School of Computing and Engineering

MSc Dissertation

CP70017E

Reinforcement Learning in Crypto Trading

Supervisor Name: *Professor Jonathan Loo*

Student ID: *21361642*

Student Name: *Nikolay Nikolaev Ninov*

Table of Contents

Chapter 1 - Introduction

People have always had to trade commodities to acquire goods that they could not produce themselves. Before people would trade food, shelter, and clothes. However, trading with commodities were difficult to store and transport. This led to people inventing money to make trades easier and faster.

Nowadays money is becoming devaluated, and inflation has increased rapidly since COVID-19. Since lockdowns have been introduced and the world economy rapidly stopping, taxes and inflation will yet further rise. About 20 per cent of all US dollars were printed in 2020 [1]. Until the 1970s money was directly linked to gold (gold standard) and afterwards it was replaced with fiat money, government-issued currency which is not backed by physical commodity [2]. On the other hand, crypto currencies have become a deflationary asset which cannot be manipulated the same way interest rates and increasing money printing are. They are also alternative currencies for countries with hyperinflation - rapid increase in prices while wages stagnate, currency purchase power decreases and living costs become more expensive [3]. Although crypto currencies have the potential to stabilize the economy, governments and banks are concerned because it is a decentralized system which records transactions on the blockchain [4].

In recent years with the swift development of computer hardware, deep learning and reinforcement learning has become faster and has gain popularity within industries such as finance and healthcare. Reinforcement learning has also been developing rapidly withing stock market and forex exchange market. Due to that it has also started developing in the crypto market as well.

The aim of this dissertation is to investigate the current state of reinforcement learning in crypto trading and develop a custom agent environment with a commonly used algorithm.

1.1 Reinforcement learning background

Reinforcement learning is learning what to do by mapping situations to actions to maximize a numerical reward. Rather than being told which actions to take, the agent must discover which actions give the most reward by trying them. Actions can have an immediate impact on the reward, next state of the environment and as well to all subsequent rewards.

The learner and decision maker are known as an agent. Everything that is outside of the agent, and it interacts with is called an environment. The agent interacts with the environment by choosing an action and the environment responds to the actions with new situation and special numerical numbers, known as rewards.

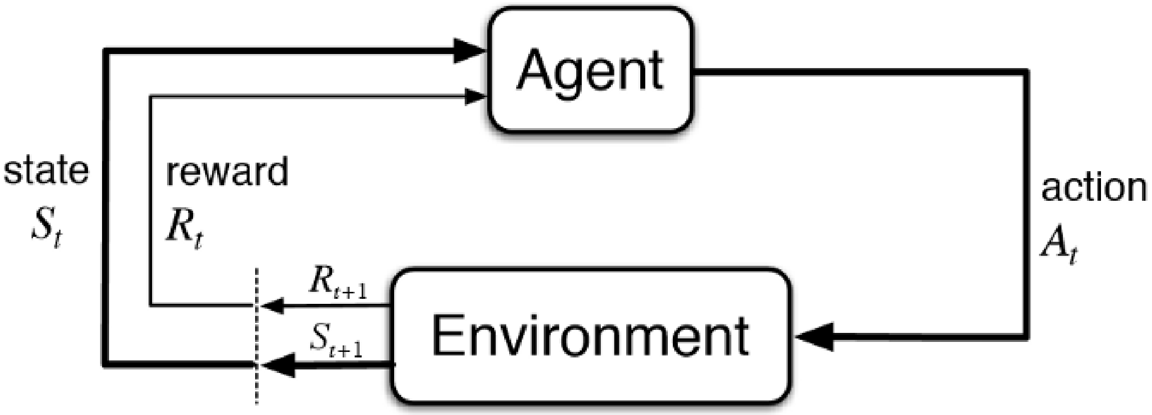


Figure 1.1: Action-reward feedback loop

One of reinforcement learning's important challenges is the trade-off between exploration and exploitation. The agent must exploit what it already has experienced but also explore to make better future decisions [15].

Even though reinforcement learning has been around since the 1960s, neural networks have started becoming popular and better than previously used techniques for approximation. Combining neural networks with reinforcement learning is known as deep reinforcement learning (DRL). DRL has achieved superhuman performances on Atari games and is even competing against humans in real games such as Go and Dota 2.

1.2 Motivation

Humans are emotional and it can seriously impact whilst trading. Personal assets that are traded can be quickly gained and lost which affects our emotion. Emotional trading is when emotions have made an impact on the person's decision-making. This practice can lead to major financial losses and is rarely helpful. On the contrary, machines are emotionless, are not overwhelmed and can process more information for a very short amount of time compared to people. Unfortunately, traditional machine learning and deep learning focus on supervised and unsupervised learning, whereas trading is a decision-based activity – a trader needs to either buy, sell, or hold an asset.

Reinforcement learning can be used to make decisions rather traders and average people. Even though the crypto currency market is the newest one, it is very volatile and can lead to either huge profits or losses for a short period of time. [16] suggests that emotions can be predicted by looking at the trading volume and return volatility. Emotions affect the total return variation process of investors and could influence the financial market by rapid price movements.

The current state of reinforcement learning for crypto currency trading is not as developed as forex exchange and stock market trading but has plenty of potential to continue expanding due to crypto's increasing security and innovative technology.

1.3 Research questions

The primary focus of this research is to improve reinforcement learning for crypto currency trading with a detailed approach in developing a custom crypto trading environment which takes different aspects of trading when receiving a reward. The main research question is formulated as follows:

What technical indicators could be applied to an agent's environment whilst trading crypto currencies?

This issue can be broken down by answering the following secondary questions:

1. *What are the most used trading strategies by traders and normal people?*
2. *What information is used when designing a trading environment for agents?*
3. *What are the current DRL designs for trading?*

1.4 Aims and objectives

The aim of this dissertation is to develop a trading environment which takes into consideration crypto related indicators so an agent can learn how to trade. This implies researching and testing algorithms that are used nowadays. The following objectives can be broken down in the following way:

1. Analyse different trading strategies which will allow focus on designing an agent's environment.
2. Evaluate existing research on reinforcement learning crypto trading.
3. Develop a custom DRL environment.
4. Test and evaluate agents' performance based on the designed environment.

Chapter 2 - Literature Review

This chapter investigates common human trading strategies, different trading techniques used for reinforcement learning trading and how they are used for crypto trading. Even though crypto currency trading is relatively new compared to forex and stock trading, it inherits strategies and trends which the other trading methods use such as day trading, range trading, etc.

2.1 Crypto currency background

2.1.1 Crypto vs Fiat money

Until 20th century, most currencies and coinage were directly converted to gold. The wealthiest countries followed the gold standard, where governments tie fixed exchange rates for national currency to gold. Countries kept sufficient reserves of gold in their vaults to safely back circulating currency supply. Currency for gold exchanges were always possible. However, during the great depression this system began to be abandoned - governments were unable to source more gold to expand their money supply and stimulate spending.

Countries switched to a fiat model - national currency which is not backed by a commodity such as gold. Now central banks can print money whenever needed. Their value is set by changes in supply, demand, and the government's strength. **People tend to buy into fiat currencies with the confidence that it will be accepted elsewhere in exchange for services and goods.** In this case confidence generates purchasing power.

The objective. of crypto currencies is to remove the problems of traditional banking:

- There is no limit to the money you can transfer using some currencies.

- Sending money from one user to another is instant by using a peer-to-peer networking structure which eliminates middlemen and transaction costs.
- Accounts are nearly impossible to hack because no financial institutions are used.
- No central point of failure.

As adoption of crypto currencies grows, the prices become more stable. Another feature that has made them appealing is that they are more resistant to inflation than fiat currencies such as the U.S. dollar. Some crypto currencies have a coin limit whereas others do not. They cannot be manipulated by governments hence the interest rates cannot be adjusted or more money cannot be printed to achieve some policy goals.

2.1.2 Blockchain

In summary blockchain is a distributed database or ledger that is shared among the nodes of a computer network. Blockchain stores information electronically in digital format. They are best known for their important role in crypto currencies. Blockchain collects the data together in groups (blocks) that hold sets of information. Blocks have specific storage capacities and when it is filled, they are closed and linked to the previously filled block. This forms a chain of data which is also called blockchain. The objective. of blockchain is to record and distribute information but not edit it.

When a new transaction is entered, it is transmitted to a network of peer to-peer computers scattered across the world. Then the network solves equations so it can confirm if the transaction is valid or not. When the transaction is confirmed to be legitimate, they are clustered together into blocks. These blocks are chained together which creates a long history of all transactions that are permanent. Finally, the transaction is complete.

2.1.3 Crypto currency usage

The goal of cryptocurrencies is to fix the problem with traditionally used currencies by putting the power and responsibilities in the holders' hands. Compared to banks, cryptocurrencies are completely free of the control of third parties.

2.1.3.1 Owned by everyone

Cryptocurrencies are a digital form of money which makes a more secure exchange. The transactions are public, irreversible, mostly unhackable and controlled by the people due to blockchain. Cryptocurrencies are decentralized - nobody owns or regulates it. Its value is not tied to a country's political views or central bank's monetary policies.

On the other hand, the U.S Federal Reserve plans to release their instant digital payment platform, FedNow, allowing consumers and businesses to instantly pay as well. This removes middlemen such as SWIFT and makes the payment service faster. This is an indirect threat to U.S citizen’s freedom because the Federal Reserve will have full access to their transaction history and people can even get into legal trouble even if no crime has been committed.

Chainalysis reports that crypto currency usage is growing even faster than before. The total transaction volume grew to $15.8 trillion in 2021, which is 567% from 2020’s total when a global pandemic was announced, and people were forced to stay in their homes and isolate [10].

2.1.3.2 Nearly impossible to forge

Cryptocurrencies operate on blockchain and are the key to the power of digital currencies. The "block" is made up of chunks of encrypted data. The "chain" is the public database where the blocks are stored and sequentially related to each other. Blocks in the blockchain have unique codes, called hash. Blocks are added in a chronological order. For example, a new block is added directly after the last block created. The database of blocks is simultaneously distributed worldwide among the thousands and millions of computers. **If a user wants to forge a single block of data on the chain, they will have to manipulate all the blocks from a point in history forward and update all the computers which hold copies of the blockchain.** Even though it is possible to create a new block it will cost a ridiculous amount of power and money to try.

2.1.3.3 Security growth over time and value

When crypto currency initially appeared, it was easier to gain control of majority of their networks when they were much smaller. However, as cryptocurrencies gained more and more attention it became difficult and nearly impossible to hack them due to the enormous amount of power and money required.

2.1.3.4 Crime

Figure 2.1 shows that the total number of illegal addresses are just 0.15% of the total transaction volume for 2021. On the other hand, 2019 has the highest percentage due to the PlusToken Ponzi scheme [11]. Nevertheless, crime decreases with time due to the crypto currency ecosystem.

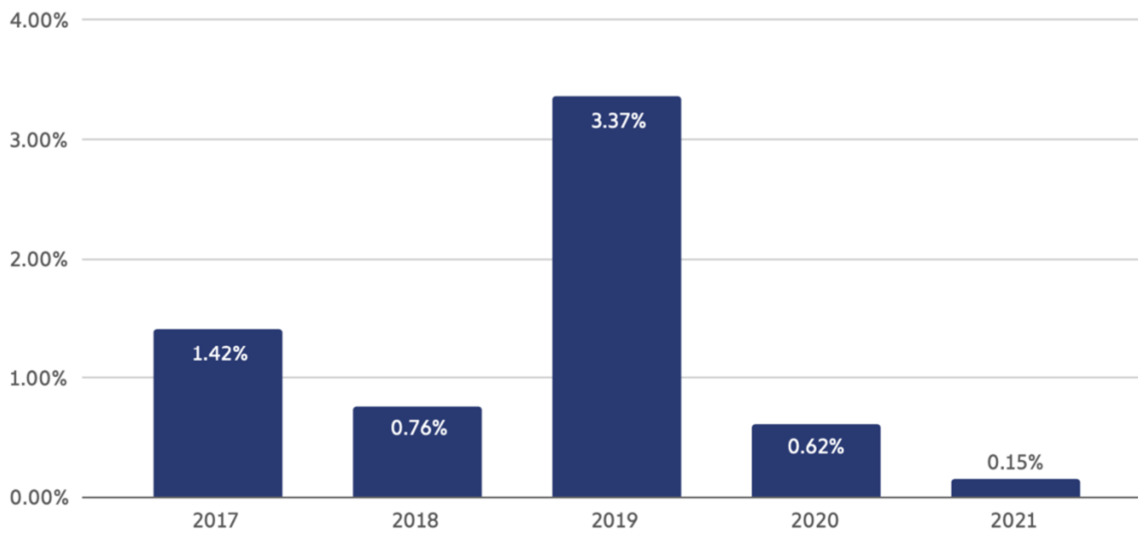


Figure 2.1: Illicit share of all crypto currency transaction volume (2017 - 2021)

The two biggest crime categories which stand out are scams and stolen funds. Scams usually happen when a crypto currency project appears to be legitimate, and enough people have purchased it and then the project's developers take the people's money and disappear. 90% of figure 2.2's value lost in 2021 is due to the fraudulent centralized exchange Thodex. Its CEO disappeared as soon as the exchange halted the users' ability to withdraw their funds.

