IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 January 2025, accepted 22 February 2025, date of publication 26 February 2025, date of current version 13 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3546099

RESEARCH ARTICLE

A Novel RMS-Driven Deep Reinforcement Learning for Optimized Portfolio Management in Stock Trading

ASMA SATTAR[1], AMNA SARWAR[2], SAIRA GILLANI[3], MARYAM BUKHARI[4], SEUNGMIN RHO[5], AND MUHAMMAD FASEEH[6]

[1]College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan
[3]Department of Information Technology and Computer Science, University of Central Punjab, Lahore 54000, Pakistan
[4]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[5]Department of Industrial Security, Chung-Ang University, Seoul 06974, South Korea
[6]Department of Electronic Engineering, Jeju National University, Jeju-si 63243, Republic of Korea

Corresponding authors: Seungmin Rho (smrho@cau.ac.kr) and Maryam Bukhari (maryambukhari09@gmail.com)

This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2025-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT Algorithmic stock trading has improved tremendously, with Reinforcement Learning (RL) algorithms being more adaptable than classic approaches like mean reversion and momentum. However, challenges remain in adequately depicting market events and generating suitable rewards to influence the trading decisions of an agent in a dynamic environment. This study proposed an improved stock market trading framework termed the RMS-Driven Deep Reinforcement Learning (DRL) model for optimal portfolio management. The research attempts to give a more comprehensive view of the market and, as a result, enhance trading decisions by including consumer information and incorporating news sentiments into the model, in addition to data from typical earnings reports. More specifically, three kinds of DRL models are presented, combined with data from stock earnings reports, Max Drawdown rewards, and sentiment indicators (RMS), known as PPO_RMS, A2C_RMS, and DDPG_RMS, respectively. The findings of this research indicate that the integrated model, mainly the DDPG_RMS effectively outperforms the baseline ^DJI index in many risk-return analyses in ratios showing better risk management, and profitability. The proposed stock trading model generates a maximum cumulative return of 27% with a Sharpe ratio of 0.66, showing an appropriate trade-off between risk and return. This approach, which incorporates sentiment analysis and Max Drawdown rewards, significantly enhances the model's performance in adapting to changing market conditions. Therefore, the results emphasize the appropriateness of integrating sentiment indices with traditional financial data to enhance a trader's performance while also offering essential information to aid in the development of improved trading tactics in continuously changing financial markets.

INDEX TERMS Deep reinforcement learning (DRL), stock market trading, portfolio management, max drawdown rewards, advantage actor-critic (A2C), deep deterministic policy gradient (DDPG), proximal policy optimization (PPO).

I. INTRODUCTION

Stock trading is the act of buying and selling a company's shares on the stock market. The primary goal of frequent purchases and sales is to maximize capital returns by exploiting the advantages of market volatility [1]. However, an

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Ni.

organization's financial condition, internal and external situations, trends in the global market, and other relevant factors all have an impact on stock prices, thus affecting stock trade decisions [2]. Stock market trading involves the use of analysis, knowledge, and management of risks in an effectively conducted market. There are several methods through which traders can assess the performance of stocks with the main ones being fundamental analysis and technical analysis [3]. Fundamental analysis involves checking the general operational capability of a specific firm, its position in the industry as well as the overall economic conditions to establish the value of the security. On the other hand, technical analysis involves the analysis of the price and volume data with the view of trying to establish some indications of future prices [4]. These analyses thus assist traders on how to procedures for purchasing at the cheap and selling at the expensive. The traders need to be more careful and flexible to look at market parameters day and night and should tweak their policies [5]. The primary objective that any investor seeks to achieve their capacity within the financial market is to maximize returns while minimizing risks. To achieve this objective, it is required to estimate the correct prices or value change rates for various market assets and correctly distribute investment among the chosen ones [6]. This procedure is challenging for humans since numerous aspects must be examined within a particular context and are always changing. Existing research shows that human trader's decision-making is typically influenced by emotions, resulting in lesser revenues than estimated [7]. Stock prices fluctuate often, which makes it challenging for human traders to adapt appropriately. Thus, the focus of the research has been to develop optimal adaptive automated trading systems that address investor's requirements and provide additional active capital to the global market.

In the existing studies, several trading techniques have been designed but they are unable to perform and yield good results over different market conditions [8]. Since the early 1990s, researchers have used artificial intelligence (AI) to analyze investments in the stock market [9]. From the context of AI, the concept of Algorithmic Trading (AT) is used by researchers to improve these strategies. Depending upon instructions and rules defined by the programmer, the AT is simply a computer program. It is observed that in comparison with human traders, these automated computer programs take less time to assist with trading decisions. Several traditional techniques include mean reversion [10], momentum-based methods [11], and rule-discovery techniques [12]. Moreover, numerous attempts have been made towards the development of such systems with the use of supervised learning methodologies dominating. These strategies make models to learn predicted market patterns using predictive models such as neural networks and random forests [13]. Moreover, most of the methods that are applied under the assumption of using supervised learning principally consider reduced prediction error risk as a target while leaving aside the threats connected with it e.g., disparities among the forecasted prices and trading decisions [14]. They sometimes fail to capture the external conditions of the marketplace, for instance, lack of cash and transaction costs which, in turn, narrows their applicability. Therefore, although SL techniques have been implemented extensively, such limitations as the inability to solve both the prediction and capital allocation problems at the same time, as well as the neglect of the constraints of the market environment, have become apparent [15]. Many forecasting systems estimate future stock prices and create trading rules [16]. Furthermore, translating predicted signals to trading positions is not straightforward; for example, forecasting horizons typically range from one to several days, depending on daily data [17].

To handle risk management and portfolios, sophisticated technologies like Reinforcement Learning (RL) are utilized in comparison with supervised learning. RL is a subfield of machine learning where an agent has to observe the feedback it gets from the environment in which it performs actions to achieve a maximum sum of rewards. While in supervised learning the training data is labeled, in RL the agent is active and is free to roam within the environment and learns through 'reinforcement' which could be in terms of rewards for the correct action or penalties for the wrong action [18]. In the context of Quantitative Trading (QT), firstly, supervised learning does not suffice for the issues that relate to long-term and delayed time rewards including trades in the financial markets [19]. In trading, future decisions are made based on the rewards that the trading agent gains in a certain period. Based on the findings of the research, it can be stated that the RL approach is more suitable for solving decision-making problems in the uncertain context of the financial markets [20]. Utilizing RL for algorithmic trading requires a continuous perception of the stock market in order to make optimal trading decisions. These algorithms can self-learn and improve their policies over time, making them appropriate for the adaptable environment of stock trading. Moreover, RL works to overcome the problems that are associated with the SL approaches when trading financial markets by combining the two steps, which are the price "prediction" step of the financial assets and the "allocation" step of the portfolio to maximize the investor's goal [21]. The stock trading agent is involved in the process of interacting with the environment which in this case is the model to make appropriate decisions. Financial data are usually highly time-dependent, which makes the Markov Decision Process (MDP), which is a primary object in solving RL problems, an ideal fit.

Therefore, in recent studies, several researchers have designed stock trading strategies with the use of RL algorithms. However, the major challenges in framing problems with RL techniques are the state representation from the environment and suitable rewards [1]. Since these elements are the major criteria to improve the performance of RL agents, for instance, a holistic view of the stock market in

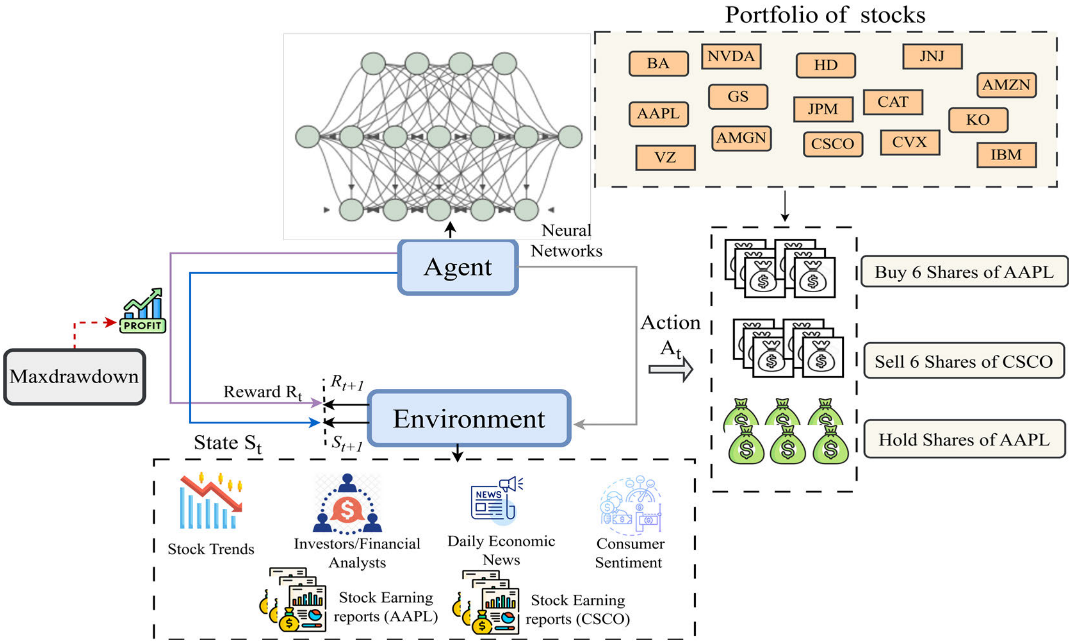


FIGURE 1. Generic overview reinforcement learning framework for stock market trading.

