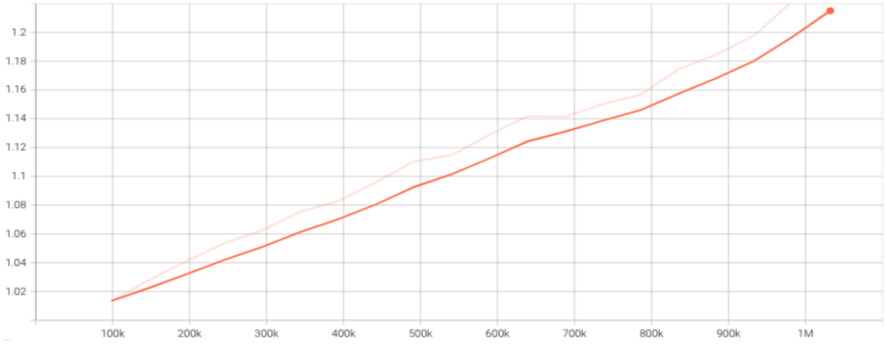

Figure 4. Visualization of Standard Deviation during Training

3.3 Result of Proximal Policy Optimization Testing

During the PPO testing phase, the trained agent was evaluated on five stocks, including BUMI.JK, ANTM.JK, BBRI.JK, BMRI.JK, and PGAS.JK-over the period from January 1, 2022, to November 14, 2022, with an initial capital of IDR 150,000,000 and transaction costs included (Table 2).

Table 2. Shows the summary result of testing

	date	cash	account_value	reward
0	2022-01-04	146003019.807	149405613.483	-594386.516
1	2022-01-05	141953933.331	148620546.149	-785067.334
2	2022-01-06	137987298.705	147962850.255	-657695.893
...	...	...	...	...
207	2022-11-09	-3089.973	180042410.026	-203500.0
208	2022-11-10	-3089.973	179050210.026	-992200.0
209	2022-11-11	-3089.973	184413910.026	5363700.0
	date	cash	account_value	reward
0	2022-01-04	146003019.807	149405613.483	-594386.516
1	2022-01-05	141953933.331	148620546.149	-785067.334
2	2022-01-06	137987298.705	147962850.255	-657695.893
...	...	...	...	...
207	2022-11-09	-3089.973	180042410.026	-203500.0
208	2022-11-10	-3089.973	179050210.026	-992200.0
209	2022-11-11	-3089.973	184413910.026	5363700.0

The cumulative return achieved was 23.43%, with an annualized return of 28.74% and a final asset value of Rp. 184,413,910, as illustrated in Figure 5. While these returns appear promising, their interpretability is limited without comparison to market benchmarks such as the Jakarta Composite Index or a simple buy-and-hold strategy for the selected stocks.

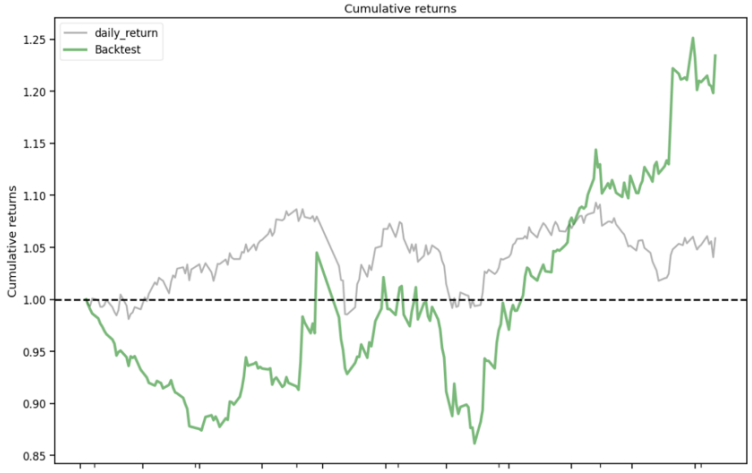

Figure 5. Backtest Plot Result PPO of Cumulative Return

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

Furthermore, Figure 5 presents cumulative returns without including critical risk metrics such as maximum drawdown or drawdown duration, and the Sharpe Ratio analysis (Figure 6) reports an average of 1.18. Without context regarding market volatility, this metric alone does not fully capture the portfolio's risk-adjusted performance.