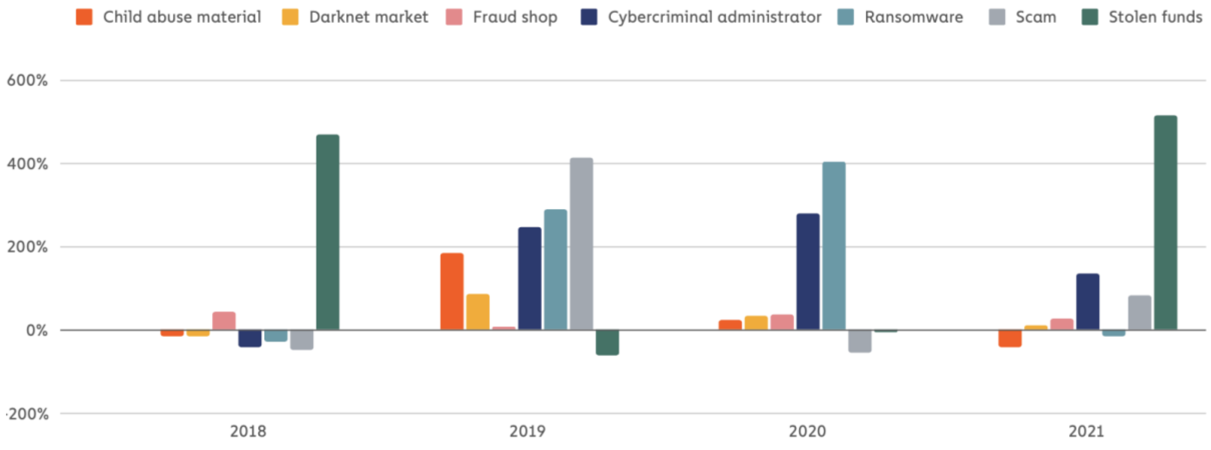


Figure 2.2: Change in value received by crime type (2018 – 2021)

In 2021 law enforcement have been seizing more crypto currencies from criminal. Some of the 2021 examples are:

- The U.S. Department of Justice (DOJ) seized $2.3 million worth of. Crypto currency from DarkSide ransomware operators who were responsible for the attack on Colonial Pipeline [12].
- IRS-CI's. cumulative seized over $3.5. billion worth of. Crypto currency [13].
- London's Metropolitan Police. Service (MPS) made the UK's largest ever seizure of crypto currency - taking £180 million worth from a suspected money launderer [14].

Crypto whales are people or entities that hold the largest amount of crypto currency in one wallet. From all crypto whales, only 3.7% are criminals which is $1 million worth of crypto currency. Most criminal whales usually receive either a very small or an extraordinarily large. Shares of their balance from illegal addresses.

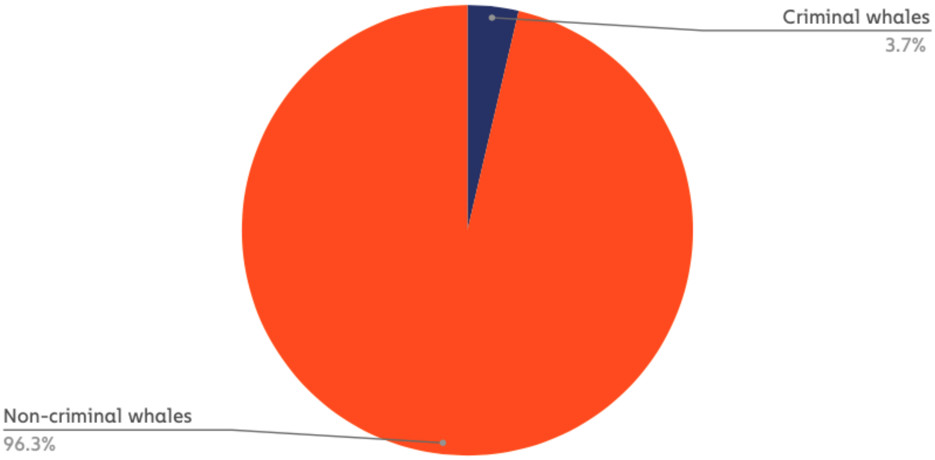


Figure 2.3: The percentage of criminal and non-criminal whales

Figure 2.4 showcases that the biggest criminal activities for whales are dominated by darknet markets, scams and stolen funds which represent 94.4% out of all the crypto currency crime.

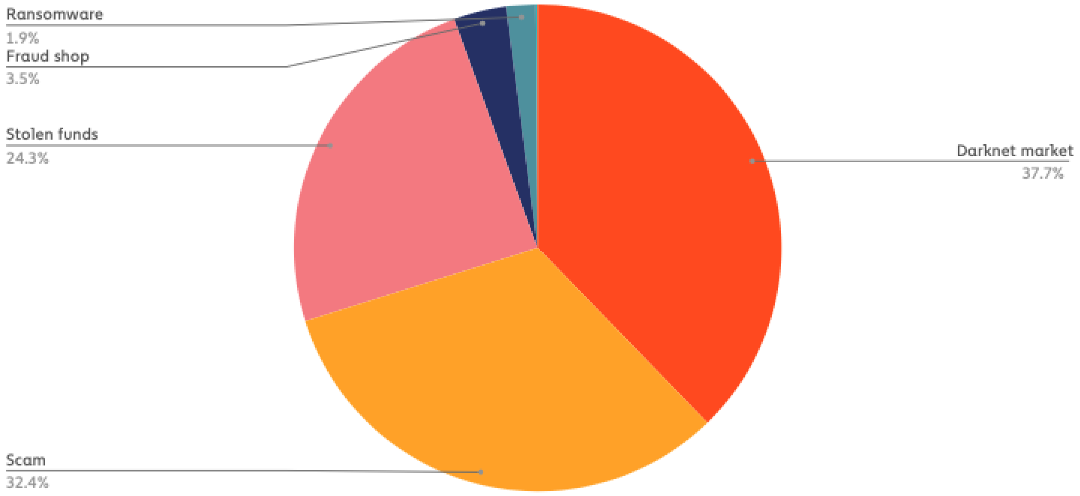


Figure 2.4: Source of illicit funds received by criminal whales

2.1.3.5 Asset backed

Like the U.S. dollar and many other fiat currencies, most crypto currencies are not backed by physical assets in a vault but find their value as a way of paying. However, the coins which are backed by physical assets are called stable coins. Crypto currencies can also be backed by assets.

Gold-backed crypto currency

These are digital currencies whose price is tied to the value of real-world gold. In this case, a trusted third party is used to store the gold. A digital exchange is used to trade the gold. Tokenized gold is much easier to trace than physical gold. What is more, the owner of the gold enjoys the liquidity that tokenization provides. It is easy to redeem the token since it is traded in several digital exchanges.

- **Tether Gold** - one full XAUt token represents one fine troy ounce of gold on a London Good Delivery bar.

- **Paxos Gold** - each token is backed by one fine troy ounce (t oz) of a 400 oz London Good Delivery gold bar, stored in Brink's vaults. If you own PAXG, you own the underlying physical gold, held in custody by Paxos Trust Company.
- **Gold Coin** - a peer-to-peer crypto currency for economic freedom and decentralization. It is backed at a ratio of 1000 GoldCoin per ounce of gold, which makes it less volatile than Bitcoin and other altcoins not pegged to any stable asset.
- **Perth Mint Gold Token** - allows blockchain users to conveniently trade and hold gold stored at The Perth Mint. Digitally manage your entitlements over the physical gold, convert and pick up gold bullion of your choice or get it delivered globally.
- **Meld Gold by Algorand** - streamlines the gold supply chain and makes gold investments accessible for every investor. Meld delivers secure, seamless access to the acquisition of gold, allowing investors to transact in real time, reducing the volatility traditionally associated with the procurement of gold.

Stable Coins

Stable coins are the most popular asset-backed crypto currencies. These coins have the liquidity of fiat currencies and the transparency and decentralization of digital currencies. Should the price of a stable coin take a nose-dive, their users can still get their reserve of fiat currency. Stable coins have proved resistant to price fluctuations. Experts are even hoping to use the concept of stable coins to end the volatility that has always been a problem in the crypto market.

- **Tether** - widely adopted stable coins, having pioneered the concept in the digital token space. They are used by investors who want to avoid the volatility of crypto currencies while holding funds within the crypto system. Tether is backed by a U.S. dollar.

- **Dai** - is the first decentralized, collateral-backed crypto currency. It maintains a stable 1 to 1 value with the U.S. dollar.
- **Binance USD** - compared to the other stable coins, Binance USD has lower fees than many other crypto currencies and is approved by the New York State Department of Financial Services.

The above mentioned are just some examples of stable coins. There are other coins that are also backed by the U.S. dollar, but they work similar to Tether and Dai.

Security Backed

Blockchain start-ups are now tokenizing securities by creating digital coins that represent ownership and generate passive income for the token holders. Tokenized securities are gaining popularity fast because more and more investors and bankers are adopting them. They allow cost-effective, safe, and secure trading. They also bring automation and liquidity to the securities industry.

- **Polymath** - a blockchain start up that tokenizes securities. The platform is based on smart contracts which gives developers a place to launch security tokens. Polymath connects different actors in a developing chain such as Know Your Customer (KYC) developers, smart contract developers and blockchain investors. This platform eliminates middlemen.
- **Gibraltar Blockchain Exchange** - Blockchain firm Valereum is currently acquiring 80 to 90 percent of Gibraltar Stock Exchange (GSX) so a fully regulated, integrated fiat and digital exchange can be built. Since 2019 GSX allows financial firms to list blockchain securities on the GSX market platform. If the Gibraltarian financial regulator approves the deal this would make Gibraltar the first stock exchange that also trades crypto. In addition, this will decrease crypto currencies' taxes.

2.2 Human trading strategies

To understand how reinforcement learning agents are designed and how they should interact with the environment, we need to investigate how humans trade. The biggest challenge in trading is predicting the price movement due to external influence such as news, political events, and market manipulations. They lead how a person can enter and exit trades in the market to increase profit and reduce the exposure of risk.

2.2.1 Random walk

Random walk assumes that stock prices are independent of other factors. The stock's past movements cannot predict the future hence they take a random path. This strategy implies that prices have the same distribution and are independent of each other. Additionally, the technique suggests that it is impossible to outperform the market without the assumption of addition risk. This theory was tested on crypto currencies and their results do not reject the Random Walk theory for Bitcoin, Maker, Dash and Stellar [6]. However, this theory is criticized because prices follow patterns or trends in the long run. In addition, are many other factors that can have an impact on the price. It does not mean that a pattern does not exist because there is no clearly identified pattern.

2.2.2 Contrarian

Contrarian investing is a strategy that goes against the currently existing market trend. This strategy is not used worldwide. For example, the Chinese stock market mainly use this strategy because it mainly has individual investors rather than company investors due to the high risk [8]. This method is mainly used for long terms which ignores the short-term noise and volatility. Nevertheless, if the market does not move for a considerable period the prices will remain low and the sellers are missing out on the profit of increased prices. It also takes a significant amount of effort to study the market, industry, and external factors.

2.2.3 Technical analysis

Technical analysis is used to predict prices by using previous evidence. The most used technical indicators are:

- **Candlestick charts** - displays four different prices for every interval - high, opening, closing and low.
- **Support and Resistance Levels** - used to identify price points between the low-level price (support level) and high-level price (resistance level).
- **Relative Strength Index (RSI)** - a momentum indicator which measures the magnitude of recent price changes to evaluate the price.
- **Average Directional Index (ADX)** - is used to determine how strong a trend is. A trend is either up or down. This is displayed by negative directional indicator (-DI) and positive directional indicator (+DI)
- **Moving Averages (MAs)** - used to identify the trend direction of a price or to determine the support and resistance levels.
- **Trend Lines** - works with support and resistance levels to determine the direction of the trend over a timeframe.

Technical analysis converts complicated interconnections into easy-to-understand history price charts which are reflected instantly. Compared to other methods, this technique does not have a limit of the period it can analyse. The drawback of this method focuses on short-term trades, requires quick actions to gain profit and it forecasts the trend but does not spot if prices will rise or fall.