state representation makes the agent more aware of different environmental variables. To address such challenges, some studies employed daily historical records, [22] and additional variables such as news sentiments [23], as well as data fusion methods such as a combination of candlestick charts and technical indicators [24] or combination of sentiments and macro-economic factors [25]. The key challenge for all of these algorithms is exact comprehension of the stock environment; the better informed an agent is regarding the stock market, the better choices it can draw.

Although, several efforts have been made to enrich data representation, however, there are still some research gaps or external elements that need to be modeled in the stock trading agent. Likewise, different suitable trading rewards for RL have also been designed, but a reward function that ensures a tradeoff between both profitability and low risks and financial losses is a major research gap to be fulfilled. Hence, in this research study, the RMS variables namely data from earning reports as well as sentiment indices containing daily news sentiments and consumer sentiments in addition to Maxdrawdown-based rewards have been modeled. In addition to daily historical data, the proposed framework also models a variety of technical indicators. The agent can be trained by use of Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO) algorithms are used in addition to proposed RMS factors referred to as A2C_RMS, DDPG_RMS, and PPO_RMS. Altogether, these algorithms guarantee potential sustainable future benefits, together with potential short-term profits, which allows the development of strategies in the form of long-term sustainable plans. The benefits of including RMS factors include, for instance, consumer sentiment is of much importance, especially to stock traders since it has an impact on spending, which has an impact on the revenues of corporations and economic growth [26]. Likewise, earning reports also contain projections or outlooks at the company's future operations and possible difficulties that may be encountered. Similarly, the proposed variations of RL agents that use Max Drawdown as a reward function offer a substantial benefit since the proposed reward is combined with risk management. By combining these RMS factors with the advantages of DDPG, A2C, and PPO, these algorithms prioritize long-term gains while still achieving potential short-term profits, allowing for the development of strategies that guarantee sustainability in the long run. This study aims to increase the effectiveness of trading systems and decision-making in the stock market. The main contributions of this study are as follows:

- This study proposed three variants of RL agents namely A2C, DDPG, and PPO with RMS factors modeling
- The proposed variants integrate both consumer sentiment as well as news sentiment indexes into its modeling approach, providing a more complete view of the stock market in addition to technical indicators and historical data
- Data from stock earnings releases is also exploited in the modeling process to create more comprehensive representations of states in RL
- A novel Max drawdown-based reward function is also proposed guaranteeing that the trading strategy not only seeks high profitability but also mitigates risks

The remainder of the article has been divided into different sections: Section II presents related work, Section III highlights the proposed work, and Section IV gives results and discussion Section V presents comparative analysis, limitations, future work, and conclusion. Figure 1 shows the generic overview of stock market trading using RL.

II. LITERATURE REVIEW

Financial markets are among the captivating advancements of recent years that have impacted several sectors including commerce, education, employment, technology, and the overall economy [27]. The inherent task of detecting trends and prices in the stock markets is very complex mainly because of the dynamic, non-linear, non-stationary, nonparametric, noisy, and chaotic nature of the markets. Mainly, AI-based systems for stock analysis involve stock price forecasting models, stock movement predictions, stock trading systems or portfolio optimization techniques [1], [28], [29]. From all these, the task of stock algorithmic trading was initially approached and developed by economists and mathematicians without the use of AI. In this regard, several methods have been developed such as mean reversion [10], momentum-based methods [11], and rule-discovery techniques [12]. However, in such methods, expert knowledge required from the domain of finance is crucial to determine underlying trends in the financial sector. Similarly, some TTR (Technical Trading rule) reliant techniques have also employed. For instance, five TTR rules have been designed such as moving average, relative strength index, momentum, etc., and findings have been reported on the Turkish stock market [30]. Likewise, some researchers employ fuzzy-rules-based techniques to predict trading signals [31], [32]. A hybrid strategy for stock trading that combines rough sets as well as genetic algorithms is also proposed in [33]. Data from the Korea Composite Stock Market Index 200 (KOSPI 200) is employed to perform experimentation to report the findings. These rules-based methods although exhibit good performance but their effectiveness is limited to generalize across a variety of situations of stock markets.

Following the aforementioned traditional trading methods, several research has used methodologies centered around supervised machine learning and deep learning algorithms. These are algorithms that are trained on labeled datasets, that is the feature or independent variables and their corresponding dependent variables and demonstrate good results not only in finance-related applications but also in other use-cases [34], [35]. More explicitly, several kinds of strategies based on stock price forecasting and stock trend predictions have been designed. For example, in [36] and [37], the stock price prediction model is designed using LSTM (Long Short-Term Memory Networks). This model also fused up additional information in the form of technical indicators including MFI, relative RSI, MACD, etc. In the next step, traders utilize an investment success score calculated from the prediction outcomes of the model in making trading decisions. Similarly, a hybrid approach of CNN and LSTM has been proposed [37], [38]. In this, the time series has been framed to predict future prices with 20 days ahead. By employing a weighted multicategory generalized eigenvalue support vector machine (WMGEPSVM) model is proposed in which the input of the algorithm is the stock's historical price data and technical indicators [38]. Their proposed model exhibits a minimal MDD value in contrast to baseline techniques.

Furthermore, RL approaches are one of the latest and third category of ML algorithms for stock market trading. For instance, two improved trading strategies that are designed to work for states and actions of financial markets: GDQN (Gated Deep Q-learning) and GDPG (Gated Deterministic Policy Gradient) have been proposed [8]. These strategies tie in a new hedonic reward function that fundamentally should provide steady returns, including in a situation where market conditions are fluctuating significantly. The analysis of the signals using techniques from mathematics and computer science shows that the models of GDQN and GDPG are more effective than the Turtle trading strategy having higher predicted profitability and more stable return rates than the current leaders based on the DRL mechanisms. Similarly, a new trading system based on the improved Deep SARSA for algorithmic trading in volatile markets is proposed [39]. This study improves the performance stability and convergence of the SARSA algorithm which is of special importance for risk control and daily trading portfolio management. A novel approach namely DRL-UTrans that is a deep reinforcement learning with the transformer and U-net for the stock trading is proposed in [40]. The transformer layers in this model learn about dynamic patterns of the market, whereas U-net architecture integrates both long and short features through skip connections. Similarly, in [41], the authors propose a new trading approach based on the DRL procedure through the application of the MADDQN mechanism. The research also highlights the drawbacks of single-agent models in simulating intricate financial relations and suggests the use of a multi-agent system to simulate various trading patterns. An attention mechanism with the RL approach is also designed for stock market trading to boost up the results followed by validating results on stocks of S&P500 [42]. Their proposed method exhibits good results in terms of the Sharpe ratio. Likewise, to acquire more abstract representations and information from stock prices, a module of Cascaded Long Short-Term Memory (CLSTM-PPO Model) is also incorporated into DRL for robust feature learning [43]. The input of this method is the fusion of daily historical records as well as common technical indicators. Subsequently, to combine conventional techniques with the latest methods, a hybrid model integrating concepts of rule-based expert systems along with the DRL model is proposed in [44]. This rule-based method makes DRL agents to learn more good trading opportunities and demonstrates superior performance.