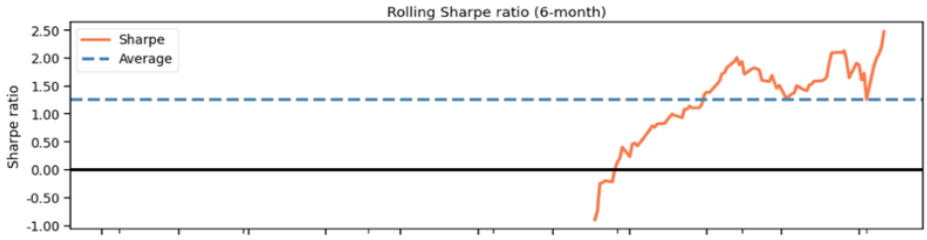

Figure 6. Backtest Plot Result Sharpe Ratio of PPO

Figure 7 provides a histogram of buy/sell/hold decisions, indicating that PGAS.JK shares were sold only in February 2022, while BUMI shares were sold in February 2022.JK shares were frequently purchased. However, this visualization lacks information on the profitability, timing, or quality of individual decisions, limiting the ability to assess the strategy's effectiveness in practical trading scenarios:

Figure 7. Plot Result History Decisions

Figure 7 above shows the automatic distribution of decisions on cumulative monthly sales and purchases. Based on the test results, it shows that there was only a sale of PGAS.JK shares in February 2022. Meanwhile, when it comes to purchases, the shares that were bought the most are BUMI.JK.

3.4 Result Comparison with Other Deep Reinforcement Learning Model

A comparison with the Advantage Actor-Critic (A2C) model was conducted using the same test data and initial capital. The A2C model produced a final account value of Rp. 167,155,600, a cumulative return of 11.74%, an annualized return of 14.25%, and an average Sharpe ratio of 0.73, as shown in Figure 8.

Figure 8. Backtest Plot Result A2C of Cumulative Return

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

While PPO outperformed A2C in this experiment, it is unclear whether the A2C model was fully optimized, and the performance differences may reflect hyperparameter selection or insufficient training rather than PPO's intrinsic superiority. In addition, neither model was tested across multiple market regimes or periods of volatility, making claims that PPO can generalize learned trading behavior to unseen data premature without extensive cross-validation and out-of-sample testing.

3.5 Discussion

The results obtained from the B-WEMA and Proximal Policy Optimization (PPO) models highlight both the potential and the limitations of applying predictive modeling and deep reinforcement learning (DRL) to stock trading. The integration of B-WEMA into the DRL scheme was intended to improve the accuracy of stock price predictions, thereby informing the PPO agent's trading decisions. In this essay, the B-WEMA model achieved a Mean Absolute Percentage Error (MAPE) of 0.021 for a 21-day prediction horizon on PGAS.JK stock. While this result appears highly favorable, it is essential to contextualize it within the complexity and unpredictability of financial markets.

Financial time series are inherently noisy, affected by exogenous events such as macroeconomic shifts, corporate news, investor sentiment, and geopolitical factors. Consequently, a MAPE of 2.1% over three weeks is unusually low, suggesting a need for more rigorous validation. Without comparison with simple benchmarks, such as naive predictions that assume no change in stock prices or basic moving-average forecasts, it is difficult to ascertain whether the model genuinely captures underlying market dynamics or merely overfits historical data. Overfitting in predictive models can lead to a false sense of accuracy, where predictions align closely with past observations but fail to generalize to new or unseen market conditions. Such a limitation is particularly relevant when integrating predictions into a DRL environment, as errors in the observation space can propagate and affect decision-making. Applying the PPO agent to this enriched dataset yielded insights into how reinforcement learning could generate automated trading strategies. The agent was trained on a historical dataset spanning over 12 years, from January 1, 2009, to January 1, 2022, using 14,875 timesteps. Remarkably, this training was completed in only 23.81 minutes, which is unusually short for a PPO algorithm, typically computationally intensive due to the requirement to repeatedly sample trajectories, calculate advantages, and perform gradient updates. It raises questions about whether the agent had sufficient opportunity to explore the state-action space and converge to a stable policy. Training depth and duration are critical factors in reinforcement learning, as inadequate exploration can lead to suboptimal policies that perform well only on the training data but fail to generalize to changing market conditions.