2.2.4 Turtle trading

Turtle trading is a strategy which follows the trend by looking at the market prices rather than depending on news to make trading decisions, being flexible with buy, sell signals and parameters and plan exit and entry strategy. Turtle trading proved that trading can be taught with a system and that by using a simple set of rules, non-traders can become successful traders. While some financial experts believe that this strategy is timeless, others claim that it would not be effective nowadays due to the increased number of trading contracts and large inflation.

2.2.5 Buy and hold

Buy and hold is a strategy which buys a stock and hold it for a long period of time without having to worry about the short time movement of the price and market's volatility. This strategy requires less hassle compared to short term trading and is effective and safer to generate more profit with a lower chance of making a wrong decision. Although this is a safe strategy, it means that the trader will not be able to access their capital for months or years. This method also takes a very long time until it returns profits with even a lesser return than expected. Lastly, if the market crashes, just like in the COVID-19 pandemic, the person will lose all their investment.

2.2.6 Trading Range Break (TRB)

TRB generates buy signals when the price is above the recent maximum and generates a sell signal when the price drops below the recent minimum. A Range is easy to spot, and it is easier to know when to enter a trade - when a price moves outside a range. This strategy is likely to also have multiple false breakouts where the prices move beyond the previous established price range and then goes back to the previous. Prices can also be corrected - when a breakout has happened and then the price reverts to its initial range which leads to a very small profit or small loss. Additionally, the price rarely has big moves to make profit.

2.2.7 Swing trading

Swing trading captures short and medium gains in stocks over a longer period and requires less time to trade compared to day trading which wraps up everything in a single day and requires more time per day. This technique focuses on trading and holding strategies which take hours, days, or weeks. There are also fewer positions with a bigger gain or loss. This trading strategy also uses technical analysis to search for trading opportunities. Nevertheless, stocks overnight and weekend have a higher market risk to change without being able to correctly trade.

2.3 Trading Problems

In this section we review what are the different problems which are faced for automatic trading such as Deep Reinforcement Learning [25]. In to understand how reinforcement learning fits into the equation, firstly we need to understand the problem it deals with. The problems are split into the following main categories: prediction problem, strategy, portfolio management and risk.

Prediction is related to accurately predicting time-steps in the future based on the past and the general dynamics of the time-series. For example, gold and oil prices do not fluctuate as much as crypto currency prices do. This can be done by using machine learning techniques [25-28] or deep learning [29-30]. With regards to Reinforcement learning, this is done by using model-based Reinforcement Learning, where a simulation of the environment is designed based on the seen data.

A trading strategy can be based on the predictions which were made and on external information such as news [31], social media sentiments [32] or other information sources which could influence the decision-making process. The trading strategy should also be customized if risk-averse or risk-seeking strategies are required. Estimating the risk can also be done through Deep Reinforcement Learning [33].

So far, the problems are related to dealing with a single asset. In contrast, portfolio management focuses on dealing with multiple assets. Initially all the crypto markets were correlated. For example, when Bitcoin prices were increasing, then the other crypto currency prices would grow as well. However recently the correlation has started to decrease which means that an investment strategy is required to trade the most favourable crypto currencies.

Risk in trading is inevitable. By using clever techniques, the reinforcement learning agent might follow risk-averse or risk-seeking strategies. The risk of the trade can be calculated by using either Sharpe or Sortino ratios which take into consideration risk-free as well.

The latter mention problems are related to designing the agent itself and designing the environment as well. Figure 2.5 displays the financial concepts and their relationship with Reinforcement Learning.

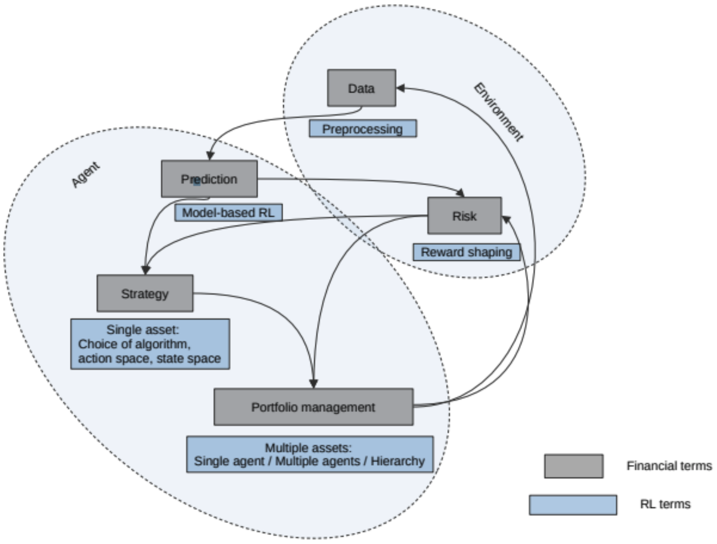


Figure 2.5: Financial concepts and their Reinforcement Learning correspondents

2.4 Common Algorithms

The most used algorithms related to trading are Actor-Critic and Policy-based algorithms such as A2C, PPO and DDPG.

Actor Critic methods consist of two parts – critic and actor. The critic is responsible for estimating the value function, either action-value (Q value) or state-value (V value). The actor is used to update the policy distribution in the direction the critic suggests. The key concept of Actor Critic measures is advantage. Advantage measures how much better it is to take a specific action compared to the average, general action at a given state.

$$A(s_t, a_t) = Q_w(s_t, a_t) - V_v(s_t)$$

This method does not require two neural networks, but it uses. The relationship the actor and critic have already established within the Bellman optimality equation. Advantage Actor Critic has two main variants - Asynchronous Advantage Actor Critic (A3C) and Advantage Actor Critic (A2C). A3C uses parallel training where multiple workers in parallel environments independently update a global value function. Asynchronous actors are efficient in exploration of the state space. On the other hand, A2C is like A3C but without having asynchronization.

Actor Critic methods are sensitive to perturbations. Small changes in the deep neural network, such as the weights, can cause large jumps in the policy space. Hence good performance can quickly become bad performance due to a small tweak. Proximal Policy Optimization (PPO) addresses this by limiting the updates to the policy network. This on-policy algorithm updates the base the update of each step on the ratio of new policy to old. The ratio is constrained within a given period to make sure that smaller steps are taken in the parameter space within the deep neural network. PPO also uses the advantage function concept which estimates how good an action is compared to average action for a specified state. This reduces the high variance within estimating the gradients between the old and new policy.

$$A(s, a) = Q(s, a) - V(s)$$

When the advantage function is positive, then the action which the agent took is good. If the advantage function is negative, that actions probability is decreased. PPO keeps track of a fixed length trajectory of memories and uses multiple network updates per data sample which use minibatch stochastic gradient ascent. This algorithm can also be parallelized like A3C. PPO can be used for either discrete or continuous action spaces.

Finally, Deep Deterministic Policy Gradient (DDPG) is an off-policy algorithm that can be used in environments with continuous action spaces. DDPG is learns a Q-function and a policy at the same time. It uses off-policy data and the bellman equation to learn the Q-function and uses the Q-function to learn the policy. DDPG's approach is very close to Q-learning and has similar motivations such as if the optimal action-value function $Q*(s,a)$ then in any given state the optimal action $a*(s)$ can be found by solving the following equation:

$$a*(s) = \underset{a}{\operatorname{argmax}} Q*(s,a)$$

2.5 Deep Reinforcement Learning Trading

In this section we review works related to reinforcement learning algorithms and how they are used for crypto currency trading. Categories which are used for trading are: Actor-Critic, Policy-Based, custom algorithms, and Hierarchical reinforcement learning. In financial services, a model-free approach with policy gradient methods could be effective at trading or portfolio management. A policy gradient method is more effective at learning the dynamics of trading rather than Q-learning which is a value-based method. In addition, risk-based Metrics like the Sharpe Ratio as reward functions outperform algorithms that use profit and loss reward functions.

The actor-critic approach has recently started being applied in trading to combine the advantages of the critic-only and actor-only advantages [34-36].

[34] proposes using MAC and RSI with a 30-day look-back window and normalizing the closing prices by using a custom technique by daily volatility adjusted to a time scale. Their agent's reward is driven by its actions rather than being affected by the market's volatility. This paper focuses on long and short trading which does not concern the to be designed agent at this time.

On the other hand, [35] goes into more details about its environment by describing how it works, its action and state space. The prices, amount of stock the agent has, and the agent's remaining balance are the states, and the agent can buy, sell, or hold any stock it has. These action increase and decrease the number of stocks the agent has. Their reward is the current portfolio amount the agent has within the new state. They split their data into three parts - train, validate and trading. On the contrary, the paper does not specify how are the buy and sell shares are generated. It is unclear if the amount of bought and sold stocks is built in the environment or if the agent selects it.

In this section we review works related to reinforcement learning algorithms and how they are used for crypto currency trading. Categories which are used for trading are: Actor-Critic, Policy-Based, custom algorithms, and Hierarchical reinforcement learning. In financial services, a model-free approach with policy gradient methods could be effective at trading or portfolio management. A policy gradient method is more effective at learning the dynamics of trading rather than Q-learning which is a value-based method. In addition, risk-based Metrics like the Sharpe Ratio as reward functions outperform algorithms that use profit and loss reward functions.

The actor-critic approach has recently started being applied in trading to combine the advantages of the critic-only and actor-only advantages [34-36].

[34] proposes using MAC and RSI with a 30-day look-back window and normalizing the closing prices by using a custom technique by daily volatility adjusted to a time scale. Their agent's reward is driven by its actions rather than being affected by the market's volatility. This paper focuses on long and short trading which does not concern the to be designed agent at this time.

On the other hand, [35] goes into more details about its environment by describing how it works, its action and state space. The prices, amount of stock the agent has, and the agent's remaining balance are the states, and the agent can buy, sell, or hold any stock it has. These action increase and decrease the number of stocks the agent has. Their reward is the current portfolio amount the agent has within the new state. They split their data into three parts - train, validate and trading. On the contrary, the paper does not specify how are the buy and sell shares are generated. It is unclear if the amount of bought and sold stocks is built in the environment or if the agent selects it.

[36] goes into vast details describing deep actor critic methods such as Critic and Q-learning, Actor and Policy Gradient and Actor Critic with Deep Networks. This paper proposes using LSTMs as a neural network, so the useful cell states are kept, and the redundant data is forgotten. This paper also takes into consideration the transaction fees, where [34-35] do not mention them. However, the action space and the reward formulation is omitted.

[37] proposes using Moving Average Convergence Divergence (MACD) with the open, high, low, close, adjusted close and volume. The agent can either buy or sell but the amount of purchasing and selling is not mentioned. The paper uses Proximal Policy Optimization (PPO), Advantage Actor Critic

(A2C), Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3). The paper uses the Sharpe Ratio to assess the overall performance of the agent and the reward is in fiat currency – Rupees.

[38] proposes using the DRLMM framework with advanced policy gradient-based methods. The agent's reward and action space differ from the above-mentioned papers. The environment uses profit and loss for the current step t as a reward and inventory ratio, total profit and loss, unrealized profit and loss, limit order distance and order completion ratio as the agent's state space. This paper also introduces normalizing the training data by using z-score normalization. As other research, this one also uses A2C and PPO. It is also mentioned that the agents trade on a second based candlestick chart.

Another known method is ensembled reinforcement learning actor-critic algorithms - Advanced Actor Critic (A2C), Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO) [39] which select the best trading action based on the Sharpe ratio. The paper's state space consists of the available balance at the current timestep, the adjusted close price, the shares owned of each stock, Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Commodity Channel Index (CCI) and Average Directional Index (ADX). The agents can buy or sell stocks within a predefined limit. The reward designed to maximize the change of the portfolio. However. The agents must be retrained every three months in parallel because one could perform well in a bullish trend but act bad in a bearish trend whereas another agent can be adjusted to the market's volatility. The suggested ensemble strategy also performed in the 2020 market crash due to a predefined turbulence index which allows the agents to cut their losses and survive.

[40] uses a completely different approach from all the above-mentioned papers. The agent's objective is to only maximize the short-term profits based on the following trading strategies – double crossover, day trading, swing trading, scalping and position trading. The model can is punished

with a negative reward if it has executed the same action for a given number of consecutive times. The data is based on an hourly candlestick pattern. Nevertheless, the paper never mentions what Reinforcement Learning algorithm except that it uses a Deep Neural Network to interact with an environment. Neither of the above-mentioned deep reinforcement learning have been used.

[41] selects top-volumed crypto currencies for Bitcoin portfolio management on a 30-minute market pattern. Bigger crypto volumes impliy better market liquidity of an asset. This also suggests that an investment can have less influence on the market. The research notices that the crypto currency market is not stable and that coins could have a sudden boost or drop in their volume for a short period of time. With that in mind, the timeframe takes into consideration the volumes of the previous 30 days. DPG is proposed to be used with either a CNN, basic RNN or a LSTM.

