Journal of Islamic Monetary Economics and Finance, Vol. 8, No.2 (2022), pp. 181 - 200
p-ISSN: 2460-6146, e-ISSN: 2460-6618

ISLAMIC STOCK PORTFOLIO OPTIMIZATION USING DEEP REINFORCEMENT LEARNING

Taufik Faturohman[1] and Teguh Nugraha[2]

[1] School of Business and Management Institut Teknologi Bandung, Indonesia,
taufik.f@sbm-itb.ac.id
[2] School of Business and Management Institut Teknologi Bandung, Indonesia,
teguh_nugraha@sbm-itb.ac.id

ABSTRACT

The Islamic principles in identifying stocks as Shari'ah principles have inevitability restrict the number of stocks that Muslims can invest in and consequently may affect the return from investment. In this paper, we examine the potential of Deep Reinforcement Learning in optimizing the portfolio returns of Islamic stocks. We model stock trading as a Markov Decision Process problem because of its stochastic and interactive nature. Then, we define the trading objective as a problem of maximization, while the DRL agents used are actor-critic algorithms. The selected portfolio consists of 30 most liquid Islamic stocks in Indonesia that constitute JII index and compare with that of the benchmark portfolio, namely the 45 most liquid conventional stocks or LQ45. The performance is compared using several algorithms The result show that trading on Islamic stocks from January 2019 to December 2020 using the DRL agents could outperform the benchmark index of conventional stocks. Using DRL agents, fund managers would be able to optimize the portfolio on daily basis, minimize risk during crisis or turbulence, and outperform the conventional stocks.

Keywords: Deep reinforcement learning, Actor-critic framework, Islamic stock.
JEL classification: C61; C63; G11.

Article history:
Received : July 31, 2021
Revised : November 22, 2021
Accepted : May 15, 2022
Available online : May 31, 2022

https://doi.org/10.21098/jimf.v8i2.1430

I. INTRODUCTION

1.1. Background

Indonesia is a country with the largest Muslim population in the world. Based on the 2018 census, more than 230 million people or 86.7% of Indonesia's population are Muslims. Furthermore, Muslims are restricted to investing only in assets that are not contrary to Islamic or shariah principles. The implementation of the principles resulted in the establishment of the Islamic capital market.

The Islamic capital market in Indonesia has grown rapidly since the launching of Islamic stock index. According to the Indonesia Stock Exchange (IDX), the number of Islamic stocks has increased by 82% from 59 to 438 or about 60% of all stocks in IDX. In parallel, the number of Islamic stock investors also increased drastically by more than 16,789% from 531 in 2011 to 89,678 in January 2021.

Figure 1 below shows the increasing market capitalization of the Jakarta Islamic Index (JII) and the Indonesian Sharia Stock Index (ISSI) over the last 20 years. Note that the maket capitalization of the Islamic stocks is about only 48% of total market capitalization of the Indonesian Composite Index. Additionally, the ratio of Islamic market capitalization to GDP (Gross Domestic Products) of Indonesia is still below 29%. Given the Islamic stock market is still low in size, it has high potential to develop to cater the investment need of the large Muslim population.

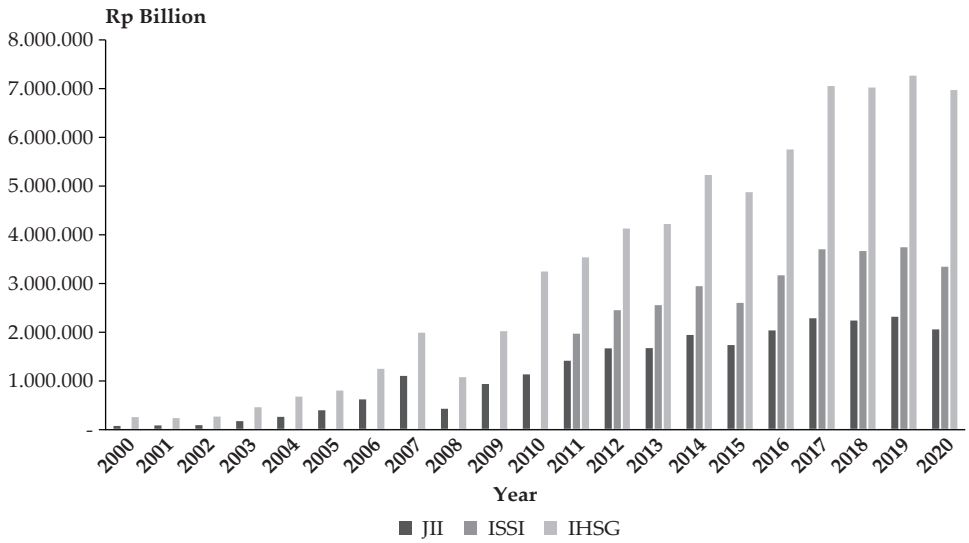


Figure 1.
The Market Capitalization of JII and ISSI Compared to IHSG

Furthermore, the comparison of the JII against the LQ45 (i.e. fourty five most liquid stocks in Indonesia) reveals that from early 2016 until Q3 of 2017, the JII index performed better. However, after Q3 2017, the LQ45 index consistently yielded higher cumulative return for more than three years. Figure 2 below shows the cumulative return of both indices in the last five years. The figure also shows that the trends of JII and LQ 45 cumulative returns are the same.

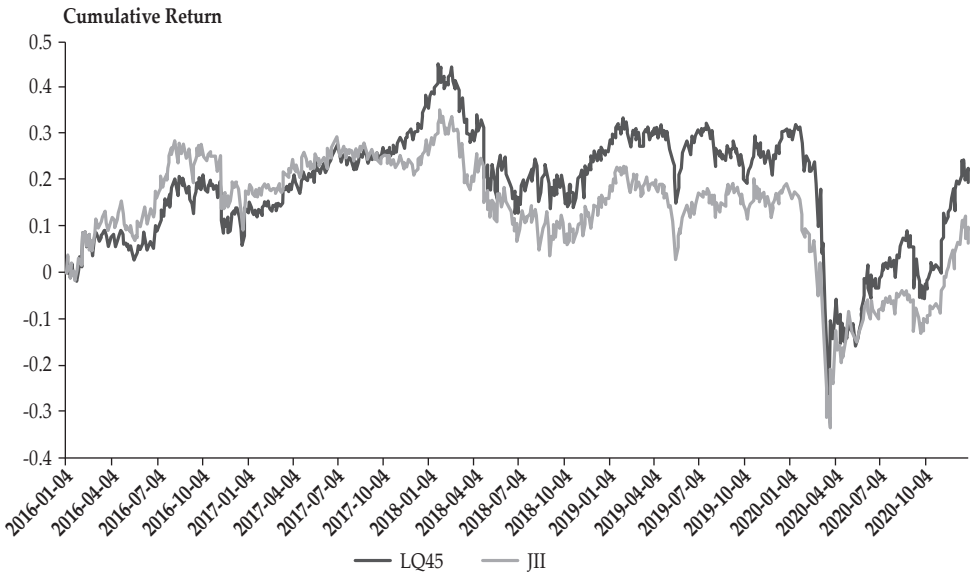


Figure 2.
Cumulative Return of JII and LQ45 Index from 2016 until 2020

To optimize capital allocation and consequently enhance expected return, a stock trading strategy is required. Estimates of a stock's future return and risk are used to maximize returns. Analysts, on the other hand, find it difficult to consider all important factors in the complex stock market. Stock trading basically involves making dynamic decisions in a highly stochastic and complex stock market, such as decisions on what to sell, at what price, and in what quantity.