A number of studies have been conducted on how DRL can be applied to stock market trading, but some research gaps exist. Also, most of the current literature focuses on historical prices and technical factors but does not consider sentiment data, especially consumer sentiments and earning reports

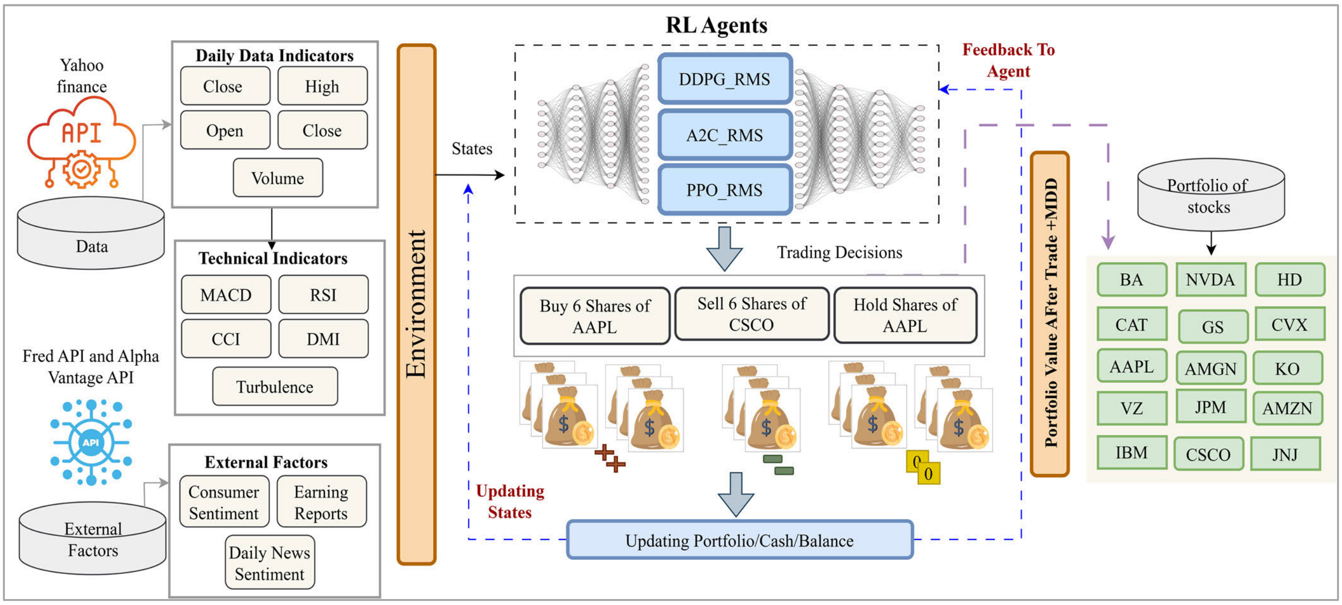


FIGURE 2. A pictorial representation of proposed RL framework for optimizing portfolios to assist investors.

of stocks or companies. This oversight means that trading strategies cannot be easily adjusted in response to changing conditions, especially in unpredictable environments. Furthermore, some models receive high theoretical scores, but when employing them in live accounts, significant issues are revealed, which makes the investigation of the practical applicability of the models important. Therefore, the research intended to address certain challenges, including evaluating the performance of DDPG, A2C, and PPO algorithms in different market environments.

To integrate dynamic elements into the model, aspects like stock price information are incorporated in our methodology, as consumer sentiment and daily new sentiments, in addition, data from earning reports and Max drawdown is identified as having a major influence on market direction that could not have been made as part of a static model. This will help us to make better models that would better suit both historical conditions and present tendencies in the market, thus improving their efficiency. Last but not least the research aspect of the study is conducted with adequate back testing and cross-validation for various market situations to assess the efficiency of the proposed strategies.

III. PROPOSED METHODOLOGY

In this section, a detailed description of the functionality of the proposed RL agent is provided step by step. A conceptual framework that illustrates the proposed work is illustrated in Figure 2.

A. BACKGROUND OF RL

Reinforcement learning (RL), which is a subfield of machine learning, has strong learning agents that aim to acquire as much cumulative reward as possible from an environment. For instance, In 2015, Alpha Go proved to be much superior to the Human professional players [45]. As for RL, it provides for effective learning in terms of objectives set along with rational choice-making. In other words, RL is a process in which agents adjust their policies through trial and error in an environment. The environment is generally described as the Markov Decision Process (MDP), defined as S, A, T, R and γ where S stands for the state, A is the actions, T for state transition, R for award and γ as the discount factor. The return refers to the aggregate of the number of future rewards over the discount factor γ whereby $\gamma \in (0, 1]$. The agent improves actions that produce favorable outcomes and avoids those that result in undesired outcomes through the process of trial and error with a view to achieving the best outcomes. In RL, the common classification of methods involves value-based, policy-based and the actor-critic. While value-based techniques involve estimating the value of actions, policy-based techniques directly aim for the optimization of policies. Actor-critic methods are a combination of value-based and policy-based methods. In RL, to solve any problem, it is mandatory to define states, rewards, actions, and environment according to the design and for achieving certain objectives. In algorithmic trading, states are not directly given but are derived from a series of observations. To handle this problem, the model is extended and an observation probability $P(ob|s, a)$ is incorporated to make what we know as the Partially Observable MDP model.

B. MARKOV DECISION PROCESS (MDP) FORMULATION

The Markov Decision Process (MDP) is a model of stochastic processes and random variables that dictate the transition of state to state with regard to certain assumptions and probabilistic laws. It is necessary to define RL and MDPs as perfect for it since the latter is mathematically well grounded. By so doing, the agent corresponds to the decision maker or learner while the environment is the world the agent exists in. In the particular time step $t \in \{1, 2, 3, \ldots, T\}$, the learning agent comes across the environment. In this work, the MDP framework has been used in stock market trading to emulate a real-life trading environment.

1) STATES MODELING

The ability to depict the state of the environment is critical to agents' learning of the best policies. In the stock market, the agent's environment is determined by the current stock market environment. The choice of the right inputs is critical in helping traders gain insight into the market as well as formulating trading rules. Similar to the trading decisions where investors may consider more factors, the proposed RL agent also includes multiple factors within its observation space.

The states of the proposed RL agent involve the daily historical stock information including the opening price (op_t), the highest price within the day (hi_t), the lowest price within the day (lo_t) and closing price (cl_t) as well as the trade volumes (tv_t) all of which are denoted as $op_t, hi_t, lo_t, cl_t, tv_t \in \mathbb{R}^n_+$. In particular, op_t stands for the initial stock's price in the certain time, hi_t and lo_t define the maximum and minimum price in the time t, cl_t is the price before the market closes, while tv_t is the total number of shares traded within the particular session. Also, using 't' to refer to the time step, $b_t \in \mathbb{R}^+$ represents the balance at any time step and the available number of shares for each stock in the portfolio is represented by the vector $h_t \in \mathbb{Z}^n_+$. In addition to these, some technical indicators (described below including MACD, CCI, DMI, RSI, and Turbulence), and data from earning reports, consumer sentiment indices, and daily news sentiment (described below) have also been modeled into the states of RL. All these indicators are combined through concatenation along with time steps of days and provided as input to the states of all three variants of proposed DRL models namely A2C_RMS, DDPG_RMS, and PPO_RMS respectively. Figure 3 shows the training and testing setup.

a: TECHNICAL INDICATORS

When making trading decisions, investors frequently use technical factors T_i, which are heuristic and mathematical computations based on past stock data. Some of the technical indicators that this study computed are as follows:

1. Moving Average Convergence Divergence (MACD)

In technical analysis, it is one of the most used indicators enabling to use of changes in intensity, direction, and duration of stock price patterns [46]. It is sometimes referred to as a "moving averages indicator" or commonly known as a "momentum indicator". We have determined the MACD for every stock in a portfolio represented as $M_p \in \mathbb{R}^t_+$ where t is the total number of stocks. It can be computed using an equation (1):

$$MACD = N.Period\_EMA - MPeriod\_EMA \tag{1}$$

In the above equation (1), the MACD is determined by subtracting the long-term EMA i.e.N from that of the short-term EMA, M. An EMA is a moving average (MA) i.e. that gives greater importance and relevance to more recent data values.

2. Relative Strength Index (RSI)

Another technical tool that investors use in the financial markets is the Relative Strength Index commonly referred to as RSI [46]. In recap, it is designed to show the past and present trends or health of a stock or market based on the last traded prices of the current trading session. This also indicates how much the current prices have recently changed. Thus, if the variations in price occur within the support line making a pattern with RSI, then it means that the stock has been oversold. From this, it means that humans are able to make judgement as to purchases. Likewise, if the price moves around the resistance level, it means it is overbought and a sale ought to be made. We have determined the RSI for every stock in a portfolio represented as $R_p \in \mathbb{R}^t_+$ where t is the total number of stocks. For all collections of stocks in a portfolio represented as $R_p \in \mathbb{R}^t_+$, initially, average gain and loss is computed:

$$RSI = 100 - \left[\frac{100}{1 + \frac{Mean\ Gain}{Mean\ Loss}}\right] \tag{2}$$

In the above equation (2), the formula uses the mean percentage gain or loss throughout a given look-back time.