Another less researched approach is Hierarchical Reinforcement Learning (HRL) [42-44]. HRL's policy is decoupled into different components where a sub-policy is responsible for solving a sub-task. The agents can interact sequentially or simultaneously. [42] consists of two agents – Order Determination and Bid Execution. The Order Determination observes the current order book and estimates the quantity to be bought, sold, or held. Bid Execution takes Order Determination's generated result and the state space, which consists of S&P500's open, close, high and low prices and the market's volume on a single minute basis. Both policies can be updated jointly or asynchronously by using on-policy or off-policy setting and TradeR's reward function is used by both agents. [43] uses two highly customised identical Deep Q Networks (DQN) which are responsible for three assets – one for cash and the other for two stocks. The custom Neural Network layers consist of 3 CNN layers, where the third layer reuses the weights of the second one, and finally a dense layer. As. [42] the normalized open, close, high, and low prices are used. These papers focus more innovative ways of solving a MDP compared to the previous ones.

Finally [46] suggests an unusual approach of pre-processing financial data for Recurrent Reinforcement Learning (RRL) by using Principal Component Analysis (PCA). And Discrete Wavelet Transform (DWT). Finally, the data is given to the agent to interact with. This paper is based on the general belief that technical analysis indicators can summarize the general pattern of the time series and reduce data noise which can be further utilized by the agent. The paper's indicators can be summarised into four different groups – momentum, volatility, cycle, and volume. As [38] the data is normalized by using z-score. The reward, as majority of the studies, is based on Sharpe ratio.

2.6 Common flaws

Most of the papers have many flaws in common. The most obvious one is that majority of the papers do not use the crypto market but use forex exchange or stock market. Even though they have the same fundamentals such as an open, close, high, and low price and the same technical analysis metrics are used, the markets are very different from each other. For example, the crypto market is very volatile and there can be significant differences within the prices for even an hour.

Another flaw is that majority of the papers stick to a single candle stick pattern such as daily, hourly, 30 minutes or 1 minute. Experienced traders understand that to get the most out of a single trade, different candle sticks must be looked. Different timescales give different kind of information and when combined a trade can be finally executed.

Some of the papers take into consideration transaction fees and other possible taxes. This still leads to less profit over time and should not be fully ignored. The more trades that are done per day, the more taxes are being paid. It might be more profitable to look for an individual bigger trade than have various small ones.

The main concern on this point is that a reinforcement learning system could possibly overfit if the data is given in chronological order. Through enough time steps the agent may start picking up the patterns themselves instead of learning from its behaviour. An agent must be able to trade without overfitting the data and perform well on unseen data. One way to. Avoid this is by randomising the data. Instead of going through the data in a chronological order, the data can be shuffled so there is no overfitting. This is based on the random walk theory which suggests that prices have the same distribution and are independent of each other. The trend, price or market's past cannot be used to predict the future [47].

Finally, the custom algorithms consist of custom neural networks that are used to generalise. The core reinforcement learning part is usually A2C, PPO, DDPPG or DQL. No research has focused on trying to edit the algorithms' core parts to try and enhance its performance specifically for trading.

Chapter 3 - Methodology

This section broadly aims to help researchers and others who are interested in replicating the work and come with new ideas of improvement. By defining a methodology, the reader understands what work is done without any second doubts.

3.1 Method of development

This research aims to find an optimal base solution for a final product. This suggests that the prototype will have a limited but essential functionality. By having the base work of a difficult task, it can help identify difficult and confusing requirements before even starting the implementation process. This research offers a deep reinforcement learning agent with a custom trading environment. Having a prototype allows taking into consideration missing functionalities and further ways to improve the research.

After a deep understanding of related work on reinforcement learning trading, the most common methods will be investigated. The agents will be A2C, PPO and DDPPG. These algorithms perform significantly better than others when trading because they make very small updates on their parameters in a way that also does not allow the policy to change as much. However, the most challenging part is building the environment and formulating a reward.

Making a successful trade requires a lot of information on different candlestick charts and every trader has a different way of determining their outcome such as pure profit and loss on exit, profit per tick, Sharpe ratio, binary wins, and losses etc. However, there is not only one correct way of determining if a trade was successful or not. The data features that are used in this research are the prices (open, high, low, close), technical indicators and different time scales (daily, hourly, etc.). The data itself is collected from the Binance API which fetches the history prices of different crypto currencies for different time frames.

3.2 Research Approach

The nature of this research is a quantitative experiment. The research is managed in a controlled environment with measurable data and the results can be compared and statistically analysed. This approach allows a comprehensive in-depth trading data exploration which could lead to meaningful decisions. The quantitative experiment includes:

Step 1: *Identify a problem*

Reinforcement learning can lead to faster and better trading solutions which will increase the profit. The data quality side has mostly been neglected in favour of the reinforcement learning point of view - how to improve existing solutions and design new ones. Although that is the main objective in the field or trading for reinforcement learning, there are currently more than enough profitable solutions but not enough focus on enhancing the dataset.

Step 2: *Find what causes the problem*

The problem is that researchers understand the technical part of the problem and the maths behind technical analysis, but they lack from a trading point of view. The way a trader and a researcher would approach trading significantly differ from each other.

Step 3: *Design a solution*

To solve the root cause of the problem the first step is to understand the technical indicators, what they do and how the crypto market works. After the indicators have been understood, they can be combined to get more useful information that can hint the agent when to buy or sell.

Step 4: *Test the solution*

The solution will be tested in Python’s openai gym environment. This is a library used for developing and comparing reinforcement learning algorithms by provides the communication between learning algorithms and environments.

Chapter 4 - Design and analysis

4.1 Dataset

The dataset used within this research is from the Binance API. Binance is the largest crypto currency exchange in the world for trading. Unfortunately, the platform has historical data available only from when Binance began offering that crypto currency's rather than having the whole history of the coin's market. For example, Binance has the full history of Bitcoin from 2017 rather than from 2009 when Bitcoin was initially launched. This study will focus on the BTC-USDT (Bitcoin to USD Tether) pair on a daily and hourly basis. Currently Bitcoin has the highest price, is the crypto market leader and the entire crypto market follow's its bearish and bullish trends.

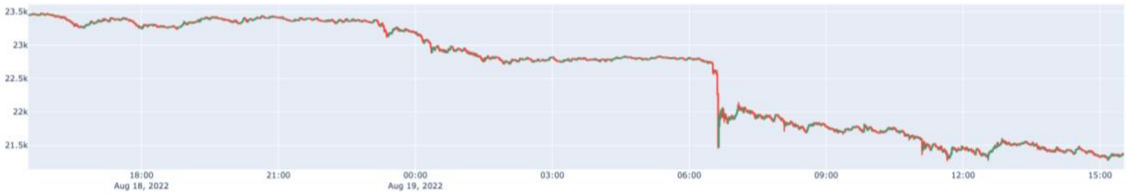
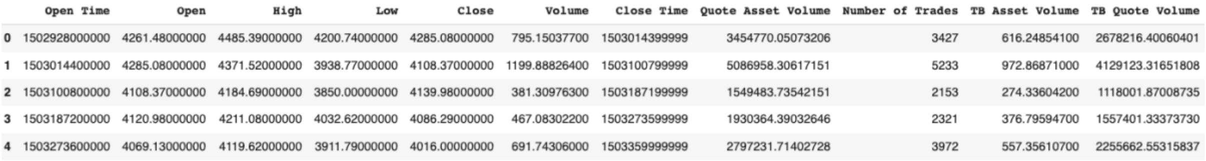


Figure 4.1: Example hourly BTC-USDT data

The data returned from Binance is in a dictionary format and it is converted to a Pandas data frame.



	Open Time	Open	High	Low	Close	Volume	Close Time	Quote Asset Volume	Number of Trades	TB Asset Volume	TB Quote Volume
0	1502928000000	4261.48000000	4485.39000000	4200.74000000	4285.08000000	795.15037700	1503014399999	3454770.05073206	3427	616.24854100	2678216.40060401
1	1503014400000	4285.08000000	4371.52000000	3938.77000000	4108.37000000	1199.88826400	1503100799999	5086958.30617151	5233	972.86871000	4129123.31651808
2	1503100800000	4108.37000000	4184.69000000	3850.00000000	4139.98000000	381.30976300	1503187199999	1549483.73542151	2153	274.33604200	1118001.87008735
3	1503187200000	4120.98000000	4211.08000000	4032.62000000	4086.29000000	467.08302200	1503273599999	1930364.39032646	2321	376.79594700	1557401.33373730
4	1503273600000	4069.13000000	4119.62000000	3911.79000000	4016.00000000	691.74306000	1503359999999	2797231.71402728	3972	557.35610700	2255662.55315837

Figure 4.2: Example Binance data converted to a Pandas data frame

The columns from left to right are as follows:

- **Open Time** - the opening time of the currency
- **Open** - the open price of the currency.
- **High** - the highest price of the currency for the day.
- **Low** - the lowest price of the currency for the day.
- **Close** - the closing price of the currency for the day.
- **Volume** - number of units traded in a market during a given time.
- **Close time** - the closing time of the currency.
- **Quote asset volume** - total amount (volume) in USDT (or another asset).
- **Number of trades** - total number of trades at a particular price or period of time.
- **Taker buy base asset volume** - total amount (volume) of base asset (eg. LUNA, SHIBAinu) for Taker (market) buy order.
- **Taker buy quote asset volume** - same as taker buy base asset volume but for quoted asset (in this case USDT).

The Binance API has every ticker which is available on Binance's trading platform. The API also gives access to the user's personal information such as their available crypto volume.

4.2 Exploratory Data Analysis

Before setting up the agent's environment the data will be explored to try finding any useful information that can be used when doing the designing process.

4.2.1 Timeframe of most executed trades

Unlike FX and the stock market, the crypto market is open 24/7 which means that trades can be executed any times. Figure 4.3 showcases the volume per hour of day. The quietest hours are from 1 am to 7 am and from 5 pm to 11 pm. The dataset's time zone is Greenwich Mean Time (GMT).

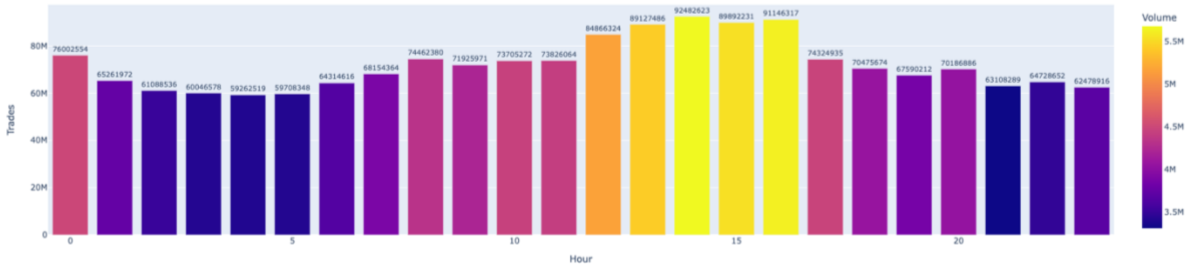


Figure 4.3: BTC-USDT transactions and volume grouped by hour of day from 17-08-2017 to 30-08-2022

The agent can trade on a 24-hour basis, which gives it more data to interact with.

4.2.2 Trading metrics

The most used metrics in Reinforcement Learning trading research are – Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Commodity Channel Index (CCI) and Average Directional Index.

Moving Average Convergence Divergence (MACD)

Moving averages are trading indicators that aim to 'smooth out' price fluctuations to help separate trends from general market activity. MACD follows the momentum between two moving averages of a price. The primary signals for MACD are:

- **MACD** - the value of an exponential moving average (EMA) subtracted from another EMA with a shorter lookback period. Common values are 26 days for the longer EMA and 12 for the shorter. This is referred to as the 'slow' signal line of the MACD. This is the blue line on figure 4.4.

- **Signal** - the EMA of the MACD of a period shorter than the shortest period used in calculating the MACD. Typically, a 9-day period is used. This is referred to as the 'fast' signal line. This is the red line on figure 4.4.
- **Difference** – the difference between the MACD – Trigger line is used to represent current selling pressures in the marketplace. This value is commonly charted as a histogram overlaying the MACD + Trigger signals. A positive value for this signal represents a bullish trend whereas a negative value indicates a bearish market.