Additionally, Figure 4 illustrates an increase in standard deviation during training, rising from 1.02 to 1.2. While the original interpretation suggested that this increase represents a balance between exploration and exploitation, it is equally plausible that the agent was experiencing instability in learning, which could undermine the reliability of the resulting policy. In reinforcement learning, a stable standard deviation over training is often desirable, as excessive fluctuations may indicate inconsistent updates or insufficient learning. In the testing, the PPO agent was applied to five stocks, including BUMI.JK, ANTM.JK, BBRI.JK, BMRI.JK, and PGAS.JK-over a period from January 1, 2022, to November 14, 2022, starting with a capital of IDR 150,000,000. The agent achieved a cumulative return of 23.43%, an annualized return of 28.74%, and a final asset value of Rp. 184,413,910 (Figures 5-7). At first glance, these results appear promising, indicating that the PPO agent generated profitable trades in the backtest environment. However, interpreting these figures requires caution. Financial backtests often present inflated performance metrics due to several factors, including look-ahead bias, survivorship bias, and the absence of real-world trading frictions such as latency, slippage, and liquidity constraints.

Moreover, the observed returns are presented without reference to market benchmarks. For instance, the Jakarta Composite Index or a simple buy-and-hold strategy over the same period could have achieved returns similar to or better than the PPO agent's, making it difficult to assess whether the PPO agent's performance represents an actual improvement over standard investment strategies. In the absence of such comparisons, the reported cumulative and annual returns remain difficult to contextualize. In addition to returns, risk-adjusted performance metrics are crucial for evaluating any trading strategy. Hence, in this essay, the average Sharpe Ratio during the testing period was reported as 1.18 (Figure 6), which is generally considered acceptable. However, without information on the volatility of the market or benchmark assets during the same period, the Sharpe Ratio alone provides limited insight. The metric is most informative when compared against alternative strategies or indices over the same timeframe.

Henceforth, Figure 5 presents cumulative returns but does not provide information on maximum drawdown, drawdown duration, or intra-period volatility, all of which are essential for understanding the strategy's risk profile. High cumulative returns may be accompanied by periods of significant losses, which could be unacceptable for risk-averse investors. Therefore, the lack of detailed risk metrics limits the ability to evaluate the PPO agent's suitability for practical deployment fully. The analysis of trading decisions, shown in Figure 7, provides a histogram of buy, sell, and hold actions over the testing period. While this visualization indicates trading frequency, it does not capture the profitability, timing, or strategic rationale of each decision. For instance, the agent sold PGAS.JK shares only in February 2022, while BUMI.JK shares were purchased frequently. However, frequent transactions do not necessarily translate to higher profits and may even incur greater transaction costs in real trading scenarios. The quality of each trade, in terms of both timing and market context, is critical to understanding whether the agent's policy is truly effective or simply opportunistic under favorable historical conditions. A comparison with the Advantage Actor-Critic (A2C) model was conducted using the same test data and initial capital. The A2C model achieved a final account

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

value of Rp. 167,155,600, a cumulative return of 11.74%, an annualized return of 14.25%, and an average Sharpe ratio of 0.73 (Figure 8). While PPO outperformed A2C in this experiment, several factors must be considered before concluding that PPO is superior. First, it is unclear whether the A2C model was fully optimized; differences in hyperparameters, network architecture, or training duration could account for its lower performance. Second, both models were evaluated on a single market period, limiting conclusions about generalizability. Market conditions are dynamic, and a strategy that performs well in one period may underperform under different volatility regimes, macroeconomic conditions, or unexpected events. Therefore, claims that PPO can generalize learned trading behavior to unseen data are premature without additional testing, such as k-fold cross-validation, rolling window backtesting, or evaluation across multiple market environments. Several broader limitations must be emphasized. The use of historical data for both model training and evaluation introduces the risk of overfitting, leading to strategies that appear profitable in backtests but fail in live markets.