3. Commodity Channel Index (CCI)

This measure is derived from functions that involve high, low, and closing prices and moves with an oscillator that incorporates momentum [47]. This measure illustrates the current price in relation to the mean of the prices of a given window size of the purchasing and selling options. It also means that when this measure is zero then the current price is higher than the mean of the previous prices. We calculated CCI for each stock in a portfolio, represented as $C_p \in \mathbb{R}^t_+$, where t represents the number of stocks higher than the mean of the previous prices. We calculated CCI for each stock in a portfolio, represented as $C_p \in \mathbb{R}^t_+$, where t represents the number of stocks.

4. Directional Movement Index (DMI)

This measure is employed to determine the trend of an asset price. It may be estimated using the high, low, and closing prices and describes the direction as well as the intensity of trading prices [48]. These are the directional movement lines (−DI) and (+DI), the average directional index (ADX), the directional index (DX), and the EMA of ADX, among others. In this research, DX has been calculated for every stock in a portfolio represented as $C_p \in \mathbb{R}^t_+$, where t is the total number of stocks. The following equations(3-5) are showing

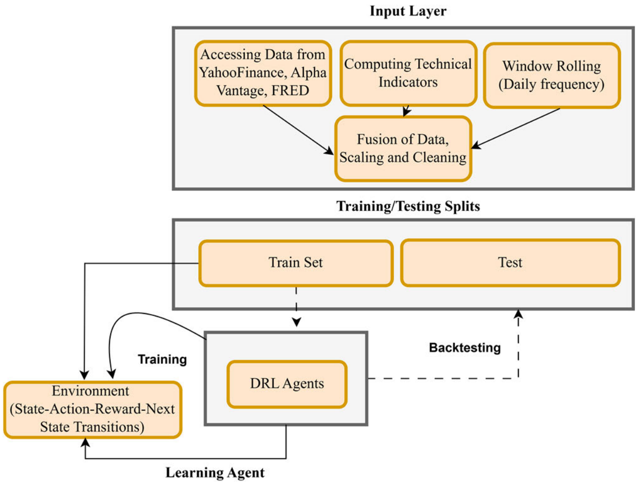

Input Layer
Accessing Data from YahooFinance, Alpha Vantage, FRED
Computing Technical Indicators
Window Rolling (Daily frequency)
Fusion of Data, Scaling and Cleaning
Training/Testing Splits
Train Set
Test
Training
Backtesting
DRL Agents
Environment (State-Action-Reward-Next State Transitions)
Learning Agent

FIGURE 3. A pictorial representation of training and testing setup.

the mathematical formulation of DX computation:

$$+DI = \left[\frac{\sum_{i=1}^{n} DM - (\sum_{I=1}^{N} \frac{DM}{14}) + Current\ DM}{ATR} \right] \times 100 \tag{3}$$

$$-DI = \left[\frac{\sum_{i=1}^{n} DM - (\sum_{I=1}^{N} \frac{DM}{14}) - Current\ DM}{ATR} \right] \times 100 \tag{4}$$

$$DX = \left[\frac{|+DI - -DI|}{|+DI \pm DI|} \right] \times 100 \tag{5}$$

5. Turbulence

This measure is referred to as periods characterized by high volatility, unpredictability, and conflicts in the financial markets.

In this research, we have calculated the turbulence for every stock in a portfolio represented as $T_p \in \mathbb{R}^t_+$, where t is the total number of stocks. Mathematically, it is defined as

$$d_t = (y_t - \mu) \sum^{-1} (y_t - \mu)' \tag{6}$$

In the above equation (6), d_t is the turbulence over a specific time period t, where y_t symbolizes the vector of asset returns, μ shows the mean value of historical returns, whereas shows the covariance matrix of returns.

b: SENTIMENT INDICES

Sentiment Indices are indicators that quantify the sentiment of news and other textual information associated with financial markets [49]. These indices are frequently extracted from articles or posts from news sources or social networks that shape people's perceptions of the market.

1. Daily news Sentiment

Daily news sentiment is very useful for trading because it gives an opportunity to invest based on the current global mood of the market. Markets are not only determined by the supply and demand factors but also depend on the sentiments of the investor, which are in most cases, swayed by the news. When the news sentiment is positive the investor gains confidence to invest, and the volume of buying increases the stock prices of the share [50]. On the other hand, negative sentiment often leads to a selloff due to investors wanting to reduce the amount of money they are likely to lose. In this research, we have collected the daily news sentiment to combine with daily stock data as an additional indicator as follows:

$$X_t = concat\left[(op_t, hi_t, lo_t, cl_t, tv_t \in R^n_+), (T_i \in R^n_+), (S_{news} \in R^n_+)\right] \tag{7}$$

In the above equation (7), (S_{news}) feature is fused with daily historical stock as well as technical indicators. The daily news sentiment is extracted from the Federal Reserve Bank of San Francisco [51].

2. Consumer Sentiment

Consumer sentiment is defined as the outlook that consumers have about the current and future state of the economy with personal financial prospects. It is mainly assessed by the consumer expectations concerning future changes, their expenditure pattern, and their view on economic conditions. High consumer sentiment also gives an indication that the consumers are optimistic about the economic status thus they will spend more on aspects that assist in economic growth. On the other hand, low consumer confidence can be attributed to a consumer being conservative which translates into less spending and possibly slower economic growth. Consumer sentiment is of much importance, especially to stock traders since it has an impact on spending, which has an impact on the revenues of corporations and economic growth [26]. In this research, we have collected consumer sentiment to combine with daily, technical, as well as news sentiment. The above equation is revised to include the consumer sentiment:

$$X_t = concat\left[\left(op_t, hi_t, lo_t, cl_t, tv_t \in R^n_+\right), (T_i \in R^n_+), \left(S_{news} \in R^n_+\right)\left(\left(S_{consumer} \in R^n_+\right)\right]\right. \tag{8}$$

In the above equation (8), $(S_{consumer})$ feature is fused with daily historical stock, technical indicators, as well as news sentiment. This data of consumer sentiment is extracted through FredAPI [52].

3. Stock Earning Reports

Stock earning reports are reports that are issued by most companies as they adopt the public company model of operation and contain the balance sheets of the company for a given period of time. Common ones are the revenues and net income, earnings per share, and the operating expenses to mention but a few. They contain a summary of a company's financial performance and are of vital interest to investors, analysts, and traders. Earning reports also contain projections or outlooks for the company's future operations and possible difficulties that may be encountered [53]. Earning reports are particularly significant in stock trading because they offer first-hand information on how a company is performing in terms of stock. In this research, we have collected the consumer sentiment to combine with daily, technical, as well as news sentiment. The above equation is revised to include the consumer sentiment:

$$X_t = concat\left[\left(op_t, hi_t, lo_t, cl_t, tv_t \in R_+^n\right), (T_i \in R_+^n\right), \left(S_{news} \in R_+^n\right)(\left(S_{consumer} \in R_+^n\right)\left(E_{Reports} \in R_+^n\right)\right] \tag{9}$$

In the above equation (9), $(E_{Reports})$ feature is fused with daily historical stock, technical indicators, daily news, and consumer sentiment to generate a combined feature representation denotes as X_t. The data of earning reports of stocks is extracted from AlphaVantage API [54].

2) TRADING ACTIONS

The behavior that is exhibited by the agent when interacting with the trading environment can be expressed in action. In this, the agent sees the fundamental, technical, and the daily data of the stock market then decides to either buy, sell or hold. These actions $a \in A$ are represented as $a \in \{-1, 0, 1\}$, where 0 represents purchase, 1 shows hold, and -1 shows sell. Since the actions are taken over several shares, the action space in this case is $\{-q, \ldots, -1, 0, 1, \ldots, q\}$, where symbol q denotes the total shares. For this study, q equals 100.

1. Proposed MDD Reward

A reward in the case of stock trading most often means the yield or profit on capital. This desirable outcome is actively achieved by traders due to the intentional acquirement and divestiture of equities. Since the principal motivation of the traders is to ensure that they earn higher profits with minimal risks, the concept of reward is very relevant to the trader. In this study, we have employed the max drawdown as an additional variable with portfolio values in the reward function. The max drawdown is the single biggest loss that a portfolio can endure from, a high point to a low point and then back to a high before a new low point. It measures the maximum percentage loss from the highest value of an investment and gives the investor a worst-case view of their portfolio [55]. The max drawdown is defined below in equation (10):

$$MDD = \frac{Trough\ value - Peak\ value}{Peak\ value} \tag{10}$$

In stock trading, max drawdown is very handy since it lets a trader measure the risk that is attached to it in various investments and strategies. Possible drawdown can be useful for traders to have some understanding to improve their portfolio, that is to avoid risking more than is needed.