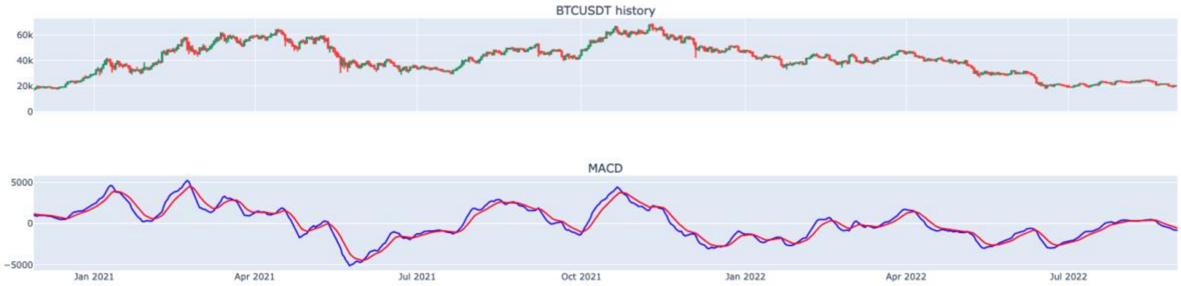


Figure 4.4: BTC-USDT price history and MACD from 28/11/2021 to 29/08/2022

When there is a crossover between the Signal and MACD lines that means the market is trending. However, when the MACD line crosses above the Signal line this is a buy indicator and when the MACD line crosses below the Signal line this is a sell indicator. The MACD and Signal check will compare two consecutive days to see where the lines cross.

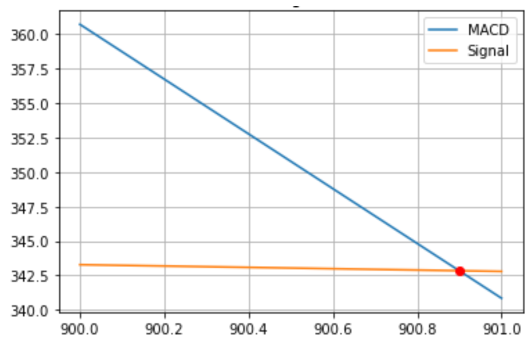


Figure 4.5: MACD x Signal - Sell

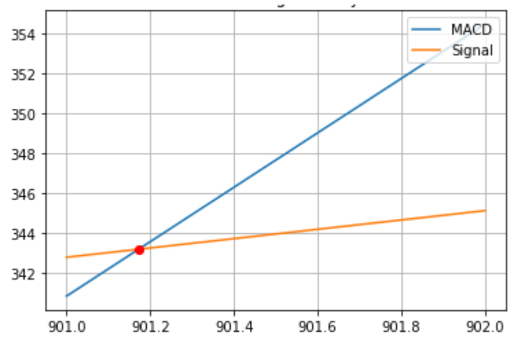


Figure 4.6: MACD x Signal - Buy

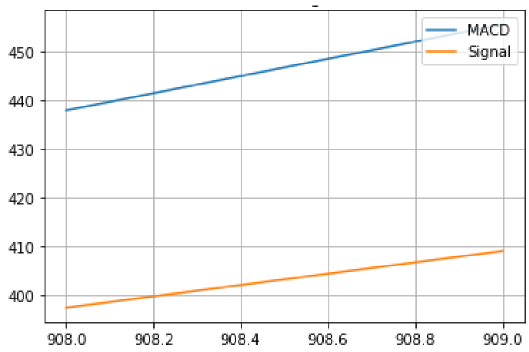


Figure 4.7: MACD x Signal – no buy or sell signal

Relative Strength Index (RSI)

An effective trading strategy would be to see which way the trend is pointing - is it going up or is it going down. RSI is a momentum indicator that describes the current price relative to average high and low prices over a previous trading period. This indicator estimates overbought or oversold status and helps spot trend reversals, price pullbacks, and the emergence of bullish or bearish markets. Traditionally the RSI is considered overbought when above 70 and oversold when below 30.

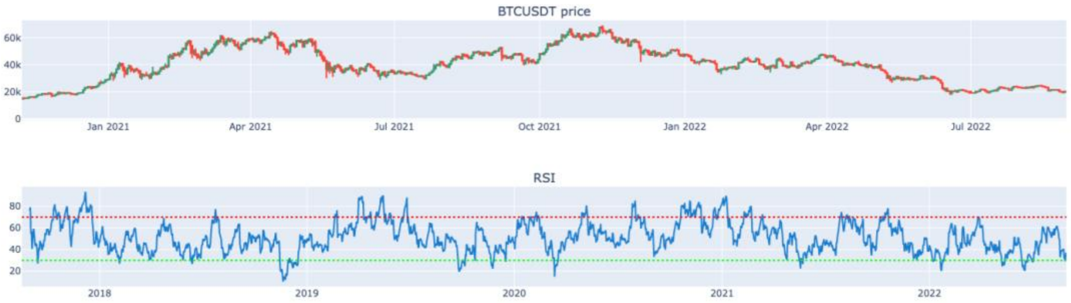


Figure 4.8: BTC-USDT price history and RSI

Average Directional Index (ADX)

ADX is used to determine the strength of a trend. The trend can be either up (positive directional indicator +DI) or down (negative directional indicator -DI). This metric consists of three different lines which are used to further support if a trade taken at all or if the trade should be taken short or long. The trend has strength when ADX is above 25 and the trend is weak / prices are trendless when ADX is below 20. Non-trending can also mean that the price is making a trend change or is currently too volatile for a clear direction. When +DI rises above -DI this means that it is time to buy, whereas when -DI rises above +DI this indicates that it is time to sell.

ADX Value	**Trend Strength**
0 - 25	Weak
25- 50	Strong
50 – 75	Very strong
75 – 100	Extremely strong

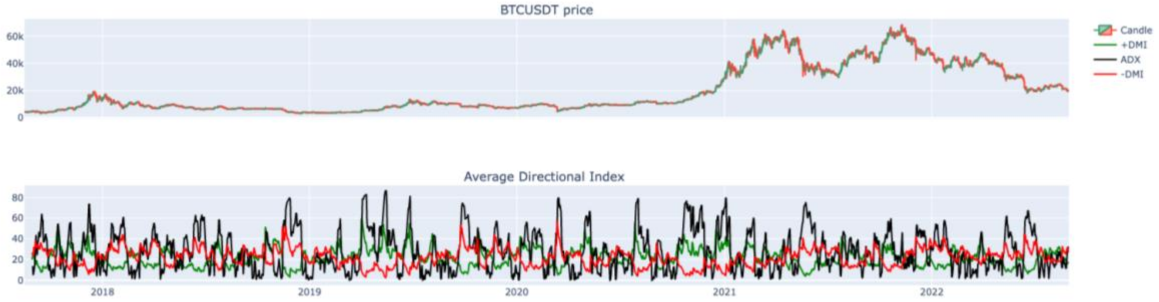


Figure 4.9: BTC-USDT price history and ADX

Commodity Channel Index (CCI)

This is a lagging momentum oscillator that signals overbought and oversold levels by measuring the difference between current and historical prices. This metric does not have set limits (like RSI). Usually, the conditions are identified when the line(s) are outside a range of +100 (overbought) or -100 (oversold).

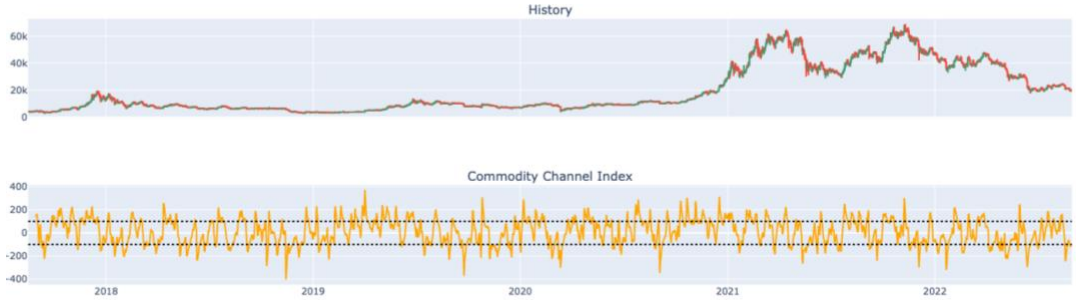


Figure 4.10: BTC-USDT price history and CCI

4.2.3 Volatility

Volatility is a variation of a trading price series over time. Crypto currencies are popular for being very volatile. Volatility is a category of indicators that are used to measure how volatile the price is in a specific period.

Double Bollinger Bands

Bollinger Bands are used as overbought and oversold indicators but given the trending nature of crypto, there are more efficient ways to use the bands. Bollinger Bands consist of three lines - the 20-period moving average and two standard deviations. One of the STDs is above and the other is below the moving average. The Bollinger Band Squeeze occurs when volatility falls to low levels and the Bollinger Bands narrow. According to John Bollinger, periods of low volatility are often followed by periods of high volatility. Therefore, a volatility contraction or narrowing of the bands can foreshadow a significant advance or decline.

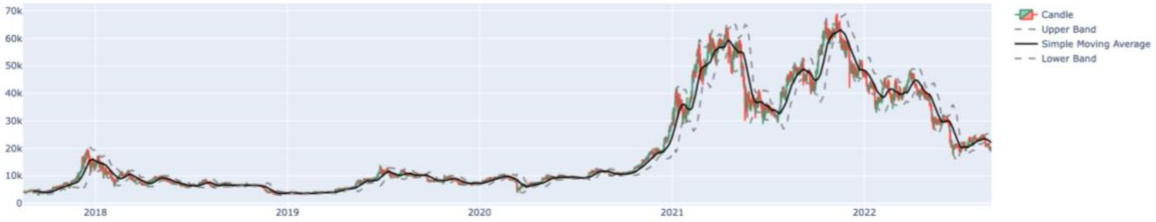


Figure 4.11: BTC-USDT price history and Double Bollinger Bands

Average True Range (ATR)

ATR can tell how volatile an asset has been on average over a specified period. This metric is useful for setting exit levels as part of a risk management strategy. It also tells how strong price movements are to identify a trend. True Range is a way of measuring an asset's daily trading range that accounts for gap openings and gives a sense of how much the asset's price has moved over time.

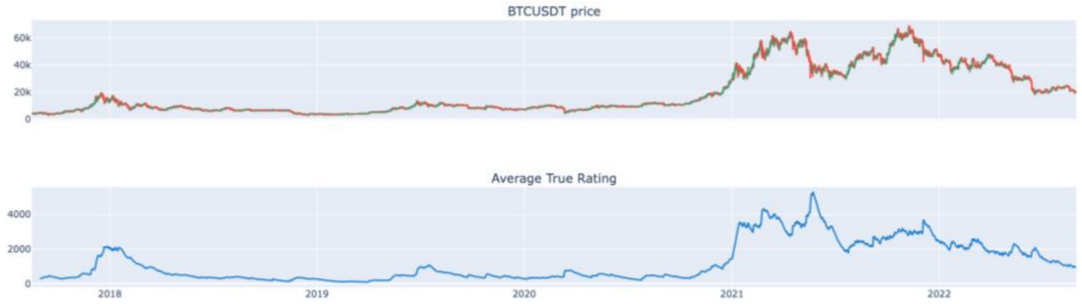


Figure 4.12: BTC-USDT price history and ATR

4.3 Data Preprocessing

The agent will trade on the 1-hour candle sticks. They offer flexibility in term what the agent can do, and the agent is given enough time to analyse trading opportunities. The 1-hour candle sticks also give more chances to trade to maximize profits. The daily charts will be used to generate signals if it is a good or bad day to trade whereas the actual trading will occur on an hourly chart for the days for the agent to learn to trade.

4.3.1 Used timeframe

It is a common practice by traders to use different candlestick timeframes to make the best trading decision. We will use the daily and 1 hour candlestick charts. The daily chart will be used to determine the market's trend, bearish or bullish, and the hourly chart will be used to determine the price's volatility and will be given to the agent to make a buy / sell / hold decision. To produce an accurate signal about the market’s buy or sell signal all three indicators must be aligned.

The daily chart will give us an insight on the daily market trend. We will use 3 different metrics to determine the market's trend - MACD, ADX and Ichimoku Cloud.

Ichimoku Kino Hyo

Reinforcement Learning researchers have not used Ichimoku Kino Hyo (Ichimoku Cloud) but it combines the market's momentum, trend, support and resistance into one indicator. The five lines of Ichimoku combine to form a cloud which collapses the information from all five linesi into an easier to read formation.

- **Tenkan-Sen (Baseline)** - is the moving average (red line) and represents. the trend (upwards, downwards, or sideways).
- **Kijun-Sen (Conversion line)** - acts as support and resistance (blue line). It is similar to Tenkan-Sen (Baseline) but has a longer timeframe. This line usually lags the Tenkan-Sen (Baseline).
- **Senkou A (Leading Span A)** - is the average of highs and lows of the latter two lines (orange line).
- **Senkou B (Leading Span B)** - is an expanded version of Senkou A (Leading Span A). If Senkou A (Leading Span A) is above Senkou B (Leading Span B) then the "cloud" is green and if Senkou A (Leading Span A) is below Senkou B (Leading Span B) the "cloud" is red.
- **Chikou Span (Lagging Span)** - shifts the current price 26 periods leftward (grene line).