In addition, the simplicity of the B-WEMA prediction model may overlook other critical market signals, including macroeconomic indicators, investor sentiment, and news events, which could affect real-world trading performance. The impact of transaction costs, execution delays, and slippage in live trading is another critical consideration, as these factors can erode the theoretical profitability reported in backtests. Furthermore, the lack of detailed risk metrics, such as maximum drawdown, drawdown duration, and conditional value-at-risk, limits the ability to assess the trading strategy's robustness under adverse market conditions. Despite these limitations, the study provides important insights into the potential integration of predictive models with DRL for automated trading.

The combination of B-WEMA predictions and PPO-based decision-making suggests that time-series forecasting can enhance reinforcement learning. The results suggest that, when properly tuned and validated, such integration could contribute to more informed decision-making and potentially higher returns than baseline strategies. However, achieving robust performance requires careful attention to model validation, hyperparameter optimization, risk management, and benchmark comparisons. Future research should explore these aspects, including testing across different market conditions, incorporating additional market features, extending the evaluation period, and assessing the DRL agent's sensitivity to prediction errors and market volatility. In sum, the essay demonstrates that while the B-WEMA and PPO combination shows potential for generating profitable trades, the results must be interpreted cautiously. Unusually low prediction errors, rapid training times, and the absence of robust benchmark comparisons limit the conclusiveness of the findings.

Nonetheless, integrating predictive modeling with DRL represents a promising avenue for automated trading research. Addressing the limitations identified in this study, including rigorous validation, comprehensive risk assessment, and multi-period testing, will be essential for establishing confidence in the practical applicability of such models. To conclude, future work should also explore the interplay between prediction accuracy and reinforcement learning policy stability, as well as the impact of real-world constraints on strategy performance, to ensure that observed backtest results translate into sustainable, risk-adjusted returns in live trading environments.

4. CONCLUSION

This essay investigated integrating the B-WEMA predictive model with a Deep Reinforcement Learning (DRL) framework using Proximal Policy Optimization (PPO) to support automated stock trading. While the experimental results suggest that the PPO agent could achieve favorable cumulative and risk-adjusted returns compared to the A2C model, a careful inspection indicates that these outcomes must be interpreted with caution. The B-WEMA model produced an unusually low prediction error for a 21-day horizon, and the PPO agent was trained in an exceptionally short period, raising questions about the depth of training and the generalizability of the resulting policy. Moreover, the lack of rigorous benchmark comparisons with market indices or simple trading strategies limits the ability to assess whether the model genuinely outperforms conventional approaches. The essay also reveals that the current actualisation is relatively simplistic. The evaluation of trading decisions, presented as the frequency of buy, sell, and hold actions, provides little insight into the quality, timing, or profitability of individual trades. Critical risk metrics, such as maximum drawdown, drawdown duration, and volatility-adjusted performance, were not comprehensively assessed, leaving uncertainty about the agent's robustness under adverse market conditions. While PPO outperformed A2C in cumulative returns and the Sharpe ratio, this does not necessarily indicate a fundamental superiority of PPO. Differences in performance may be influenced by hyperparameter selection, model architecture, or the specific market period used for testing, rather than the inherent advantage of one algorithm over the other. In addition, the essay stresses several limitations of the current model and the caution required when interpreting its potential. The absence of portfolio management strategies, dynamic risk management mechanisms, and explicit rules for cutting losses or taking profits indicates that the system is not yet suitable for real-world deployment. Integration with a stockbroker for live trading, while an appealing idea, remains premature due to insufficient validation, risk analysis, and simulation of execution constraints, such as latency, slippage, and liquidity limitations. These factors are critical in determining whether backtested performance can translate into sustainable results in a live trading environment. Despite these limitations, the research demonstrates the potential of combining predictive modeling with DRL-based decision-making. The integration of B-WEMA forecasts into the agent's observation space represents a promising approach for guiding sequential trading decisions. Yet, the findings underscore the need for further work to establish the model's