Markowitz (1952) describes a standard strategy with two phases. The expected stock return and the stock price covariance matrix are computed first. The best portfolio allocation strategy can therefore be found by maximizing return for a specific risk ratio or minimizing risk for a predetermined return. However, since portfolio managers may want to change their selections at each step and have other considerations such as transaction cost, this technique is complex and costly to implement.

Machine learning and deep learning algorithms have been widely used for financial market prediction and classification models. The algorithms are coupled with earning reports and market data (market sentiment, credit card transactions, etc.) to develop new investment alphas or forecast a company's stock price. However, instead of distributing the assets or shares between stocks, these methodologies are only focused on choosing high-performing stocks.

In contrast to a regression/classification model that predicts the likelihood of future events, deep reinforcement learning (DRL) model utilizes a reward function to optimize future rewards. In many difficult games, DRL programs can surpass human players. For instance, DeepMind's AlphaGo, which uses a DRL algorithm, defeated world champion Lee Sedol in the game of Go in March 2016. Large labeled training datasets are also not required for DRL. This is a key benefit because, as

the amount of data rises exponentially, labeling a huge dataset becomes extremely time- and labor-intensive.

The exploration-exploitation method in DRL strikes a balance between trying out new things and leveraging on what has already been discovered. In comparison to other learning algorithms, this is a unique feature. Furthermore, the agent is encouraged to explore areas that have yet to be explored by human experts during the exploration phase. The stock trading process can be modelled as a Markov Decision Process (MDP), which is the backbone of DRL.

According on Bekiros (2010) and Xiong et al. (2018), the actor-critic method of DRL has recently been used in finance. Xiong et al. (2018) mention that the actor-critic method has been used for video games like Doom and has proved to be able to learn and adapt to large and complex environments. As a result, the actor-critic method is ideally suited to stock trading with a big portfolio. Three DRL algorithms of the actor-critic method are Advantage Actor Critic (A2C), Deep Deterministic Policy Gradient (DDPG), and Proximal Policy Optimization (PPO).

In this study, the three DRL algorithms (A2C, DDPG, and PPO) that automatically learn a stock trading strategy by maximizing investment return are explored to optimize the Islamic stocks portfolio return. Their performance metrics will be compared and evaluated.

1.2. Objective

This research evaluates and determines the best trading agent for Islamic stocks in the JII based on three DRL algorithms, namely A2C, DDPG, and PPO. We also compare the performance metrics of Islamic stock trading using the DRL algorithms against conventional index LQ45. Using only Islamic stocks in the portfolio will reduce the investment universe and create a less diversified investment portfolio (Faturohman et al. 2021). By narrowing the investment universe and making the investment less diversified, the portfolio will have higher idiosyncratic risk (Barnett & Salomon, 2006). Furthermore, portfolio with no constraints will most likely outperform constrained portfolios (Adler & Krizman (2008). Therefore, this study fills the gap in the literature by determining the best trading agent for limited investment universe, which in this case is the Islamic stocks, to outperform conventional stocks in Indonesia.

II. BACKGROUND AND LITERATURE REVIEW

2.1. Indonesia Stock Exchange and Islamic Stock Index

The Indonesia Stock Exchange, commonly referred as IDX, is a stock exchange based in Jakarta, Indonesia. It was originally known as the Jakarta Stock Exchange (JSX), but after merging with the Surabaya Stock Exchange (SSX) in 2007, it was renamed. According to IDX monthly statistics as of June 2021, 734 stocks are listed with market capitalization more than IDR 7.107 trillion.

The Islamic Capital Market refers to capital market activities that do not violate Islamic principles. Indonesia's Islamic capital market is a segment of the Islamic financial industry regulated by the Financial Services Authority (OJK), specifically the directorate of Islamic capital markets. The OJK issued the Regulation Number 15/POJK.04/2015 regarding the implementation of Islamic principles in the capital market in order to make the application of Islamic principles in the Indonesian capital market more binding and have legal certainty. The regulation establishes the *akad* (contracts) that can be utilized in every issue of Islamic securities in the Indonesian capital market. As long as the contracts comply with the Islamic principles, they can be utilized in the issuing of Islamic securities. *Ijara*, *istisna'*, *kafala*, *mudarabah*, *musharakah*, and *wakala* contracts can be utilized in the issuing of Islamic securities in the Indonesian capital market.

With the introduction of the Indonesia Sharia Stocks Index (ISSI), the growth of Indonesia's Islamic capital market has already begun. The ISSI, which was launched on May 12, 2011, is a composite index of Islamic stocks traded on the IDX. The index is a measure of how well Indonesia's Islamic stock market is performing. ISSI Constituents are Islamic stocks that are listed on IDX and are included in the OJK's List of Islamic Securities (DES). Out of 734 stocks traded in the IDX, there are 438 stocks that constitute the ISSI as of June 2021.

Following the DES review schedule, the ISSI constituents are re-selected twice a year, in May and November. As a result, there are Islamic equities that have exited or entered the ISSI constituents at each selection period. The ISSI calculation technique uses the weighted average based on market capitalization similar to the other IDX stock index computation method, with December 2007 as the base year.

On July 3, 2000, the Jakarta Islamic Index (JII) became the first Islamic stock index to trade on the Indonesia stock exchange. The 30 most liquid Islamic shares listed on IDX make up the JII constituents. The constituents are re-selected twice a year, every May and November. The following are the liquidity criteria utilized by the IDX in the selection of 30 Islamic stocks on the JII constituents:

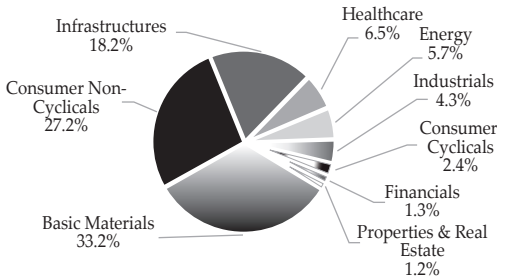
- The stocks are consistently shariah-compliant or re-selected as Islamic stocks for the last 6 months
- Sixty stocks with the highest average market capitalization in the past 1 year are selected.
- Of the 60 stocks, then 30 stocks with highest average daily transaction value in the regular market are selected.