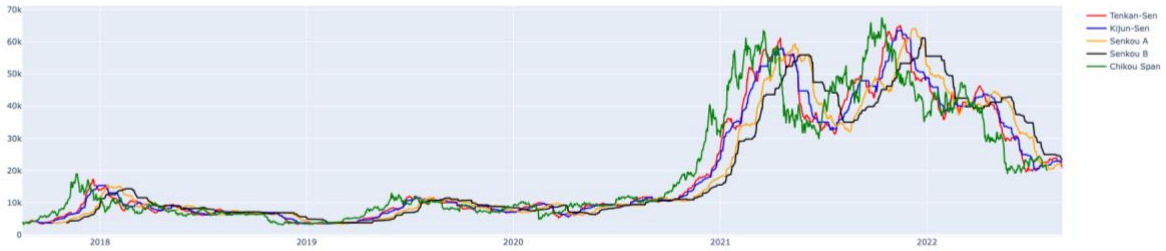


Figure 4.13: BTC-USDT Ichimoku Kino Hyo

The Ichimoku signals are as follows:

1. **Buy**
    a. Prices rise above the cloud
    b. Cloud turns from red to green
    c. Price moves above Kijun-Sen (Conversion line)
    d. Tenkan-Sen (Baseline) rises above Kijun-Sen (Conversion line)

2. **Sell**
    a. Prices fall below the cloud

b. Cloud turns from green to red
c. Prices moves below Kijun-Sen (Conversion line)
d. Tenkan-Sen (Baseline) falls below Kijun-Sen (Conversion line)

Fear and Greed Index

Another famous crypto currency metric is the Fear and Greed Index. It gives traders way to gauge stock market movements and whether stocks are fairly priced. The theory is based on the logic that excessive fear tends to drive down share prices, and too much greed tends to have the opposite effect.

The Fear and Greed Index is a compilation of seven different indicators that measure some aspect of stock market behaviour. They are market momentum, stock price strength, stock price breadth, put and call options, junk bond demand, market volatility, and safe haven demand. The index tracks how much these individual indicators deviate from their averages compared to how much they normally diverge. The index gives each indicator equal weighting in calculating a score from 0 to 100, with 100 representing maximum greediness and 0 signalling maximum fear.

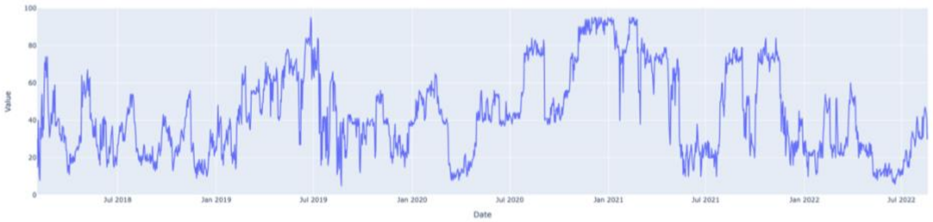


Figure 4.14: Fear and Greed Index

On the other hand, the hourly chart gives the agent more opportunities during the day. As previously discussed, the crypto currency market is 24/7 and there is no period with a big drop of trading. Even though ADX was used to generate a buy or sell signal on the daily candlesticks, this indicator will also be used on an hourly basis as well, so the trend's strength is generated per hour rather than daily.

4.3.2 Missing data

Fear and greed had missing data for 14/04/2018, 15/04/2018 and 16/04/2018. The mean of the previous five days from the current missing values will be used to generate the missing ones. The other missing values will be dropped out for simplicity. They are generated when the indicators are calculated because the calculations are based on a specific number of candlesticks until indicator values are produced. The data from Binance did not have any missing values when initially downloaded.

4.3.3 Data transfer

After the indicators have been generated for the hourly and daily candlesticks, the daily indicators are copied to the hourly ones based on the day. This means that every day (24 hours) there will be hourly indicators, which slightly differ from each other, and daily ones, which are the same for all 24-hour datapoints. This makes the dataset take place between 01/02/20218 and 25/08/2022.

4.3.4 Label encoder

The signal generated columns from Ichimoku cloud, MACD and ADX have string values of Buy, Sell or None. These target values are converted to class numbers so they can be given to the agent.

Ichimoku Signal	**Encoding**
Buy	0

Sell	1
MACD Signal	
Buy	0
Sell	1
ADX Signal	
Buy	0
Sell	1

4.3.4 Data normalization

Different value ranges and needs to be scaled down to a common scale without distorting differences in the ranges of or losing information. A common practice for scaling financial data is Z-Score. In Finance the Z-Score is used to predict the likelihood of a company declaring bankruptcy. The Z-Score refers to how many STDs a value is from the mean of the data. A Z-Score of 1 means that the value is 1 STD from the mean whereas a Z-Score of 2 means 2 STDs. This metric is useful when comparing data points from different datasets.

$$Z = \frac{x - \mu}{\sigma}$$

Z - standard score

X – observed value

μ - mean of the sample

σ − standard deviation of the sample

4.4 Data split

The modified data is 1666 days - from 01/02/2018 00:00:00 to 25/08/2022 13:00:00 on an hourly basis. The data will be split into three parts - train, validate and test.

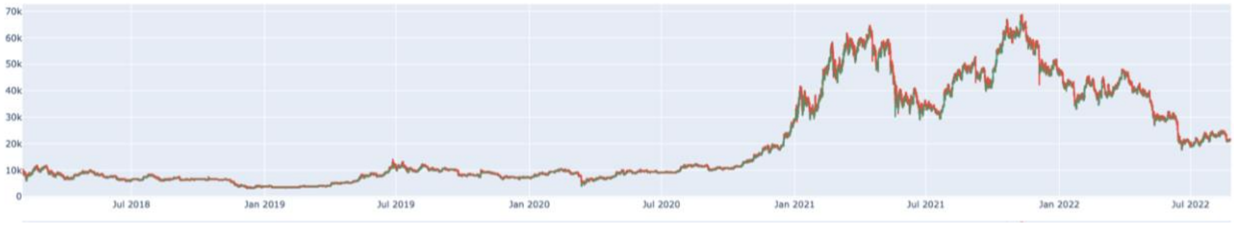


Figure 4.15: BTC-USDT price history from 01/02/2018 to 25/08/2022

The training data is given to the agent so they can learn the environment and learn to make decisions. The validate data segment is used to see how the agent will interact with never seen data. This gives us the opportunity to go back and tweak the agent's performance. The test dataset is used to evaluate the agent as if they were trading in live time.

The validation and test sets must be big enough so the model's performance can be accurately measured. To get most of the dataset, it will be split in the following way:

- **train** - 85% of the dataset
- **validate** - 7.5% of the dataset
- **test** - 7.5% of the dataset

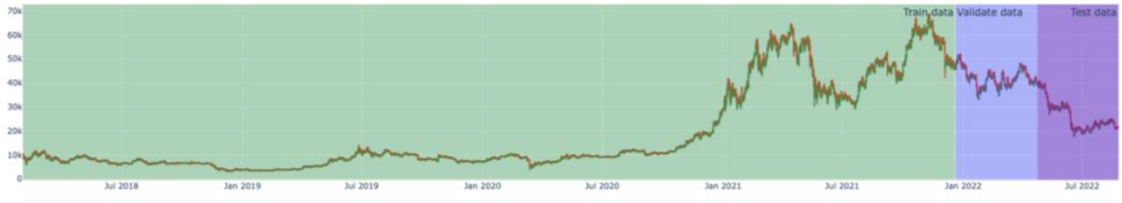


Figure 4.16: Data train, validate and test split from 01/02/2018 to 25/08/2022

4.5 Environment

To develop the agent's training environment the **gym** Python library will be used. gym is an open-source library for developing reinforcement learning algorithms by providing a standard API that communicates between the agents and the environment.

4.5.1 Actions

The agent will be able to perform the following actions:

- **Buy (0)** - buy Bitcoin with fiat currency (USD). The agent will an initial starting amount. The agent will be able to buy Bitcoin between 0 to 100% of the current money it has as fiat currency. When the agent performs a buy operation the fiat currency is converted to Bitcoin USD.
- **Sell (1)** - buy Bitcoin from USD Tether (USDT) to fiat currency (USD). The agent will use Bitcoin's USD current price (at the time of the trade). The agent can sell between 0 to 100% of the Bitcoin that it currently has. When the agent performs a sell operation the Bitcoin USD is converted to fiat currency (USD).
- **Hold (2)** - neither buy nor sell at the current price.

4.5.2 Observation space

The agent will have the current amount of fiat currency in percent percentages (between 0 and 1) and it will also be able to observe the following Z-score normalized columns:

- **Hourly based information** - Open, High, Close, Volume, ADX and Volatility. The volatility is calculated by extracting the lower Bollinger band from the upper Bollinger band. The higher the number, the higher the volatility and vice versa.

- **Daily based information** - Tenkan-Sen, Kijun-Sen, Senkou A, Senkou B, Ichimoku Signal, MACD, MACD Signal, MACD History, MACD Signal, +DMI, -DMI, ADX Signal and Fear and Greed

To reduce overfitting, the data will be randomized. This is based on the random walk theory which suggests that prices have the same distribution and are independent of each other. The trend, price or market's past cannot be used to predict the future.

4.5.4 Reward

The agent's reward depends on different factors, and it changes based on the amount of money the agent has. If the money falls below a given threshold, which is provided by the user. The important indicators for all rewards are - amount of buy or sell signals, ADX, Fear and Greed and the amount of USD the agent has bought one Bitcoin for.

The agent can give a number to buy or sell between 0 and 1. This is linked to how much the agent would like to buy Bitcoin from with its fiat currency and how much the agent would like to sell from its Bitcoin portfolio based on the market price at that time. The positive or negative reward is based on the average price the agent has bought 1 Bitcoin for. If the current day's reward is positive, then the agent is either making a profitable buy or sell action. This is known as today's reward.

$$mean = \frac{\sum_{i=1}^{i} current\ amount}{\sum_{i=1}^{i} current\ volume}$$

mean – average buying Bitcoin price of the agent based on their purchase

current amount – the amount the agent has spent when purchasing Bitcoin at a specific time *current volume* – the volume the agent has bought with *current amount* when purchasing Bitcoin

$$reward = \frac{\text{current} - \text{mean}}{mean}$$

current – current price of Bitcoin when taking an action

The reward will use the amount of buy or sell signals there are from columns MACD Signal, ADX Signal and Ichimoku Signal. The reward is multiplied based on how many buy and sell signals there are and what action the agent took. The ADX indicator gives only the strength of the trend and will be used as a reward multiplier. The Fear and Greed metrics will also be used as a reversed multiplier. The lower the Fear and Greed index is, the more people are "scared" to buy or sell Bitcoin. The last common indicator that is used to calculate the reward is the volume the agent will buy or sell. This is a number between 0 and 1 which corresponds to 0 and 100% of the crypto portfolio or the agent's fiat money depending on the performed action.

$$new\ reward = \begin{cases} ADX * \left(1 - (Fear\ and\ Greed * 001)\right) * reward, & signals = 0 \\ signals * \ ADX * \left(1 - (Fear\ and\ Greed * 001)\right) * reward, & signals > 0 \end{cases}$$

From this point onwards the agent’s action will impact the reward it gets. The objective of this reward is to encourage buying when the price is below the average price that the agent has bought and sell when the price is above. A threshold will give us additional reward tweaks. If the agent buys or sells below the threshold the agent should get a less harsh punishment and try to go above the threshold.

1. **Buy** (0)

$$buy\ reward = \begin{cases} new\ reward * \dfrac{current}{10}, & current > threshold \\ new\ reward * \dfrac{current}{10} * \dfrac{new\ amount}{initial}, & current \leq threshold \end{cases}$$

2. **Sell** (1)

$$sell\ reward = \begin{cases} new\ reward * \dfrac{100 - current}{10}, & current > threshold \\ new\ reward * \dfrac{current\ - 100}{10}, & current > 100\% \\ \dfrac{new\ reward}{threshold - current} * \dfrac{100 - current}{10}, & current \leq threshold \end{cases}$$

3. **Hold** (2) – the agent is only given a punishment if it chooses. To hold when it could have made some profit from selling Bitcoin. This encourages the agent to trade more often than just buy and hold until a new high price has reached.

$$positive\ punishment = \begin{cases} new\ reward, & signals = 0 \\ signals * new\ reward, & signals > 0 \end{cases}$$

$$hold\ reward = \begin{cases} new\ reward, & signals \leq\ 0 \\ signals * new\ reward, & signals > 0 \end{cases}$$

signals – buy or sell signals, depending on the agent's action

ADX - the ADX signal which is a number between 0 and 100

current – the agent's current amount of fiat money in percentages

threshold – a threshold to indicate when to reduce the rewards received

current amount – the current amount of fiat money the agent has at the time of executing the trade *new amount* – the amount of fiat money the agent would have after purchasing Bitcoin

initial amount – the initial amount the agent starts with

There are two extreme cases that could be reached within the environment - when the agent does not have any fiat money remaining there will be a negative reward of the initial started amount. For example, if the agent started off with £100000, the punishment would be -100000. The other extreme scenario is when the agent has reached more than a "maximum" amount which should be the objective of the agent in this case the reward will be a positive number based on the desired amount of USD.