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

reliability and applicability. Future research should focus on robust cross-validation across multiple market periods, the inclusion of comprehensive risk management and portfolio allocation strategies, and rigorous benchmarking against naive and index-based strategies. A detailed analysis of the agent's trading behavior, including profitability and decision timing, will also be essential to assess whether the learned policy generalizes effectively across diverse market conditions. In short, while the current implementation of B-WEMA and PPO shows potential for generating profitable trades under backtesting conditions, it remains an exploratory model. Its simplicity and incomplete risk management highlight that the results should be interpreted as preliminary. Advancing this research will require extensive validation, improved risk assessment, and the integration of more sophisticated trading and portfolio management strategies to move closer to practical, real-world application.

REFERENCES

[1] A. Maharani and F. Saputra, "Relationship of Investment Motivation, Investment Knowledge and Minimum Capital to Investment Interest," *Journal of Law, Politic and Humanities*, vol. 2, no. 1, pp. 23–32, 2021, doi: 10.38035/jlph.v2i1.84.

[2] A. F. Kamara, E. Chen, and Z. Pan, "An ensemble of a boosted hybrid of deep learning models and technical analysis for forecasting stock prices," *Inf. Sci. (N Y).*, vol. 594, pp. 1–19, May 2022, doi: 10.1016/j.ins.2022.02.015.

[3] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications," *International Journal of Financial Studies*, vol. 11, no. 3, p. 94, Jul. 2023, doi: 10.3390/ijfs11030094.

[4] D. A. Daniswara, H. Widjanarko, and K. Hikmah, "THE ACCURACY TEST OF TECHNICAL ANALYSIS OF MOVING AVERAGE, BOLLINGER BANDS, AND RELATIVE STRENGTH INDEX ON STOCK PRICES OF COMPANIES LISTED IN INDEX LQ45," *Indikator: Jurnal Ilmiah Manajemen dan Bisnis*, vol. 6, no. 2, p. 16, Apr. 2022, doi: 10.22441/indikator.v6i2.14806.

[5] U. W. Chohan and S. Van Kerckhoven, *Activist Retail Investors and the Future of Financial Markets*, 1st ed., vol. 1. London: Routledge, 2023. doi: 10.4324/9781003351085.

[6] A. Coloma-Carmona, J. L. Carballo, F. Miró-Llinares, and J. C. Aguerri, "Not all traders gamble, but some gamblers trade: a latent class analysis of trading and gambling behaviors among retail investors," *Public Health*, vol. 244, p. 105742, Jul. 2025, doi: 10.1016/j.puhe.2025.105742.

[7] K. D. Pradnyani, I. M. S. Sandhiyasa, and I. M. A. O. Gunawan, "Optimising Double Exponential Smoothing for Sales Forecasting Using The Golden Section Method," *Jurnal Galaksi*, vol. 1, no. 2, pp. 110–120, Aug. 2024, doi: 10.70103/galaksi.v1i2.21.

[8] D. P. Anggraeni, "Optimisation of Inventory Management Through Time Series Analysis of Inventory Data with Double Exponential Smoothing Method," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 3, pp. 1693–1700, Jul. 2024, doi: 10.47709/cnahpc.v6i3.4410.

[9] S. K. Sahu, A. Mokhade, and N. D. Bokde, "An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges," *Applied Sciences*, vol. 13, no. 3, p. 1956, Feb. 2023, doi: 10.3390/app13031956.

[10] M. Saberironaghi, J. Ren, and A. Saberironaghi, "Stock Market Prediction Using Machine Learning and Deep Learning Techniques: A Review," *AppliedMath*, vol. 5, no. 3, p. 76, Jun. 2025, doi: 10.3390/appliedmath5030076.

[11] K. Olorunnimbe and H. Viktor, "Deep learning in the stock market—a systematic survey of practice, backtesting, and applications," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2057–2109, Mar. 2023, doi: 10.1007/s10462-022-10226-0.