Figure 3 below describes the characteristics of JII constituents based on data from idx.co.id as of Jan 2021. JII has market capitalization of IDR 1,965 trillion or about 29% of the total in the IDX. The index is dominated by basic materials, consumer non-cyclicals, and infrastructure sectors, which contribute 33.2%, 27.2% and 18.2% of the market capitalization respectively.

INDEX CHARACTERISTICS

Number of Constituents	30
Market Cap Total (in tr. IDR)	1,965.13
Largest	308.08
Samallest	10.14
Average	65.50
Median	45.28
Index MC/IDX MC (%)	29.06%

SECTOR WEIGHTS



TOP 10 CONSTITUENTS

Code	Company Name	Market Cap (in trill. IDR)	Index Weight	Sector
TLKM	Telkom Indonesia (Persero) Tbk.	308.08	15.68%	Infrastructures
UNVR	Unilever Indonesia Tbk.	264.19	13.44%	Consumer Non-Cyclicals
TPIA	Chandra Asri Petrochemical Tbk.	181.46	9.23%	Basic Materials
ICBP	Indofoof CBP Sukses Makmur Tbk.	106.12	5.40%	Consumer Non-Cyclicals
CPIN	Charoen Pokphand Indonesia Tbk.	94.29	4.80%	Consumer Non-Cyclicals
UNTR	United Tractors Tbk.	85.23	4.34%	Industrials
BRPT	Barito Pacific Tbk.	82.65	4.21%	Basic Materials
INKP	Indah Kiat Pulp & Paper Tbk.	70.58	3.59%	Basic Materials
KLBF	Kalbe Farma Tbk.	68.67	3.49%	Healthcare
SMGR	Semen Indonesia (Persero) Tbk.	62.87	3.20%	Basic Materials
	Total (Top 10 Constituents)	**1,324.14**	**67.38%**	

Data prior to the launch date is back-tested data. Data as of: Jan 29, 2021
Source: idx.co.id

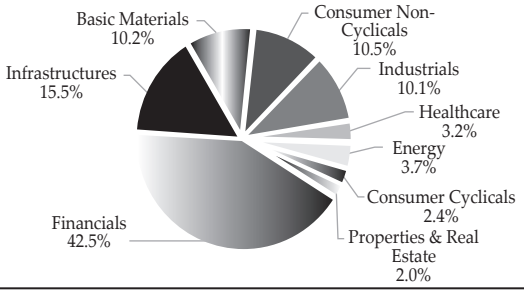
Figure 3.
JII Stocks Characteristics

The LQ45 index is an index provided by the IDX that tracks the performance of 45 stocks with a large market capitalization, plenty of liquidity, and solid fundamentals. The index also consists of non-shariah stocks. Since the stocks with largest market capitalization are conventional banks, the LQ45 index is dominated by financial sector as seen in Figure 4 below. Among top 10 constituents of LQ45, there are 4 conventional banks: BBRI, BBCA, BMRI, and BBNI. The financial sector contributes about 42.5% to the market capitalization of LQ45, followed by infrastructure sector and consumer non-cyclicals at 15.5% and 10.5% respectively. The market capitalization of LQ45 is about 61.35% of the market capitalization in IDX.

INDEX CHARACTERISTICS

Number of Constituents	45
Market Cap Total (in tr. IDR)	1,439.42
Largest	219.74
Samallest	1.73
Average	31.99
Median	12.46
Index MC/IDX MC (%)	61.35%

SECTOR WEIGHTS



TOP 10 CONSTITUENTS

Code	Company Name	Market Cap* (in trill. IDR)	Index Weight	Sector
BBRI	Bank Rakyat Indonesia (Persero) Tbk.	219.74	15.27%	Financials
BBCA	Bank Central Asia Tbk.	214.46	14.90%	Financials
TLKM	Telkom Indonesia (Persero) Tbk.	147.42	10.24%	Infrastructures
BMRI	Bank Mandiri (Persero) Tbk.	121.51	8.44%	Financials
ASII	Astra Iternational Tbk.	111.35	7.74%	Industrials
BBNI	Bank Negara Indonesia (Persero) Tbk.	40.99	2.85%	Financials
UNVR	Unilever Indonesia Tbk.	38.10	2.65%	Consumer Non-Cyclicals
CPIN	Charoen Pokphand Indonesia Tbk.	37.83	2.63%	Consumer Non-Cyclicals
UNTR	United Tractors Tbk.	34.46	2.39%	Industrials
SMGR	Semen Indonesia (Persero) Tbk.	30.79	2.14%	Basic Materials
	Total (Top 10 Constituents)	**996.63**	**69.24%**	

Data prior to the launch date is back-tested data. * Adjusted Market Capitalization Data as of: Jan 29, 2021
Source: idx.co.id

Figure 4.
LQ45 Index Characteristics

2.2. Markov Decision Process

Stock trading can be modelled as a Markov Decision Process (MDP) problem because of its stochastic and interactive nature. MDP's goal is to teach an agent how to develop a policy that will yield the most cumulative rewards from a sequence of actions in one or more states. Figure 5 below describes the MDP model.

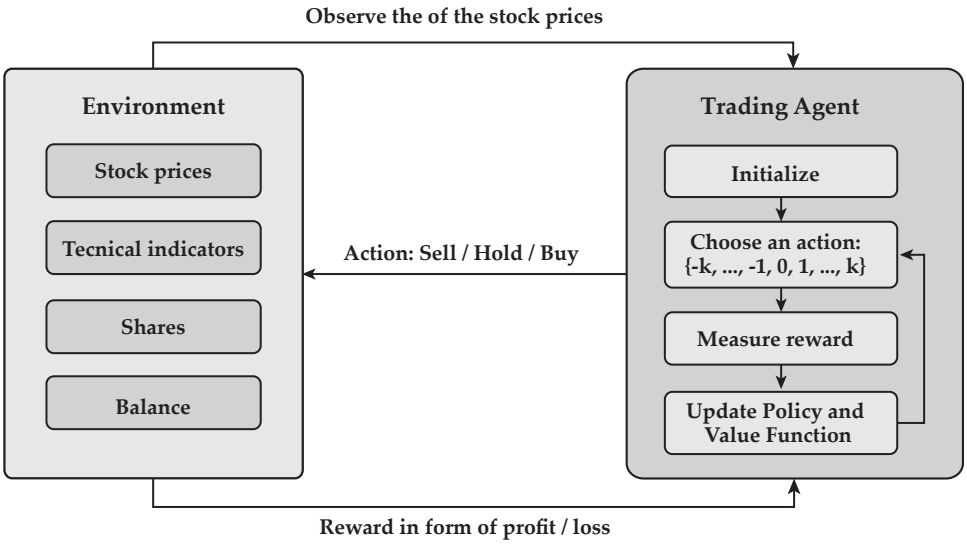


Figure 5.
Conceptual Framework of Stock Trading as a Markov Decision Process

The MDP model is a tuple of ($\boldsymbol{s}$, $\boldsymbol{a}$, r, π, γ) where:

- State $\boldsymbol{s} = [\boldsymbol{p}, \boldsymbol{h}, b]$ is a vector consisting of stock prices $\boldsymbol{p} \in \mathbb{R}_+^D$, stock shares $\boldsymbol{h} \in \mathbb{Z}_+^D$, and the remaining balance $b \in \mathbb{R}_+$, where D denotes the number of stocks and $\mathbb{Z}_+$ denotes non-negative integers.
- Action a is a vector of action over D stocks. The allowed actions on each stock include *selling*, *buying*, or *holding*, which result in decreasing, increasing, and no change of the stock shares h, respectively.
- Reward $r(s, a, s')$ is the direct reward from action a at state s and arriving at the new state s'.
- Policy $\pi(s)$ is the trading strategy at state s, which is the probability distribution of actions at state s.
- The future rewards are discounted by a factor of $0 < \gamma < 1$ for convergence purpose.

In order to simulate the trading of multiple stocks, a continuous action space is employed. The portfolio is presumed to consist of 30 stocks in total. The state space of numerous stocks trading environments is represented as a 181-dimensional vector with seven components of information at time t: $[b_t, \boldsymbol{p}_t, \boldsymbol{h}_t, \boldsymbol{M}_t, \boldsymbol{R}_t, \boldsymbol{C}_t, \boldsymbol{X}_t,]$, where:

- $b_t \in \mathbb{R}_+$ is the remaining balance of portfolio.
- $\boldsymbol{p}_t \in \mathbb{R}_+^{30}$ is the adjusted close price of each stock.
- $\boldsymbol{h}_t \in \mathbb{Z}_+^{30}$ is the number of shares of each stock.
- $\boldsymbol{M}_t \in \mathbb{R}_+^{30}$ is Moving Average Convergence Divergence (MACD) that is calculated using the close price. MACD uses momentum indicator to identifies moving averages.

- $\boldsymbol{R}_t \in \mathbb{R}_+^{30}$ is Relative Strength Index (RSI) based on the close price. RSI measures how much recent price has changed. If the stock's price fluctuates around the support line, it is oversold, and one should purchase it. If the stock's price swings around the resistance level, it is overbought, and best to sell it.
- $\boldsymbol{C}_t \in \mathbb{R}^{30}$ is the Commodity Channel Index (CCI) based on high, low, and close prices. It compares the current price to the average price over a period of time to determine whether to purchase or sell.
- $\boldsymbol{X}_t \in \mathbb{R}_+^{30}$ is the Average Directional Index (ADX) calculated using high, low and close price. ADX indicator determines trend strength by calculating the amount of price change.

For all stock $d \in [1, D]$ in the portfolio, an action is taken at each state. The three possible actions are as follows:

- Selling $\boldsymbol{k}[d] \in [1, \boldsymbol{h}[d]]$ shares result in $\boldsymbol{h}_{t+1}[d] = \boldsymbol{h}_t[d] - \boldsymbol{k}[d]$, where $\boldsymbol{k}[d] \in \mathbb{Z}_+$.
- Holding $\boldsymbol{h}_{t+1}[d] = \boldsymbol{h}_t[d]$.
- Buying $\boldsymbol{k}[d]$ shares result in $\boldsymbol{h}_{t+1}[d] = \boldsymbol{h}_t[d] + \boldsymbol{k}[d]$.

The action space for a single stock is described as {-*k*, ..., 1, 0, 1, ..., *k*} where *k* and -*k* represent the number of shares to be bought and sold, respectively, and $k \leq h_{max}$ represents the maximum number of shares for any buying action. As a result, the total size of the action space is $(2k + 1)^{30}$.

It is important to note that the portfolio value at time *t* is $b_t + \boldsymbol{p}_t^T \boldsymbol{h}_t$. After some actions at time *t* are taken and the stock prices are updated at $t + 1$, the portfolio value may change to $b_{t+1} + \boldsymbol{p}_{t+1}^T \boldsymbol{h}_{t+1} - c_t$ after deducting transaction cost c_t. The MDP's goal is to maximizes the portfolio's final value. As a result, the reward function is defined as the change in portfolio value when action α_t is taken in state s_t and the new state s_{t+1} is reached.

$$r(s_t, a_t, s_{t+1}) = (b_{t+1} + \boldsymbol{p}_{t+1}^T \boldsymbol{h}_{t+1}) - (b_t + \boldsymbol{p}_t^T \boldsymbol{h}_t) - c_t, \tag{1}$$

Based on the action at time *t*, the stocks are divided into sets for selling *S*, buying *B*, and holding *H*, where $S \cup B \cup H = \{1, \ldots, D\}$ and nothing intersects. One can substitute b_{t+1} and h_{t+1} with the value in time *t* as follows:

$$\boldsymbol{h}_{t+1} = \boldsymbol{h}_t^H - \boldsymbol{k}_t^S + \boldsymbol{k}_t^B, \tag{2}$$

$$b_{t+1} = b_t + (\boldsymbol{p}_t^S)^T \boldsymbol{k}_t^S - (\boldsymbol{p}_t^B)^T \boldsymbol{k}_t^B, \tag{3}$$

Suppose r_H, r_S, and r_B denote the change of the portfolio value comes from holding, selling, and buying shares from time *t* to $t + 1$, respectively. Then the reward function can be rewritten as

$$r(s_t, a_t, s_{t+1}) = r_H - r_S + r_B - c_t, \tag{4}$$

where

$$r_H = (\boldsymbol{p}_{t+1}^H - \boldsymbol{p}_t^H)^T \boldsymbol{h}_t^H, \tag{5}$$

$$r_S = \left(\boldsymbol{p}_{t+1}^S - \boldsymbol{p}_t^S\right)^T \boldsymbol{k}_t^S, \tag{6}$$

$$r_B = (\boldsymbol{p}_{t+1}^B - \boldsymbol{p}_t^B)^T \boldsymbol{k}_t^B, \tag{7}$$

The stock values at time 0 are set to p_0, while the beginning fund amount is set to b_0. Suppose Q-value $Q_\pi(s,a)$ is the expected reward from action a at state s following policy π. The values of h and $Q_\pi(s,a)$ are both zero, and $\pi(s)$ is evenly distributed among all actions for each state. Then, by engaging with the stock market environment, $Q_\pi(s_t,a_t)$ is updated.