4. State Transitions

The RL agent chooses an action from the policy given its current state and it executes this action in the market (e.g., buying or selling a specified quantity of asset). For instance, before the agent purchases 100 shares in a company the financial status of the company as well as the agent's portfolio is altered through a market action. Consequently, after the action, the RL states for the next day include shares, dollar amount of shares, balance, and trading shares. That is, if the agent invests in stocks, the current balance is reduced by the amount invested to make such an acquisition, while the dollar value of stocks owned in the given asset increases. After trading shares, and rebalancing the portfolio, the RL agent goes to the next state in the state space. The state also enables the agent to have a real-time view of the environment and one's position in it once a particular activity has been accomplished.

C. DRL STOCK AGENTS

This study employs three reinforcement learning models namely: Advantage Actor Critic (A2C), Deep Deterministic Policy Gradient (DDPG), and Proximal Policy Gradients (PPO). The weights of these DRL agents involving neural networks are optimized through Adam optimizer. The states are comprised of technical indicators, sentiment indices, consumer indices and daily indicators and these features are provided as input to DRL-agents. Figure 4 shows the detail internal working of architecture and following is a detailed description of them:

1. Deep Deterministic Policy Gradient (DDPG)

Deep Deterministic Policy Gradient (DDPG) is used to optimize investment returns. DDPG is developed from the Deterministic Policy Gradient (DPG) algorithm [56], which is a combination of Q-learning [57] and policy gradient framework [58]. However, DDPG applies neural networks as the function approximators. The DDPG algorithm is intended for the Markov Decision Process (MDP) model of the stock trading market. Q-learning is fundamentally a method for learning about the environment. Instead of updating $Q\left(s_y, a_y\right)$ using the expected value of $(s_{y+1}a_{y+1})$, Q-learning employs a greedy action a_{y+1} that maximizes $Q(s_{y+1}a_{y+1})$ for the next state s_{y+1} i.e.,

$$Q_\pi\left(s_y, a_y\right) = \mathbb{E}_{s_{y+1}}[\omega(s_y, a_y, s_{y+1}) + \gamma_{a_{y+1}}^{max} Q(s_{y+1}a_{y+1})] \tag{11}$$

Equation (11) represents the Q-function in reinforcement learning is the expected return of taking an action a_y in state s_y under policy π. It is calculated by taking the expected value of the immediate reward $\omega(s_y, a_y, s_{y+1})+$ plus the discounted maximum future reward, where γ is the discount factor and $\gamma_{a_{y+1}}^{max} Q\left(s_{y+1}a_{y+1}\right)]$ represents the highest expected return from the next state s_{y+1} considering all possible future actions a_{y+1}. As for the Deep Q-network (DQN) architecture, functions are approximated using neural networks, and states are encoded into value functions. However, DQN cannot be applied to this problem because of the scale of action space and how it has been reduced from the previous problem. Several trading actions are feasible for each stock, and

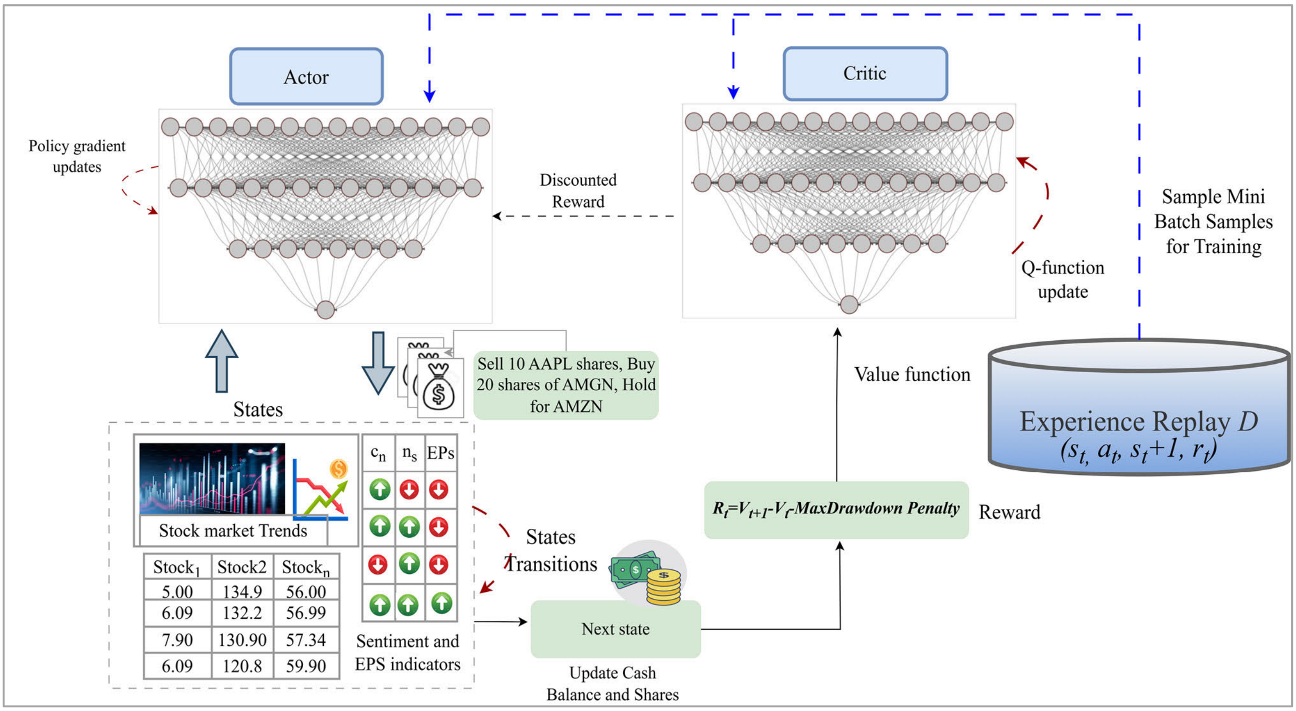


FIGURE 4. Detailed internal working of rl framework for stock market trading.

with respect to the total number of stocks, the sizes of the action spaces increase exponentially which leads to a curse of dimensionality [15]. To deal with this problem, there needs to be an algorithm that can map states to action deterministically, and this is where the DDPG algorithm comes in.

The target networks Q' and σ' are created by duplicating the actor and critic networks, ensuring stable temporal difference updates. These networks are updated iteratively. At each timestep, the DDPG agent performs an action a_y based on the state s_y, and subsequently receives a reward corresponding to the new state s_{y+1}. The transition $(s_y, a_y, s_{y+1}\gamma_y)$ is stored in the replay buffer $\mathcal{R}$. A minibatch of $\mathbb{N}$ sample transitions are then drawn from $\mathcal{R}$ and $y_i = r_i + \gamma Q'(s_{y+1}, \sigma'(s_{i+1}|\theta^{\sigma'})|\,\theta^{Q'})$ is computed for $i = 1, \ldots .N$. The critic network is subsequently updated by minimizing the expected difference $L(\theta^Q)$ between the outputs of the target critic network Q' and the critic network Q, specifically:

$$L\left(\theta^Q\right) = E_{s_y,a_y,\gamma_y,s_{y+1}\sim buffer}[(\gamma_y + \gamma Q'(s_{y+1}, \sigma(s_{y+1}|\theta^{\sigma})|\theta^{\sigma'}) - Q(s_y, a_y|\theta^Q))^2 \tag{12}$$

The parameters θ^σ of the actor network are subsequently defined as follows:

$$\begin{aligned}\nabla_{\theta^\sigma} J &\approx E_{s_y,a_y,\gamma_y,s_{y+1}\sim buffer}[(\nabla_{\theta^\sigma} Q(s_y, \sigma(s_y|\theta^\sigma)|\theta^Q)) \\ &= E_{s_y,a_y,\gamma_y,s_{y+1}\sim buffer}[(\nabla_{\theta^\sigma} Q(s_y, \sigma(s_y)|\theta^Q)\nabla_{\theta^\sigma}\sigma(s_y|\theta^\sigma)]\end{aligned} \tag{13}$$

Following the updates to the critic and actor networks using transitions from the experience buffer, the target actor and target critic networks are updated in the following manner:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \tag{14}$$

$$\theta^{\sigma'} \leftarrow \tau\theta^\sigma + (1-\tau)\theta^{\sigma'} \tag{15}$$

where $\boldsymbol{\tau}$ is the learning rate.