4.5.5 Hyperparameters

In this section all the hyperparameters will be shown with their default values for this research.

Variable	**Value**
Threshold	25
Initial balance	10^5
Max balance	Initial balance + (initial balance * 0.5)
Min balance	Initial balance * 0.01
Initial volume	1
Transaction fee	0.001
Total timesteps	10^6

4.6 Results

The agents that were used for this environment are A2C, PPO and DDPG from stable baseline3. The results are recorded by using stable baseline3's TensorBoard. The agents were trained on 1 million timesteps. Figure 4.17 represents Bitcoin's price history from 20/12/2021 until 23/04/2022 when the validation takes place. The agent begins with 1 Bitcoin which costs $50,000. However, price has been mainly going below $50,000.

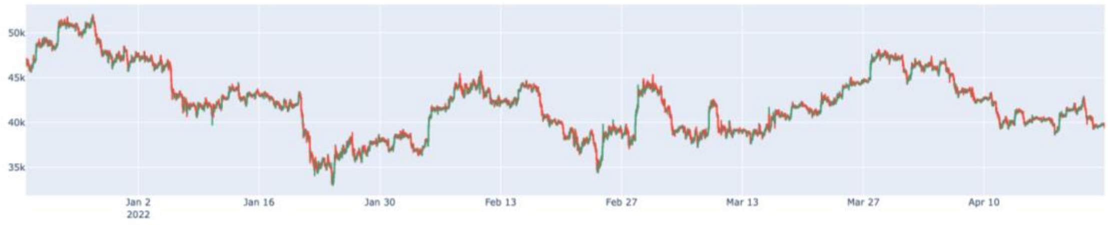


Figure 4.17: Agent's validation data - 20/12/2021 to 23/04/2022

In this period, Bitcoin's price started getting very volatile and unpredictable, mostly losing its high price in a couple of months. Now of writing this dissertation Bitcoin's price is holding at $20,000. Figure 4.18 showcases that the agent has decided to hold from start to finish because there were no profitable actions to take. All the agents did the same action - hold.

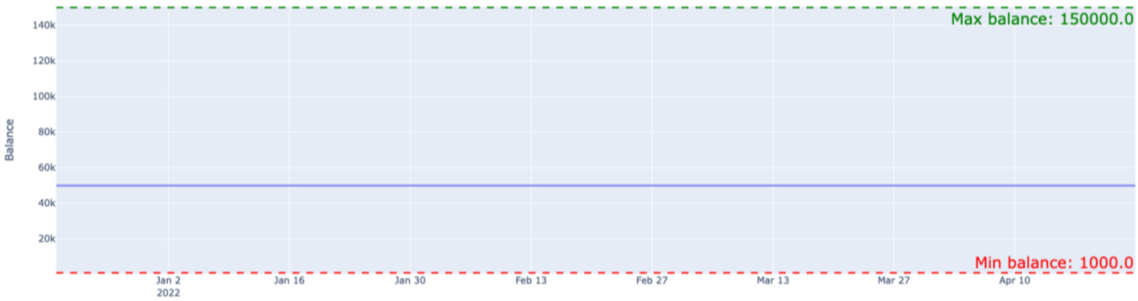


Figure 4.18: Agents' balance for validation data - 20/12/2021 to 23/04/2022

4.6.1 A2C

With more timesteps the A2C's average episode length increased however the reward was lowest before the agent reached 100K steps. As Figure 4.18 shows, A2C's reward was always a negative number, however towards the end of the training the model learned to keep negative rewards to a minimum.

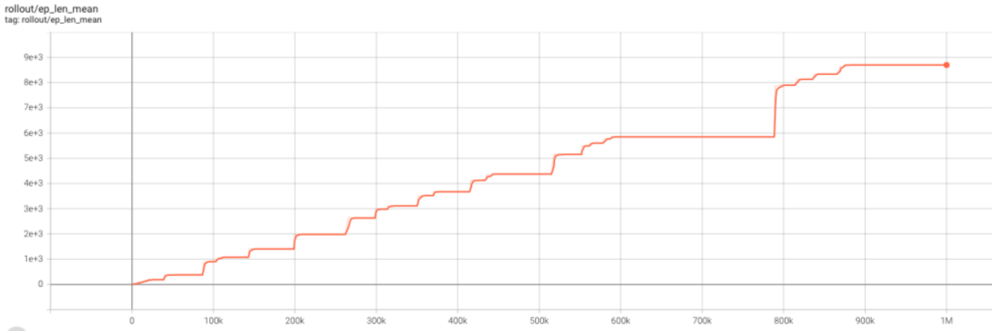


Figure 4.17: A2C average episode time

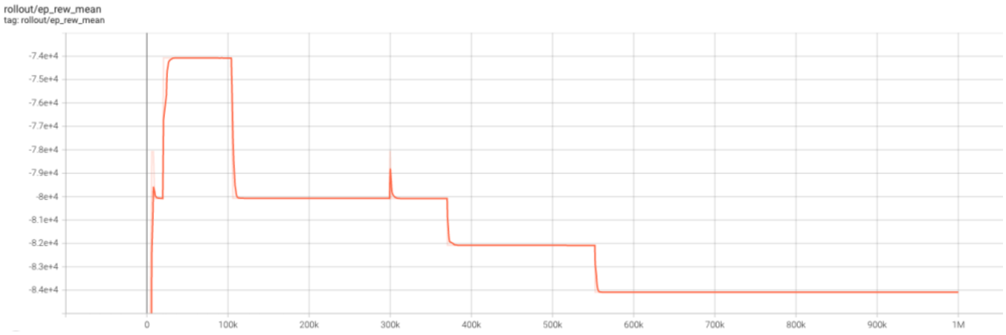


Figure 4.18: A2C average reward

4.6.2 PPO

As A2C, PPO’s average episode length increased with time but not as much as A2C. PPO’s reward was lowest at 200K steps and towards the end it also learned how to avoid keep negative rewards to a minimum.

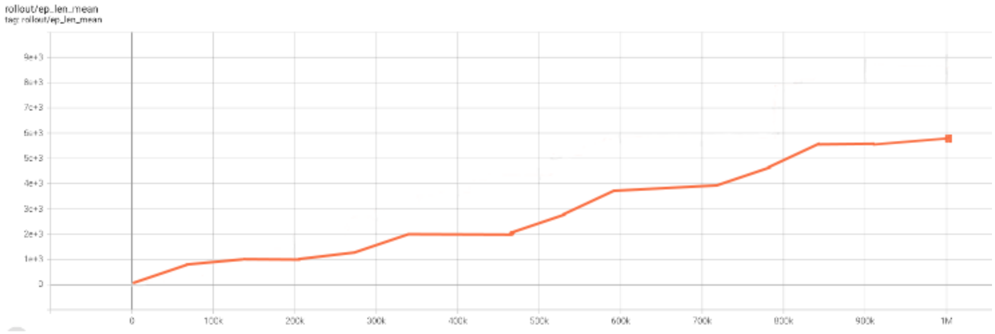


Figure 4.19: A2C average episode time

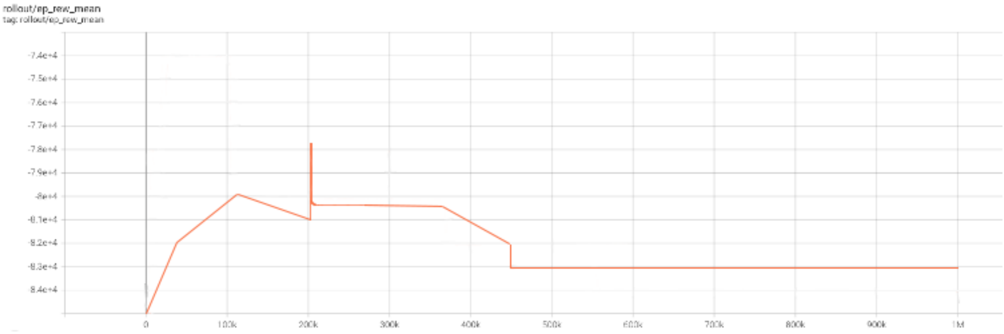


Figure 4.20: A2C average reward

4.6.3 DDPG

DDPG's results do not differ as much as A2C and PPO. DDPG's average episode time was highest during iterations 700K and 800K however it slightly decreased. The reward's performance is no different from the other agents. At the end the agent learns to minimise the negative reward as much as possible.

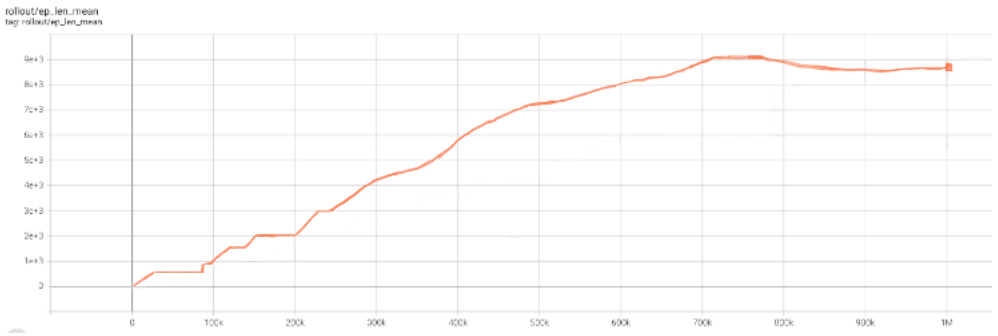


Figure 4.21: DDPG average episode time

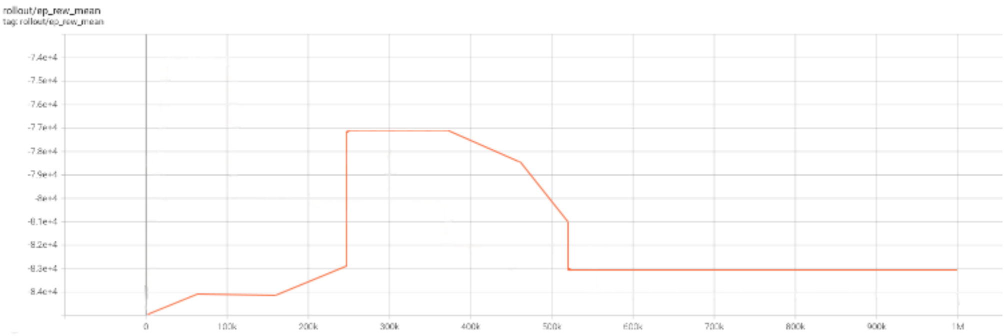


Figure 4.22: DDPG average reward

Chapter 5 - Conclusion

This chapter discusses the outcomes from the research, its limitations and the further work that can be done.

5.1 Discussion

A custom agent environment is built with a reward function based on the agent's balance, amount of bitcoin owned, current bitcoin price and most used technical indicators. A2C, PPO and DDPG are used to interact with the environment. Overall the agents did not lose any money when validating but also did not gain any money because the reward focuses on the price the agent has purchased one Bitcoin for and the validation data is in the period where the data continued falling. During training neither of the agents was able to make it to a positive reward. The reward. The agents successfully minimised the negative reward towards the end of training.

5.2 Limitations

The most important limitation is that the dataset consists only of Bitcoin prices, but other famous coins can be included as well – Ethereum, BNB, Solana, etc. The reward of the agent depends on the agent's starting data. The agent starts with 1 Bitcoin as if it has purchased it at that timestep. If the starting price is close to the highest amount in the dataset, then the agent will be only penalized if it does trades. This makes the environment somewhat unstable, and a new strategy is required. The agent's current volume and Bitcoin amount can be passed to the observation space as well.

5.3 Further work

The environment could also be further improved by using other risk measurement strategies such as Sharpe ratio, Sortino ratio and Profit and Loss. Further testing can also be done on the data's normalization and observation space. Instead of using one agent, hierarchical reinforcement learning can be used where an agent is responsible for solving a subproblem rather than the whole problem at once. For the solution to be more stable it, ensemble RL could be included. Using multiple agents to select a single action is more stable and less prone to losing money. Finally, research can be done on custom function approximations and even editing the basis of the RL algorithms, so they are more suited towards crypto currency trading than for more generic tasks.