The Bellman Equation determines the best strategy, with the expected benefit from taking an action in state s_t equaling the total of the direct reward $r(s_t, a_t, s_{t+1})$ and the future reward in the following state s_{t-1}. If the future benefits are discounted by a factor of $0 < \gamma < 1$ for the sake of convergence, the Q-value function can be written as

$$Q_\pi(s_t,a_t) = \mathbb{E}_{s_{t+1}}\left[r(s_t,a_t,s_{t+1}) + \gamma \mathbb{E}_{a_{t+1}\sim\pi(s_{t+1})}[Q_\pi(s_{t+1},s_{t+1})]\right] \quad (8)$$

The Deep Reinforcement Learning (DRL) approach is used to address the challenge of designing a trading strategy that maximizes the positive capital gain $r(s_t, a_t, s_{t+1})$ in a dynamic environment.

2.3. Actor-critic Learning Method

Based on Fischer (2018), the implementations of DRL in financial markets uses one of three learning approaches: critic-only method, actor-only approach, or actor-critic method. In this section, the three approaches are reviewed.

The most frequent strategy, critic-only learning, handles a discrete action space issue by training an agent on a single stock or asset using Deep Q-learning (DQN) and its refinements (Chen & Gao, 2019). The critic-only technique aims to learn the best action-selection strategy that maximizes the predicted future reward given the current state of using a Q-value function. Instead of creating a state-action value table, DQN minimizes the difference between estimated and target Q-values over a transition and estimates the function with a neural network. However, the critic-only technique can only function with discrete and finite state and action spaces, and it is inconvenient for a big portfolio of stocks because prices are continuous.

Moody & Saffell (2001), Deng et al. (2017), and Jiang & Liang (2017) have applied the actor-only agent. The actor-only agent learns the best policy on its own. Rather than learning the Q-value, the policy is learned by the neural network. The policy is a probability distribution that is basically a strategy for a given situation, specifically the probability of taking a permitted action. Moody & Saffell (2001) introduce recurrent reinforcement learning to escape the dimensional curse and boost trade efficiency. Continuous action space settings can be handled with an actor-only method.

According on Bekiros (2010), and Xiong et al. (2018), the actor-critic method has recently been used in finance to update both the policy actor network and the value function critic network at the same time. The value function is estimated by the critic, and the actor updates the policy probability distribution using policy gradients, which are led by the critic. The actor improves his acts through time, while the critic improves his ability to evaluate those acts. The actor-critic method has been utilized to play famous video games like Doom (Wu & Tian, 2017), and has shown to be capable of learning and adapting to big and complicated world. As a result, the actor-critic method holds promise in multiple-stock portfolio trading.

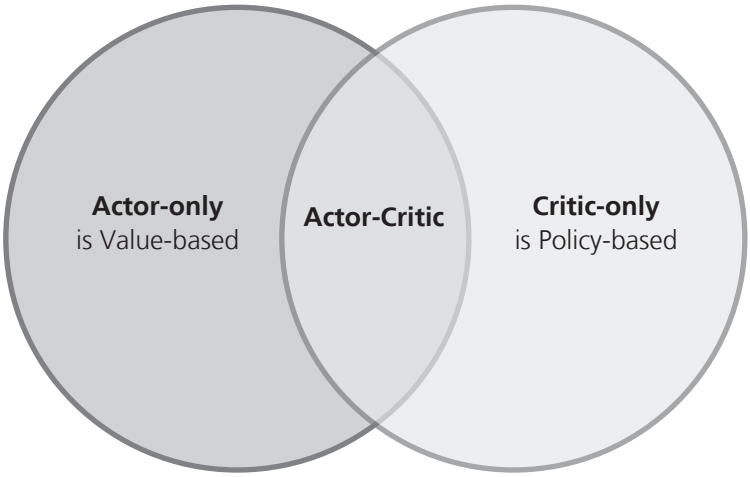


Figure 6.
The Actor-Critic Method Updates Both the Policy and the Value Function

Implementing a DRL trading strategy is quite a challenging process. The development and debugging procedures are time-consuming and prone to mistakes. To solve this issue, Liu et al. (2020) develop a three-layered FinRL library that streamlines the development of stock trading strategies. FinRL provides common building blocks that enable strategy developers to create virtual stock market environments, train deep neural networks as trading agents, measure trading performance through comprehensive back testing, and add key market frictions. In this study, FinRL library is utilized to implement the DRL framework.

III. METHODOLOGY

3.1. Assumptions and Constraints

Transaction costs, market liquidity, risk aversion, and other concerns are addressed under the following assumptions and limitations:

- The orders can be placed at the close price and the stock market is not affected by the trading agent.
- The actions taken should not result in a non-negative balance.
- The transaction costs charged by brokers may vary. The selected transaction costs are 0.19% of each buy value and 0.29% of each sell value.
- To manage risk in a worst-case situation like the global financial crisis of 2008, the financial turbulence index *turbulanc* e_t is used, which monitors excessive asset price volatility.

$$turbulance_t = (\boldsymbol{y_t} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_t} - \boldsymbol{\mu})' \quad (9)$$

where $\boldsymbol{y}_t \in \mathbb{R}^D$ denotes the stock returns for current period t, $\boldsymbol{\mu} \in \mathbb{R}^D$ denotes the average historical returns, and $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{D \times D}$ denotes the covariance of historical returns. The trading agent simply stops buying or selling all shares when *turbulanc*

e_t exceeds a threshold, indicating severe market situations. Once the turbulence index falls below the threshold, it continues trading.

3.2. Trading Agent Algorithms

The trading agents are based on three actor-critic based algorithms: A2C, DDPG, and PPO. Advantage Actor Critic (A2C) is an actor-critic algorithm that is used to reduce the variance of the policy gradient by using an advantage function. Evaluation of an action considers not just how effective it is, but also how much better it could be. As a result, the high variance of the policy networks is reduced, and the model becomes more robust.

Deep Deterministic Policy Gradient (DDPG) is an actor critic-based algorithm that is used to optimize investment return. Q-learning and policy gradient frameworks are combined in the DDPG, which employs neural networks as feature approximators.

Proximal Policy Optimization (PPO) is used to keep track of the policy gradient update and guarantee that the new policy does not deviate too far from the prior one. Large policy changes outside of the clipped period are discouraged by PPO. As a result, by limiting policy updates at each training phase, PPO increases the stability of policy network training. PPO is utilized for stock trading because it is reliable, fast, and simple to set up.

Before starting the agent training, some parameters should be determined. The first parameter is the date range for training and trading. For this study, the data for training are from January 2010 to December 2018. The trading is from January 2019 to December 2020.

Then, for the model environment, the initial amount of balance is set to IDR 150 million, which is close to USD 10,000 value that is commonly used as an initial financial investment. Considering the initial amount, the number of shares for each stock is limited to a maximum of 100 lot. The transaction cost is set based on IPOT (PT Indopremier Sekuritas broker) fees, which is 0.19% of buy value and 0.29% of sell value. The turbulence index threshold is set to 100 that is 95th percentile approximation of the turbulence index for the last two years.