2. Policy Gradient Methods

Policy gradient methods are a subclass of the algorithms used in learning reinforcement that intend to find the policy that optimizes the given function by calculating the estimator of the gradient of the policy. The policy gradient estimator is represented as $\hat{g} = E_y[\nabla_\theta log\pi_\theta(a_y|s_y)\hat{A}_y]$, where π_θ denotes the stochastic policy and $\hat{A}_y$ is an estimator of the advantage function at a given timestep y. Proximal Policy Optimization (PPO) with Actor-Critic is one of the policy gradient methods which involves optimizing the policy by combining multiple components: precisely, the clipped surrogate loss, the value function loss, and an entropy bonus are chosen. It incorporates the notions of A2C and trust region policy optimization (TRPO), allowing numerous epochs of minibatch updates. TRPO is a sophisticated technique applied in the reinforcement learning process to optimize policies without resulting in large updates which can destabilize the system. The cost function of the TRPO objective function aspires to maximize the expected advantage which can be expressed as:

$$maximize_{\theta}\hat{\mathbb{E}}_t\left[\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta\_old}\left(a_t \mid s_t\right)}\hat{A}_t\right] \tag{16}$$

$$\text{Subject to}: \quad \hat{\mathbb{E}}_t\left[KL\left[\pi_{\theta_{old}}\left(. \mid s_t\right), \pi_\theta\left(a_t \mid s_t\right)\right]\right] \leq \delta \tag{17}$$

The given constrained optimization problem is solved under the conjugate gradient framework of linear and quadratic approximations. However, theoretical considerations indicate that perhaps it is more appropriate to use a penalty term β on the KL divergence, as follows:

$$maximize_{\theta}\hat{\mathbb{E}}_t\left[\frac{\pi_\theta\left(a_t\mid s_t\right)}{\pi_{\theta_{old}}\left(a_t\mid s_t\right)}\hat{A}_t\right. \\ \left.-\beta\left[KL\left[\pi_{\theta_{old}}\left(.\mid s_t\right),\pi_\theta\left(a_t\mid s_t\right)\right]\right]\right] \tag{18}$$

However, TRPO uses a constraint due to difficulties in choosing an optimal β. PPO optimizing the policy by combining multiple components: precisely, the clipped surrogate loss, the value function loss, and an entropy bonus are chosen. The objective function $L_{CLIP+VR+S}(\theta)$ combines these elements to allow a stable and balanced policy update.

$$L_{CLIP+VR+S}(\theta) = E_y[L_{CLIP}(y,\theta) - c_1 L_{VR}(y,\theta) \\ + c_2 S[\pi_\theta(s_y)] \tag{19}$$

Here L_{CLIP} is the clipped surrogate loss, L_{VR} is the value function loss, $S[\pi_\theta]$ is the entropy bonus, c_1 and c_2 are coefficients. The clipped surrogate loss L_{CLIP} reduces the impact of large changes to the policy by bounding the probability ratio, and the value function loss L_{VR} is supposed to decrease the variance of the value function estimate and the entropy bonus $S[\pi_\theta]$ which helps to explore by adding randomness of the policy.

```
Algorithm (Pseudocode)
  Input: Historical stock data, consumer sentiment, news sentiment, earnings reports Reward $R_i$ incorporating portfolio value and Max Drawdown, Action space $\{-q,\ldots,-1,0,1,\ldots,q\}$, DRL agents: DDPG, PPO, A2C. Replay buffer denoted as $D$
  Output: Optimized trading strategy $T_S$
  Initialize: Portfolio value $V_0$, balance, environment $E$ with portfolio of stocks
  For each episode $i = 1,\ldots\ldots\ldots, max\_episodes$ :
      Initialize: Build state $s_0 = \{s_{0,1}, s_{0,2,\ldots.}s_{0,m}\}$ for $m$ stocks
      For each time step $t = 1,\ldots\ldots\ldots, max\_steps$ :
          Action $\leftarrow a_t \in \{a_{t,1},\ldots..a_{t,m}\}$
          Execute Trade $\leftarrow a_t$ buy, sell, hold
          Update portfolio $\leftarrow$ Adjust balance and shares
          Reward and Next state $\longleftarrow R_t$ and$S_{t+1}$
          Store Transition $D \longleftarrow (s_t, a_t, r_t, s_{t+1})$
          Sample mini batch from $D$
          Update actor-critic networks: Using DDPG, PPO, or A2C gradient updates.
      End For
      Evaluate portfolio performance using $T_S$
  End For
```

Coefficients c_1, c_2 balance the importance of two terms that are the value function loss and entropy bonus correspondingly. In the case of advantage estimation, PPO uses Generalized advantage estimation or a clipped version of the same to estimate the advantage function $\hat{A}_y$ expressed below:

$$\hat{A}_y = \delta_y + (\gamma\lambda)\delta_{y+1} + \ldots\ldots + (\gamma\lambda)^{Y-y+1}\delta_{Y-1} \tag{20}$$

where:

$$\delta_y = \gamma_y + \gamma V(s_{y+1}) - V(s_y) \tag{21}$$

This method approximates the advantage by adding the amount of the instant reward to the product of the discount factor and the difference between consecutive estimated values. Since LE provides a mean advantage estimator, the generalized advantage estimation is an estimator that offers balance between bias and variance by the introduction of a discount factor γ andand a smoothing parameter λ. This approach yields a superior and more stable estimate of the advantage which is of prime importance for updating policies.

3. Advantage Actor-Critic (A2C)

The combination of value-based and policy-based methods will lead towards actor-critic algorithms of RL. These actor-critic methods tackle the challenges of high variance which is usually happening during the backpropagation of policy-gradient methods. Therefore, combining actor-critic techniques with generalized advantage estimation significantly reduces the unpredictability of gradient updates. In this regard, one of the most popular algorithms of RL algorithms is A2C which boosts up policy gradient updates. This variation in policy gradient is tackled through the inclusion of the advantage function. This advantage function is calculated by the critic agent, instead of only approximating the value function. The evaluation of a trading agent's action is based not just on the positive outcomes of the action, such as selling, buying, or holding stocks, but additionally on its possibilities for additional improvement. With this inclusion, the robustness of the model increases while on the other hand, the variance of policy agents is going to decrease. By using the various data examples, the A2C model modifies the gradients using identical network architectures for all agents. Each agent separately takes action and interacts with the stock market environment on their own. The average of all gradients from all models is passed towards the global model and are processed by the coordinator in A2C in each iteration. This global network is then subsequently utilized in the actor-critic networks. For large data of stocks, A2C employed synchronized gradient updates for fast execution. Because of such advantages, the actor-critic model is considered to be more reliable for this task. The mathematical formulation of the objective function of A2C is given below:

$$\nabla J_\theta(\theta) = \mathbb{E}\left[\sum_{t=1}^{T}\nabla_\theta log\pi_\theta\left(a_t\mid s_t\right)A\left(a_t\mid s_t\right)\right] \tag{22}$$

In the above equation, $\pi_\theta\left(a_t\mid s_t\right)$ represent the network of policy, however, the second term shows the advantage function as $A(a_t \mid s_t)$ which can be re-defined as follows:

$$A(a_t \mid s_t) = Q(a_t \mid s_t) - V(s_t) \tag{23}$$

Or

$$A(a_t \mid s_t) = r(s_t, a_t, s_{t+1}) + \gamma V(s_{t+1}) - V(s_t) \tag{24}$$

In the above equation, the temporal difference error is computed through $r(s_t, a_t, s_{t+1}) + \gamma V(s_{t+1})$ on state-value function. The architecture of actor-critic networks is based on the multi-layer neural networks.

IV. EXPERIMENTS AND RESULTS

This section discusses the dataset, the assessment criteria used for examining the model, the proposed technique's results, and the analysis of findings.

A. DATA COLLECTION

B. This research proposed a multi-stock trading strategy using DRL agents by utilizing Dow Jones 30 constituent stocks. For the purpose of this study, the data used was obtained from Yahoo Finance API with a timeframe from January 1, 2010, to April 01, 2024. Training data was used from January 1, 2010, through October 1, 2021, and backtesting was done from October 1, 2021, through April 01, 2024. Moreover, data from earning reports were collected from Alpha Vantage API. Table 1 contains descriptive statistics for Dow Jones stocks' historical data, comprising significant metrics such as mean, standard deviation (std), count, minimum (min), maximum (max), and percentiles (25%, 50%, 75%). This Table 1 summarizes the dataset in terms of observations. Following on, Figure 5 shows the plots of the top 5 volatile stocks of DOW Jones i.e. closing prices, daily returns, and their 20-day moving average trend. It is observed from Figure 5 that a 20-day moving average can help discover fundamental trends while filtering out daily market noise. Figure 6 shows the trading volume i.e. quantity and number of shares of each stock listed on DOW Jones and it is observed that AAPL has the highest trading volume emphasizing its strong trading and investor interest. Subsequently, Table 2 shows the full names of stock symbols that were used in the analysis, and Table 3 lists the indicators and the original sources where the data was acquired. Furthermore, the training details of the proposed framework include the programming language Python, and all simulations are run on Google Colaboratory using the Free GPU facility, which includes NVIDIA's T4 GPUs. On colaboratory, the most recent versions of Python 3.10, NumPy, and OS libraries are pre-installed, with the exception of some unique libraries such as ''yfinance''.