5.4 Summary

This research is a build-up to a doctoral proposition related to reinforcement learning for crypto currency trading. The aims of this research are to investigate what crypto currency is, how it is used for crypto currency trading and to get familiar with the Python libraries.

There are a lot of misconceptions by the public about crypto currencies, how secure they are and their future. The main objective of them is to put the financial responsibilities and freedom in the user's hands rather than to banks. Crypto currencies' value grows with time whereas to fiat money's value is decreased.

The most used trading reinforcement learning algorithms were used on a highly customised environment. The agents were not able to get any positive rewards during training but learned how to minimise it. With the validation data, neither of them bought or sold Bitcoin.

Chapter 6 - References

[1] Robertson, H. (2020) *'Almost a fifth of ALL US dollars were created this year'*

Available at: https://www.cityam.com/almost-a-fifth-of-all-us-dollars-were-created-this-year/

(Accessed: 25th April 2022)

[2] Chen, J. (2022) *'Fiat Money'*

Available at: https://www.investopedia.com/terms/f/fiatmoney.asp

(Accessed: 25th April 2022)

[3] Binance (2022) *'Cryptocurrency and Inflation: Everything You Need to Know'*

Available at: https://www.binance.com/en/blog/fiat/cryptocurrency-and-inflation-everything-you-need-to-know-421499824684902682

(Accessed: 25th April 2022)

[4] Floyd, D. (2021) *'How Bitcoin Works'*

Available at: https://www.investopedia.com/news/how-bitcoin-works/

(Accessed: 25th April 2022)

[5] Philips, D., Chipolina, S. (2022) *'What backs Bitcoin?'*

Available at: https://decrypt.co/resources/what-is-bitcoin-backed-by

(Accessed: 25th April 2022)

[6] Hayes, A. (2022) *'Blockchain Facts: What Is It, How It Works, and How It Can Be Used'*

Available at: https://www.investopedia.com/terms/b/blockchain.asp

(Accessed: 25th April 2022)

[7] Chu, M. (2020) *'What is the Point of Cryptocurrency? (4 Reasons Why You Should Care)'*

Available at: https://dataoverhaulers.com/purpose-point-of-cryptocurrency/

(Accessed: 25th April 2022)

[8] Zafarm, T. (2021) *'Crypto vs Banking: Which Is a Better Choice?'*

Available at: https://www.entrepreneur.com/money-finance/crypto-vs-banking-which-is-a-better-choice/399503

(Accessed: 25th April 2022)

[9] Thompson, C. (2022) *'Federal Reserve's FedNow Real-Time Payments. Set for Mid-2023 Debut'*

Available at: https://uk.news.yahoo.com/federal-fednow-set-mid-2023-204948566.html

(Accessed: 25th April 2022)

[10] Chainalysis (2022) *'The 2022 Crypto Crime Report'*

Available at: https://go.chainalysis.com/rs/503-FAP-074/images/Crypto-Crime-Report-2022.pdf

(Accessed: 25th April 2022)

[11] Chainalysis (2020) *'PlusToken Scammers Didn't Just Steal $2+ Billion Worth of Cryptocurrency. They May Also Be Driving Down the Price of Bitcoin.'*

Available at: https://blog.chainalysis.com/reports/plustoken-scam-bitcoin-price/

(Accessed: 25th April 2022)

[12] Department of Justice (2021) *'Department of Justice Seizes $2.3 Million in Cryptocurrency Paid to the Ransomware Extortionists Darkside'*

Available at: https://www.justice.gov/opa/pr/department-justice-seizes-23-million-cryptocurrency-paid-ransomware-extortionists-darkside

(Accessed: 25th April 2022)

[13] Bellusci, M. (2021) *'IRS. Seized $3.5B in Cryptocurrency During Fiscal 2021'*

Available at: https://uk.finance.yahoo.com/news/irs-seized-3-5b-cryptocurrency-201544440.html

(Accessed: 25th April 2022)

[14] BBC (2021) *'Met Police seize record £180m of cryptocurrency in London'*

Available at: https://www.bbc.co.uk/news/uk-england-london-57816644

(Accessed: 25th April 2022)

[15] Sutton, R. and Barto, A. (2018) *'Reinforcement Learning: An Introduction',* MIT Press, pp. 25-26.

(Accessed: 25th April 2022)

[16] DeepMind (2016) *'Deep Reinforcement Learning'*

Available at: https://www.deepmind.com/blog/deep-reinforcement-learning

(Accessed: 25th April 2022)

[17] Ahn, Y., Kim, D. (2021) *'Emotional trading in the cryptocurrency market'* in Finance Research Letters Volume 42

Available at:

https://doi.org/10.1016/j.frl.2020.101912

[18] Thomas, M. (2019) *'How AI Trading Technology Is Making Stock Market Investors Smarter'*

Available at:

https://builtin.com/artificial-intelligence/ai-trading-stock-market-tech

(Accessed: 25th April 2022)

[19] Palamalai, S., Kumar., Krishna, Maity, B. (2021) '*Testing the random walk hypothesis for leading crypto currencies'* in *Borsa Istanbul Review,* Volume 21, Issue 3

Available at: https://doi.org/10.1016/j.bir.2020.10.006

(Accessed: 25th April 2022)

[20] Dupernex, S. (2007) *'Why might share prices follow a random walk?'* in Student Economic Review, Volume 21

Available at:
https://www.tcd.ie/Economics/assets/pdf/SER/2007/Samuel_Dupernex.pdf

(Accessed: 25th April 2022)

[21] Yu, L., Fung, H., Leung, W. (2019) *'Momentum or contrarian trading strategy: Which one works better in the Chinese stock market'* in International Review of Economics & Finance, Volume 62

Available at: https://doi.org/10.1016/j.iref.2019.03.006

(Accessed: 25th April 2022)

[22] Corbet, S., Eraslan, V., Lucey, B., Sensoy, A. (2019) *'The effectiveness of technical trading rules in cryptocurrency markets'* in Finance Research Letters, Volume 32

Available at: https://doi.org/10.1016/j.frl.2019.04.027

(Accessed: 25th April 2022)

[23] Grobys, K., Ahmed, S., Sapkota, N. (2020) *'Technical trading rules in the cryptocurrency market'* in Finance Research letters (2020), Volume 32

Available at: https://doi.org/10.1016/j.frl.2019.101396

(Accessed: 25th April 2022)

[24] Sattarov, O., Muminov, A., Lee, C., Kang, H., Oh, R., Ahn, J., Oh, H., Jeon, H. (2020) *'Recommending Cryptocurrency Trading Points with Deep Reinforcement Learning Approach'* in Computing and Artificial Intelligence

Available at: https://doi.org/10.3390/app10041506

(Accessed: 25th April 2022)

[25] Millea, A. (2021) *'Deep. Reinforcement Learning for Trading - A Critical Survey'* in Featured Reviews Of Data Science Research

Available. at: https://doi.org/10.3390/data6110119

(Accessed: 25th April 2022)

[26] Hernique, B., Sobrerio, V., Kimura., H. (2018) *'Stock price prediction using support vector regression on daily and up to the minute prices'* in The Journal of Finance and Data Science, Volume 4

Available at: https://doi.org/10.1016/j.jfds.2018.04.003

(Accessed: 25th April 2022)

[27] Vijh, M., Chandola, D., Tikkiwal, V., Kumar, A. (2020) *'Stock Closing Price Prediction using Machine Learning Techniques'* in Procedia Computer Science, Volume 167

Available at: https://doi.org/10.1016/j.procs.2020.03.326

(Accessed: 25th April 2022)

[28] K. Rathan, S. V. Sai and T. S. Manikanta, (2019) *'Crypto-Currency price prediction using Decision Tree and Regression techniques'*, 3rd International Conference on Trends in Electronics and Informatics (ICOEI), doi: 10.1109/ICOEI.2019.8862585.

(Accessed: 25th April 2022)

[29] Hu, Z., Zhao, Y., Khushi, M. (2021) *'A Suervey of Forex and Stock Price Prediction Using Deep Learning'* in Feature Paper Collection in Applied System Innovation

Available at: https://doi.org/10.3390/asi4010009

(Accessed: 25th April 2022)

[30] Wang, J., Sun, T., Cao, Y., Wang, D. (2021) *'Financial. Markets Prediction with Deep Learning'* in 2018 18th IEEE International Conference on Machine Learning and Applications (ICMLA)

Available at: https://doi.org/10.48550/arXiv.2104.05413

(Accessed: 25th April 2022)

[31] Vo, A., Nguyen, Q., Ock, C. (2019) *'Sentiment Analysis. Of. News. For Effective Cryptocurrency Price Prediction'* in International Journal of Knowledge Engineering, Volume 5

Available at: http://www.ijke.org/vol5/116-MK032.pdf

(Accessed: 25th April 2022)

[32] D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel and B. K. Lama, *'Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis'* 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), doi: 10.1109/CCCS.2018.8586824

(Accessed: 25th April 2022)

[33] Clements, W., Delft, B., Robaglia, B., Slaoui, R., Toth, S. (2019) *'Estimating Risk and Uncertainty in Deep Reinforcement Learning'* in Machine Learning (cs.LG)

Available at: https://doi.org/10.48550/arXiv.1905.09638

(Accessed: 25th April 2022)

[34] Zihao Z., Stefan Z., Roberts, S., (2019) 'DToeep Reinforcement Learning for Trading, Oxford-Man Institute of Quantitative Finance'

Available at:

https://www.oxford-man.ox.ac.uk/wp-content/uploads/2020/06/Deep-Reinforcement-Learning-for-Trading.pdf

(Accessed: 25th April 2022)

[35] Xoiong, Z., Liu, X., Zhong, S., Yang, H., Walid, A. (2018) *'Practical Deep Reinforcement Learning Approach for Stock Trading'* in Machine Learning (cs.LG)

Available at: https://doi.org/10.48550/arXiv.1811.07522

(Accessed: 25th April 2022)

[36] Li, J., Rao., R., Shi, J., (2018) *'Learning to Trade with Deep Actor Critic Methods'* in International Symposium on Computational Intelligence and Design (ISCID), doi: 10.1109/ISCID.2018.10116

(Accessed: 25th April 2022)

[37] Vishal, M., Satija, Y., Babu, B. (2021) *'Trading Agent for the Indian Stock Market scenario using Actor-Critic. Based Reinforcement Learning'* in IEEE International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS),

doi: 10.1109/CSITSS54238.2021.9683467

(Accessed: 25th April 2022)

[38] Sadighian, J. (2019) *'Deep Reinforcement Learning in Cryptocurrency Market Making'* in Trading and Market Microstructure

Available at: https://doi.org/10.48550/arXiv.1911.08647

(Accessed: 25th April 2022)

[39] Yang, H., Liu, X., Zhong, S., Walid, A. (2020) *'Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy'*

Available at: http://dx.doi.org/10.2139/ssrn.3690996

(Accessed: 30th June 2022)

[40] Sattarov, O., Muminov, Z., Lee, C., Kang, H., Oh, R., Ahn, J., Oh, H., Jeon, H. (2020) *'Recommending Cryptocurrency Trading Points with Deep. Reinforcement Learning Approach'*

Available at: https://doi.org/10.3390/app10041506

(Accessed: 30th June 2022)

[41] Jiang, Z., Xu, D., Liang, J. (2017) '*A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem*' in Computational Finance (q-fin.CP)

Available at: https://doi.org/10.48550/arXiv.1706.10059

(Accessed: 30th June 2022)

[42] Suri, K.; Saurav, S. Attentive Hierarchical Reinforcement Learning for Stock Order Executions. Available online:

https://doi.org/10.48550/arXiv.2104.00620

(Accessed: 9h July 2022)

[43] Gao, Y.; Gao, Z.; Hu, Y.; Song, S.; Jiang, Z.; Su, J. (2021) *'A Framework of Hierarchical Deep Q-Network for Portfolio Management'* in

Proceedings of the ICAART

Available at: https://www.scitepress.org/Papers/2021/102332/102332.pdf

(Accessed: 9h July 2022)

[44] Suri, K.; Shi, X.Q.; Plataniotis, K.; Lawryshyn, Y. (2021) *'TradeR Practical Deep Hierarchical Reinforcement Learning for Trade Execution'* in Trading and Market Microstructure (q-fin.TR)

Available at: https://doi.org/10.48550/arXiv.2104.00620

(Accessed: 9h July 2022)

[45] Li, L. (2021) *'Financial Trading with Feature Preprocessing and Recurrent Reinforcement Learning'*

Available at: https://doi.org/10.48550/arXiv.2109.05283

(Accessed: 13th August 2022)

[46] Scott, G. (2020) *'Random Walk Theory'*

Available at:

https://www.investopedia.com/terms/r/randomwalktheory.asp

(Accessed: 13th August 2022)