[12] D. Sheth and M. Shah, "Predicting stock market using machine learning: best and accurate way to know future stock prices," *International Journal of System Assurance Engineering and Management*, vol. 14, no. 1, pp. 1–18, Feb. 2023, doi: 10.1007/s13198-022-01811-1.

[13] H. Noor, M. Shahbaz, and W. Ali, "Risk-Aware Proximal Policy Optimization for Time-Series Options Trading," *Multiagent and Grid Systems*, vol. 21, no. 3–4, pp. 209–227, Nov. 2025, doi: 10.1177/15741702251398696.

[14] H. Feng, Y. Wang, S. Zhong, T. Yuan, and Z. Quan, "Federated Reinforcement Learning in Stock Trading Execution: The FPPO Algorithm for Information Security," *IEEE Access*, vol. 13, pp. 25074–25086, 2025, doi: 10.1109/ACCESS.2025.3538859.

[15] W. Wen, Y. Yuan, and J. Yang, "Reinforcement Learning for Options Trading," *Applied Sciences*, vol. 11, no. 23, p. 11208, 2021, doi: 10.3390/app112311208.

[16] S. Sha, Y. Liu, and B. Huo, "Dynamic proximal policy optimization: Enhancing PPO with adaptive entropy and smooth clipping," *Neurocomputing*, vol. 674, p. 132861, Apr. 2026, doi: 10.1016/j.neucom.2026.132861.

[17] Z. Wang, W. Jiang, R. Peng, Q. Kou, L. Wan, and X. Lan, "Improving Sample Efficiency Through Stability Enhancement in Deep-Reinforcement Learning," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 55, no. 9, pp. 6164–6176, Sep. 2025, doi: 10.1109/TSMC.2025.3578050.

[18] Y. Cheng, Q. Guo, and X. Wang, "Proximal Policy Optimization With Advantage Reuse Competition," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3915–3925, Aug. 2024, doi: 10.1109/TAI.2024.3354694.

[19] J. Zhang and J. Xie, "Adaptive Portfolio Optimization via PPO-HER," *Journal of Global Trends in Social Science*, vol. 2, no. 4, pp. 23–30, Apr. 2025, doi: 10.70731/6xd2xq47.

[20] A. A. S. Gunawan, S. Bilqis Ashifa, R. Y. Rumagit, and H. Ngarianto, "Development of Stock Market Price Application to Predict Purchase and Sales Decisions Using Proximal Policy Optimization Method," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, IEEE, 2021, pp. 431–437. doi: 10.1109/ICCSAI53272.2021.9609714.

[21] F. Espiga-Fernández, Á. García-Sánchez, and J. Ordieres-Meré, "A Systematic Approach to Portfolio Optimization: A Comparative Study of Reinforcement Learning Agents, Market Signals, and Investment Horizons," *Algorithms*, vol. 17, no. 12, p. 570, Dec. 2024, doi: 10.3390/a17120570.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2416−2425
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9349

[22] M. Kong and J. So, "Empirical Analysis of Automated Stock Trading Using Deep Reinforcement Learning," *Applied Sciences*, vol. 13, no. 1, p. 633, Jan. 2023, doi: 10.3390/app13010633.
[23] Y. Ansari *et al.*, "A Deep Reinforcement Learning-Based Decision Support System for Automated Stock Market Trading," *IEEE Access*, vol. 10, pp. 127469–127501, 2022, doi: 10.1109/ACCESS.2022.3226629.
[24] C. Quintero, D. Leon, J. Sandoval, and G. Hernandez, "Deep Reinforcement Learning in Continuous Action Spaces for Pair Trading: A Comparative Study of A2 C and PPO," *SN Comput. Sci.*, vol. 6, no. 5, p. 407, Apr. 2025, doi: 10.1007/s42979-025-03854-0.
[25] J. Zou, J. Lou, B. Wang, and S. Liu, "A novel Deep Reinforcement Learning based automated stock trading system using cascaded LSTM networks," *Expert Syst. Appl.*, vol. 242, p. 122801, May 2024, doi: 10.1016/j.eswa.2023.122801.