B. EVALUATION METRICS

In this study, different assessment measures are utilized in order to evaluate the performance of the proposed agent in making stock trading decisions. The metrics utilized are as follows:

- **Cumulative Return:** This is the total return that is accumulated by the time of the close of the trading day.

Mathematically, it is defined in equation (25):

$$R = \frac{v - v_0}{v_0} \tag{25}$$

In the above equation, v represents the most recent portfolio value, whilst v_0 is the start of capital.

Sharpe Ratio: This metric assesses the performance of the agent by calculating the returns on investments in addition to risk consideration. It is computed using an equation (26):

$$S_T = \frac{mean(R_t) - r_f}{std(R_t)} \tag{26}$$

While

$$R_t = \frac{v_t - v_{t-1}}{v_{t-1}} \tag{27}$$

In the above equation, r_f is the risk-free rate while v_t is the portfolio value at time stamp t.

- **Max Drawdown:** This measures how sound the agent is since it determines the worst loss from the highest to the lowest point within a period, thus the value of an agent's portfolio at its lowest. Maximum drawdown indicates the possibility of a downside over a certain time span. It is computed by using the trough and peak values of the portfolio.

$$MDD = \frac{Trough\ value - Peak\ value}{Peak\ value} \tag{28}$$

- **Annual Return:** This measure gives the amount of profit and or loss made through an investment portfolio within one year. It is defined in equation (29):

$$r = (1 + R)^{\frac{365}{t}} - 1 \tag{29}$$

In the above equation, t denotes the total number of trading days.

- **Annual Volatility:** This figure quantifies the stability of the model by computing the standard deviation of the portfolio returns to find out how much the returns vary.

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{n} (r_i - \bar{r})^2}{n - 1}} \tag{30}$$

In the above equation, i represents the year for which annual return is calculated symbolized as r_i while $\bar{r}$ is indicating mean annual return as well as the total number of years symbolized as n.

- **Calmar Ratio:** This performance measure assesses the fund performance mainly, investment firms like hedge funds or commodities trading advisors (CTAs) in relation to annualized return and the max drawdown. It is defined below in equation (31):

$$Calmar\ ratio = \frac{R_p - R_f}{MDD} \tag{31}$$

In the above equation, R_p is the portfolio return, R_f is indicating the risk-free rate, while the MDD is the value of the maximum drawdown.

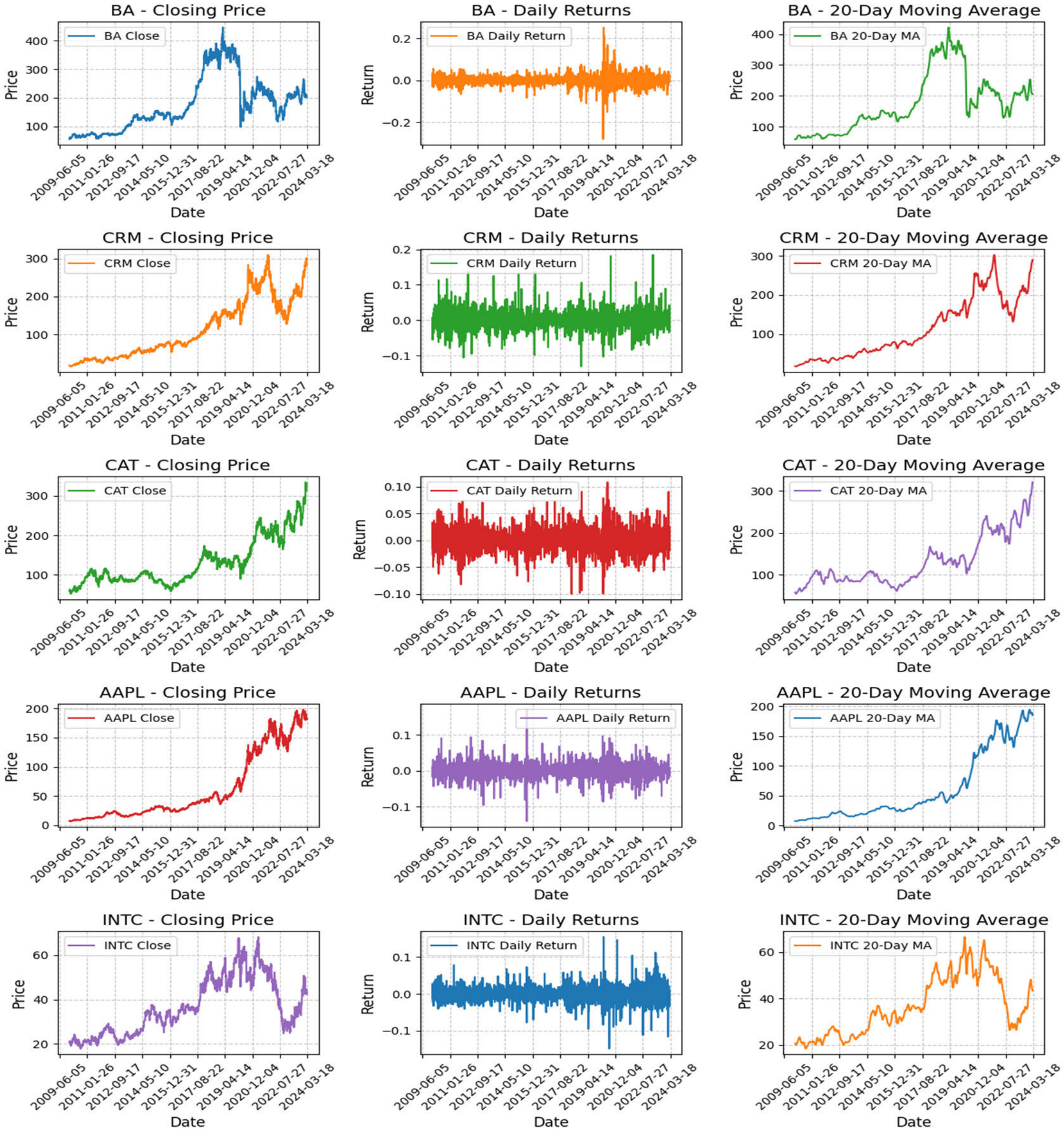


FIGURE 5. Analysis of top five volatile stocks of DOW JONES dataset in terms of closing price, moving average and daily returns.

TABLE 1. Descriptive statistics of DOW Jones stock's historical data.

index	open	high	low	close	volume	day
count	103327.0	103327.0	103327.0	103327.0	103327.0	103327.0
mean	107.48	108.47	106.48	94.2059	19340432.376	2.0261
std	79.063	79.850	78.268	77.126	59961476.165	1.3979
min	6.8704	7.0	6.795	5.792	305400.0	0.0
25%	48.544	49.0	48.119	39.9708	3934650.5	1.0
50%	87.15000	87.87000	86.4199	71.15947	7162900.0	2.0
75%	145.8699	147.1799	144.487	126.0045	14841200.0	3.0
max	555.0	558.099	550.130	546.607	1880998000.0	4.0

- **Omega Ratio:** This measure gives the volatility profile of the investment strategy managed by the trading agent. This indicator is derived by dividing the cumulative return distributions into segments with losses and gains

TABLE 2. Symbols of utilized stocks and their abbreviations.

Symbol of Stock	Abbreviation	Symbol of Stock	Abbreviation
AXP	American Express Co	**JPM**	JPMorgan Chase & Co
AMGN	AMGN Inc	**MCD**	McDonald's Corp
Apple	Apple Inc	**MMM**	3M Co
BA	Boeing Co	**MRK**	Merck & Co Inc
CAT	Caterpillar Inc	**MSFT**	Microsoft Corp
CISCO	Cisco Systems Inc	**NKE**	Nike Inc
CVX	Chevron Corp	**PG**	Procter & Gamble Co
TRV	Travelers Companies Inc	**CRM**	Salesforce Inc
VZ	Verizon Communications Inc	**WBA**	Walgreens Boots Alliance Inc
WMT	Walmart Inc	**DOW**	Dow Inc
GS	Goldman Sachs Group Inc	**IBM**	International Business Machines Corp
HD	Home Depot Inc	**INTC**	Intel Corp
HON	Honeywell International Inc	**JNJ**	Johnson and Johnson
UNH	UnitedHealth Group Inc	**KO**	Coca-Cola Co
V	Visa Inc	**DIS**	Walt Disney Co

based on a set threshold.

$$\Omega(r)\frac{\int_{r}^{\infty}(1-F(x))\,dx}{\int_{r}^{\infty}F(x)\,dx} \tag{32}$$

In the above equation, F shows the cumulative probability distribution function of the returns while r is the threshold for the target return.