The parameters set for the algorithms are the following:

1. batch_size: minibatch size for each gradient update
2. buffer_size: size of the replay buffer
3. ent_coef: entropy coefficient for the loss calculation
4. learning_rate: learning rate for Adam optimizer. The same learning rate will be used for all networks (Q-Values, Actor and Value function)

In A2C algorithm, the ent_coef is set to 0.01, and learning rate is the same as the default of 0.0007. In DDPG algorithm, the batch_size is set to 128, buffer_size is 50000, and learning_rate is the same as default of 0.001. In PPO algorithm, the batch_size is set to 128, ent_coef is 0.01, and the learning_rate is 0.00025.

3.3. Performance Metrices

The model performance is assessed using five metrics:

- Cumulative return is computed by subtracting the portfolio's final value from its initial value, then dividing by the latter.
- Annualized return is the geometric average of the amount of money made by the agent each year throughout the period.
- Annualized volatility is the standard deviation of a portfolio's performance over a year.
- Sharpe ratio is determined by subtracting the annualized risk-free rate from the annualized return and dividing by the annualized volatility.
- Max drawdown is the maximum percentage loss throughout the trading session.

Returns at the end of the trading stage are reflected in the cumulative return. The portfolio's annualized return is the return at the end of each year. The robustness of the model is measured by annualized volatility and maximum drawdown. The Sharpe ratio is a common measure of both return and risk.

3.4. Research Design

The initial step was to review the literature related to DRL in quantitative finance. Most of the studies use U.S. stocks as the object, but in this study, Islamic stocks that are constituents of the JII index are used and compared to conventional stocks in the LQ45 index. The daily close prices and trading volume for each stock from 2010 until 2020 are collected programmatically using Yahoo Finance API. Then, technical indicators such as MACD, RSI, CCI, ADX, and turbulence index are added as additional data. A virtual environment is developed to simulate the stock trading based on the stock prices, technical indicators, remaining balance, and number of shares for each day.

After the data and environment are ready, three algorithms of DRL or agents are trained using data from 2010 until 2018. The output of training is stock trading strategy by each agent. Then, to test the agent's performance, the data from the period of 2019 until end of 2020 are chosen for the trading phase. The performance is compared between the algorithms, and against LQ45 index as the benchmark.

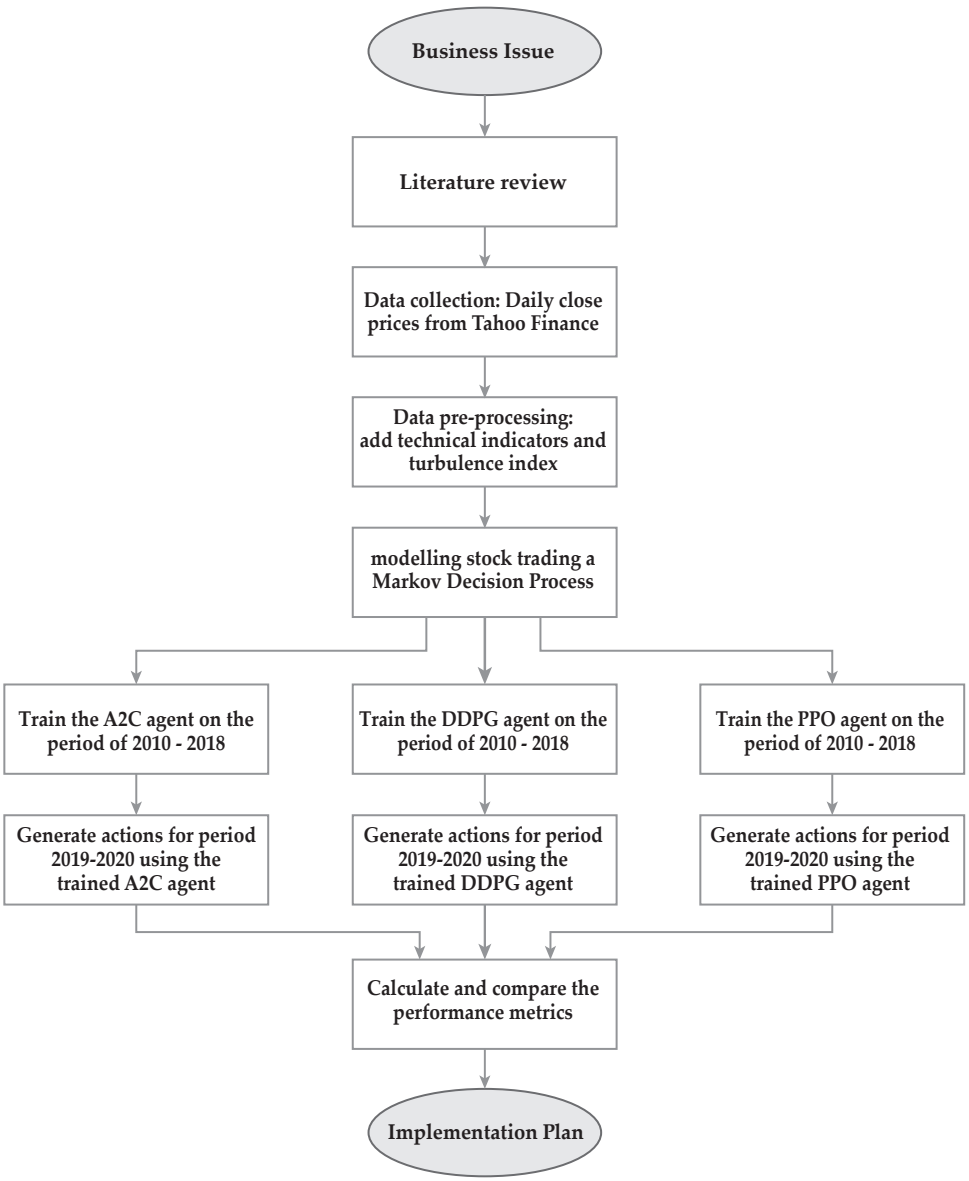


Figure 7.
Research Methodology

IV. RESULTS AND ANALYSIS

4.1. Performance Comparison for the Period 2019-2020

The performance evaluation of the suggested strategy is presented in this section. Table 1 shows that all agents' returns outperform the benchmark index LQ45. A2C has the best max drawdown of -38.54% with relatively high Sharpe ratio of 0.76. DDPG generates the lowest cumulative return, 4.00%, and annual return of 2.05%. While DDPG has the lowest annual volatility, which is 26.89%, it still gives better return than the benchmark index LQ45. PPO agent has the greatest annual return of 23.36% and cumulative return of 50.03%. However, PPO has the worst max drawdown, -44.84%, and the highest annual volatility among the three agents.

Table 1.
Performance Metrics Comparison from Jan 2019 to Dec 2020

Metrics	A2C	DDPG	PPO	LQ45	JII
Cumulative returns	41.70%	4.00%	50.03%	-5.02%	-7.96%
Annual return	19.77%	2.05%	23.36%	-2.63%	-4.20%
Annual volatility	29.02%	26.89%	36.55%	27.92%	26.39%
Sharpe ratio	0.76	0.21	0.76	0.04	-0.03
Max drawdown	-38.54%	-40.99%	-44.84%	-45.59%	-45.82%

As seen in the Figure 8 below, DDPG is the best agent in sideways market at the first quarter of 2019. When the trend is improving, the PPO and A2C agents performs better. In the market crash due to lockdown in March 2020, since the turbulence index is higher than 100, the algorithms sell all the stocks in hand and stop buying until the turbulence index returns to normal. As a result, the PPO and A2C agents could avoid the crisis quickly. Then, they respond to maximize profit as the market bounces back. In the end, A2C and PPO give higher return than the benchmark index LQ45 and JII.

The previous research by Yang et al. (2020) also uses A2C, DDPG, and PPO algorithms on 30 Dow Jones (DJI) constituent stocks. The performance comparison shows that PPO is able to generate the best cumulative return, A2C is lower but still higher than DDPG. Therefore, the order of the best agents on DJI stocks is somewhat similar to the order of the best agents on JII stocks in this study. According to Yang et al. (2020), PPO agent is effective at anticipating trends and maximizing profits, while A2C agent is more risk tolerant, and in a bullish market, DDPG can be utilized in conjunction with PPO.

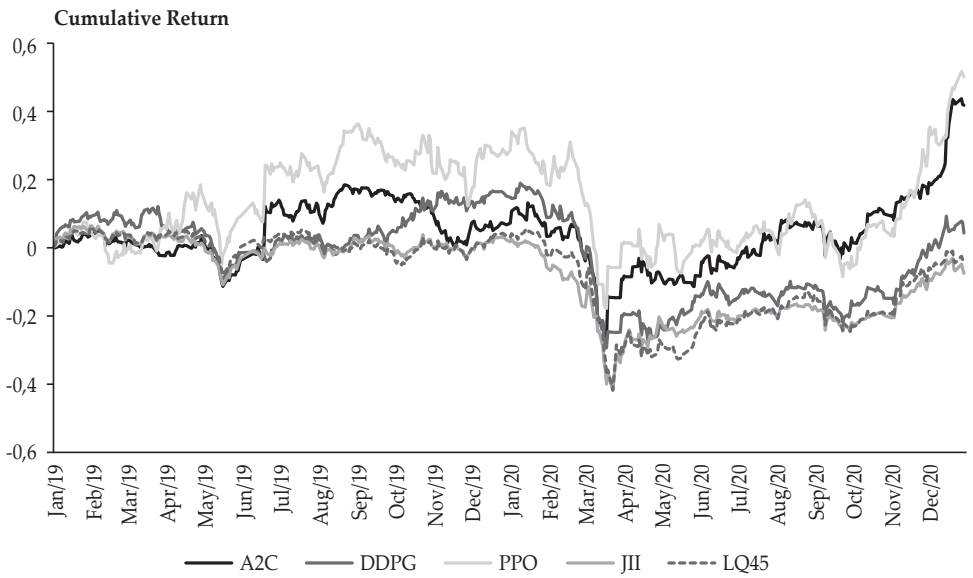


Figure 8.
Cumulative Return Comparison with Initial Portfolio Value of $10,000 from 2019 to 2020

4.2. Sensitivity Analysis

In this section, the agent algorithms are given different input to evaluate the sensitivity. One of the parameters is learning rate. The agents are given lower and higher learning rates than the previously used. As seen in the tables below, the result shows that the previous learning rate (in the middle) still generates the best performance than others for every algorithm.

Table 2.
Sensitivity Analysis of A2C Based on Learning Rate

		A2C	
Learning rate	0.0005	**0.0007**	0.001
Cumulative returns	25.17%	**41.70%**	-39.15%
Annual return	12.32%	**19.77%**	-22.67%
Annual volatility	32.43%	**29.02%**	31.82%
Sharpe ratio	0.52	**0.76**	-0.65
Max drawdown	-47.70%	**-38.54%**	-66.02%

Table 3.
Sensitivity Analysis of DDPG Based on Learning Rate

		DDPG	
Learning rate	0.0007	**0.001**	0.0025
Cumulative returns	-11.44%	**4.00%**	0.07%
Annual return	-6.09%	**2.05%**	0.04%
Annual volatility	34.81%	**26.89%**	33.54%
Sharpe ratio	-0.01	**0.21**	0.17
Max drawdown	-58.11%	**-40.99%**	-51.07%

Table 4.
Sensitivity Analysis of PPO Based on Learning Rate

		PPO	
Learning rate	0.0001	**0.00025**	0.0005
Cumulative returns	15.15%	**50.03%**	-12.56%
Annual return	7.57%	**23.36%**	-6.71%
Annual volatility	33.14%	**36.55%**	34.69%
Sharpe ratio	0.39	**0.76**	-0.03
Max drawdown	-44.80%	**-44.84%**	-62.37%

The agents' short-term performance in the full month of May 2020 is also evaluated. As seen in the Table 5, the PPO and A2C agents generate cumulative returns of 2.45% and 1.42% respectively. They are better than the benchmark index LQ45 and JII, which give 0.49% and -1.79%. On the other hand, DDPG agent has lower cumulative return at -2.91%.

Table 5.
Performance Metrics Comparison in the Month of May 2020

Metrics	A2C	DDPG	PPO	LQ45	JII
Cumulative returns	1.42%	-2.91%	2.45%	0.49%	-1.79%
Sharpe ratio	1.25	-2.18	1.44	0.55	-1.24
Max drawdown	-3.78%	-7.38%	-6.40%	-4.00%	-6.14%

4.3. Risk Mitigation Using Turbulence Index

Turbulence index is used to monitor excessive asset price volatility which indicates severe market situations if it exceeds the threshold. As seen in the Figure 9 below, the turbulence index exceeds the threshold of 100 several times. On the first day of the event, the agent stops buying and sells all the shares until the turbulence index is back to under 100.

Most events are related to COVID-19 since the trading period is between 2019 and 2020. One can observe that the highest turbulence index of 559.29 coincides with the state of emergency declared in DKI Jakarta and East Java on March 20th of 2020. This situation surely results in negative sentiment in the market. However, since the trading agent anticipates this and sells all the shares, a bigger loss has been prevented at that time.

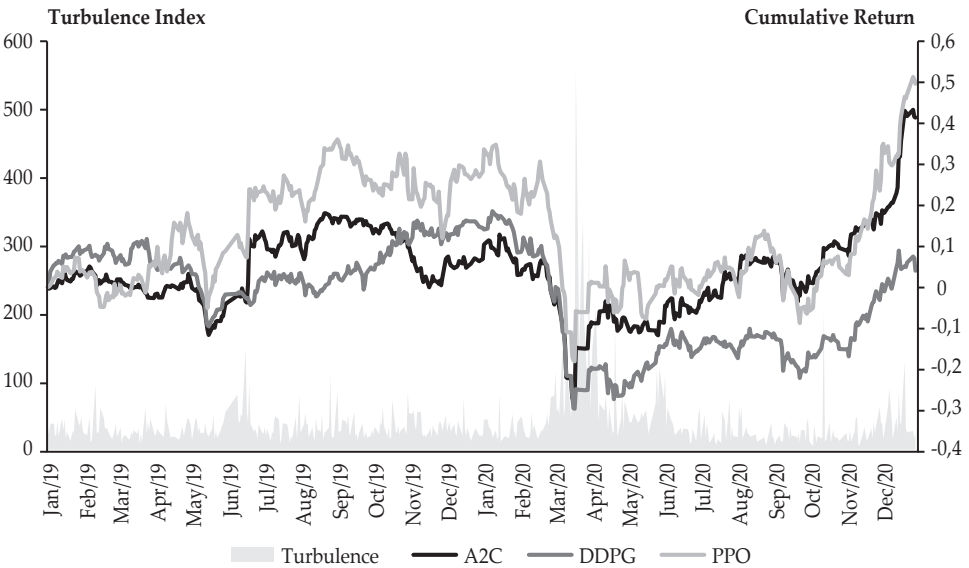


Figure 9.
Turbulence Index Impact to Prevent Loss During Crisis

V. CONCLUSION AND RECOMMENDATION

5.1. Conclusion

To learn a stock trading strategy, one can utilize actor critic-based algorithms such as Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and Deep Deterministic Policy Gradient (DDPG) agents. The PPO agent is the best at observing trends and responding to maximize profits. For dealing with a bullish market, PPO is preferable. DDPG agent has the lowest annual volatility so it is the most risk averse. A2C agent has the best max drawdown so it is more adaptive to risk while generating high return. Based on the cumulative return during 2019-2020, PPO gives the highest yield of around 50.03%, followed by A2C whose yield is 41.70% and finally DDPG generates yield of 4.00%.

By balancing risk and return together with transaction costs, each trading agent can improve the performance of 30 Islamic stocks in Indonesia and outperform the index of 45 most liquid conventional stocks in the Indonesian Stock Exchange (IDX), except DDPG in the short-term test. This suggests that Islamic stocks in the emerging market of Indonesia can also generate better results when a good trading strategy is applied.

5.2. Recommendations

For fund managers, the actor-critic DRL algorithms can be utilize to optimize the Islamic stocks portfolio on a daily basis, minimize risk during crisis or turbulence periods, and outperform the conventional stocks. To implement the technique, the fund managers should first install a code editor and the requirements of the program. Secondly, they should choose some Islamic stocks that constitute the portfolio. However, not all stocks will be traded unless the agent recommends buying the stocks. Next, the fund managers set up the program and choose a training data, for example, data of the chosen stocks from the last 10 years. Then, the DRL agents will be trained using those data and generate stock trading strategy model. The fund managers should evaluate the performance of the data from the last 2 years and change some parameters if it is still unsatisfactory. Furthermore, the fund managers choose a date range for the actual trading, set the date range to the program, and load the saved model to generate actions for stock trading. Finally, the fund managers can perform the actual trading based on the recommended actions by the agents for each stock

Policy makers and regulators may consider discussing and making the regulation related to automated trading using machine learning such as the DRL. Currently, machine learning has gained its popularity as tools for investment and trading, widely known as robotrading. However, there are many robotrading providers that harm the society. The total loss from illegal robotrading activities is estimated to reach more than IDR 2.5 trillion in 2021 (Malik, 2022). Therefore, regulators should have the standard to evaluate the methods used by automated trading providers.

In future research, it will be interesting to explore more advanced models with large-scale data using all Islamic stocks available. We can also add more variables to the state space, such as fundamental indicators, the sentiment of financial market news, and the macro economy parameters.

REFERENCES

Adler, T., & Krizman, M. (2008). The cost of socially responsible investing. *Journal of Portfolio Management, 35*(1), 52–56.

Barnett, M., & Salomon, R. (2006). Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance. *Strategic Management Journal, 27*(11), 1101–1122.

Bekiros, S. D. (2010). Heterogeneous trading strategies with adaptive fuzzy actor-critic reinforcement learning: A behavioral approach. *Journal of Economic Dynamics and Control, 34*(6), 1153–1170.

Chen, L., & Gao, Q. (2019). Application of deep reinforcement learning on automated stock trading. *In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. (pp. 29–33).

Deng Y., Bao F., Kong, Y., Ren, Z., & Dai Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653–664.

Faturohman, T., Widjaya, K. A., & Afgani, K. F. (2021). Sin stock proportion and investment manager education background in indonesian equity funds. In Barnett, W.A. & Sergi, B.S. (Eds.), Environmental, social, and governance perspectives on economic development in Asia (*International Symposia in Economic Theory and Econometrics*, *Vol. 29A*), Emerald Publishing Limited, Bingley, pp. 83-99. https://doi.org/10.1108/S1571-03862021000029A020.

Fischer, T. G. (2018). Reinforcement learning in financial markets - a survey. FAU Discussion Papers in Economics 12/2018, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics.

Jiang, Z., & Liang, J. (2017). Cryptocurrency portfolio management with deep reinforcement learning. *In 2017 Intelligent Systems Conference*. London, UK: IEEE.

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. *In Deep Reinforcement Learning Workshop, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. arXiv preprint arXiv:2011.09607*. (pp. 1-12).

Malik, A. (2022). "Bikin Ngilu! Kerugian Masyarakat Akibat Kripto Dan Robot Trading Ilegal Capai Rp6,5 Triliun." ["Make Painful! Community Losses Due to Crypto and Illegal Trading Robots Reached Rp6.5 Trillion"]. *Bareksa.Com*.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, *7*(1), 77–91.

Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks, 12*(4), 875–889.

Wu, Y., & Tian, Y. (2017). Training agent for first-person shooter game with actor-critic curriculum learning. *ICLR 2017 Conference.*

Xiong, Z., Liu, X. Y., Zhong, S., Yang, H., & Walid, A. (2018). Practical deep reinforcement learning approach for stock trading. *NIPS Workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy, Montréal, Canada.*

Yang, H., Liu, X. Y., Zhong, S., & Walid, A. (2020). Deep reinforcement learning for automated stock trading: An ensemble strategy. *ICAIF '20, October 15–16, 2020, New York, NY, USA.* Available at SSRN.