- **Tail Ratio:** This measure established the risk-adjusted value of the investment by taking into consideration the downside risk. It is a measure of the investment's mean return over the best-performing tails (95% vs 5%). It is defined as:

$$Tail\ raio = \frac{R_{95}}{|R_5|} \tag{33}$$

In the above equation, R_{95} is the 95th percentile of the return distribution, R_5 is the 5th percentile of the return distribution

- **Sortino Ratio:** Like the Sharpe Ratio, the Sortino measures the risk-adjusted performance but with an additional consideration which is that of below-target returns.

$$Sortino\ ratio = \frac{R_p - R_f}{D_V} \tag{34}$$

In the above equation, R_p is the portfolio return, R_f is indicating the risk-free rate, while the D_V is the target downside deviation.

C. RESULTS OF THE PROPOSED RMS MODELS AND ^DJI BASELINE INDEX

To validate the performance of the proposed framework comprising all DRL agents namely A2C_RMS, DDPG_RMS, and PPO_RMS, they are initially trained and later on, their performance is accessed through backtesting. Table 4 shows the results of the proposed agents in terms of different metrics. It is observed from Table 4 that A2C_RMS is more profitable on average compared to other models showing the highest annualized return at 10% and a cumulative return of 27% outperforming PPO_RMS at 8% annualized return, and 22% cumulative returns and DDPG_RMS at 7% annualized and 19% cumulative returns. The DJI index performs a little bit less, with a yearly return of 5% and a total return of 13%. This lower performance highlights the relative strength of the proposed model, which outperforms the baseline index. Even in terms of risk-adjusted return based on the Sharpe ratio; A2C_RMS has a better risk-adjusted return of 0.66 in comparison with the ^DJI baseline index having 0.42. The Calmar ratio and Omega ratio are significant in that A2C_RMS performs more effectively in the aspect of better risk-adjusted returns and also far more efficiently in terms of the number of positive returns compared to other models. Following on, if the comparison has been carried out in terms of Max drawdown, then all agent exhibits nearly equal values e.g.,

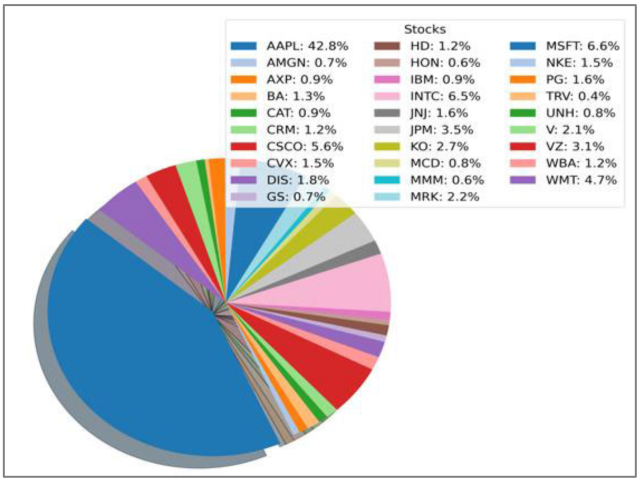


FIGURE 6. Percentage of trading volume of each stock in DOW jones.

A2C_RMS shows −0.21788 Max drawdown. Likewise, the proposed model of A2C_RMS also works well in terms of the Sortino ratio and Omega ratio which is about 0.96 and 1.11. More explicitly, the more closely the Sortino ratio is near to one or more, the more effective the model is in producing returns without subjecting the portfolio to severe negative risk. In this case, a ratio of 0.96 indicates that the A2C_RMS model is literally risk-neutral with respect to bad returns, which is normally beneficial for investors. Similarly, an omega ratio of 1.11 indicates a quite significant ability of the proposed model to generate positive returns when contrasted to losses and shows that DRL models have a good ability to make trading decisions.

TABLE 3. Sources of collected indicators.

Indicator	Source
Daily News Sentiment	Federal Reserve Bank of San Fransisco
Consumer Sentiment Index	FRED API
Earning Reports	Alpha Vantage
Fiscal date ending	Alpha Vantage
reportedEPS/estimated EPS	Alpha Vantage
Surprise	Alpha Vantage
Surprise percentage	Alpha Vantage
Reporteddate	Alpha Vantage
Reportedtime	Alpha Vantage

Subsequently, the Calmar ratio is also good having a value of 0.48 with A2C_RMS, 0.43 with PPO_RMS, and 0.36 with DDPG_RMS. These values are comparatively larger than the baseline ^DJI index having a Calmar ratio of 0.24. In the context of stability, A2C_RMS has the best stability at 0.3323, indicating greater consistency in performance than the other DRL agents designed for stock trading. Following on, PPO_RMS is somewhat stable at 0.2647, whereas DDPG_RMS as well as the DJI Baseline Index have lower stability scores of 0.2139 and 0.069647, respectively. This stability metric shows the consistency of models in generating returns and it is observed that A2C_RMS has good performance not only in the case of good Sharpe, Omega, and Calmar ratios but also in good values of stability.

In a nutshell, it can be logically deduced that the A2C_RMS is more effective in achieving the majority of the metrics under consideration. This higher performance is primarily due to the actor-critic framework used by A2C_RMS and its advantage function. Figure 7 shows the graphs of different models indicating their portfolio values over the testing data during backtesting. The portfolio value of all proposed models namely A2C_RMS and PPO_RMS models against the baseline ^DJI index during backtesting provide a clear analysis of these investment strategies at a glance. For instance, in A2C_RMS, i.e. Figure 7(a) the portfolio returns of A2C_RMS is much larger than the baseline ^DJI index with a maximum value of 1242380. In the same way, for PPO_RMS, Figure 7 (b) shows the dynamics of the portfolio value compared to the ^DJI index. A higher and more consistently rising PPO_RMS would better compare to the ^DJI line indicating better overall performance and stability. Similarly, Figure 7(c) shows the portfolio value of DDPG_RMS, however, it is observed that its performance is a little bit less in comparison with other models with the maximum portfolio value at the end of the trading period being 1198754 respectively.

D. COMPARATIVE ANALYSIS OF PROPOSED MODELS WITH DIFFERENT INDICATORS

In this experimental setup, the performance of proposed models is investigated by analyzing the impact of individual indicators namely daily news, consumer sentiment, and data from earning Reports, and any data that is a combination of these types. The examined concern and risk metrics include the following: the Sharpe ratio, the Calmar ratio, the Tail ratio, the Sortino ratio, and the Omega ratio. The graphs shown in Figure 8 provide the results of this experimental setup. More precisely, Figure 8(a) and Figure 8(b) show the performance of A2C_RMS models for each input data type along with these ratios. For instance, A2C_RMS_ALL seems to have comparable performance across all the plots, showing how the model works when all the sources of data are considered. The second graph in Figure 8 (c) and Figure 8(d) shows the results of DDPG_RMS models over different indicators individually. This also presents the comparison of several metrics (Sharpe ratio, Calmar ratio, Tail ratio, Sortino ratio, Omega ratio). The indicators involve news sentiment, consumer sentiment, earning reports, and all of them at once. The results of DDPG indicate that with all indicators, the DDPG_RMS performance is comparable. However, from the graphs, it is observed that DDPG_RMS shows more good performance if only daily news sentiment is modeled into the state representation. This indicates that by involving this external element of news sentiment in the state representation, DDPG_RMS can obtain a more sophisticated knowledge of market dynamics, highlighting the interaction between sentiment-driven market behaviors and stock price performance. It is also observed that the incorporation of other variables such as consumer sentiment and data from earnings reports is also beneficial, but it seems to reduce the model's ability (i.e. DDPG_RMS) to efficiently analyze and make trading decisions (particularly cumulative returns and annual returns is less as observed in Figure 8(d)), potentially due to greater noise or complexity in the state representation. Following on, the third comparison has been made on PPO_RMS with the same experimental settings. Figure 8 (e) and Figure 8 (f) compare the performance of different PPO_RMS using different combinations of indicators. The performance measures include various risk-adjusted return metrics. The most common measures are the Sharpe ratio, Calmar ratio, Tail ratio, Sortino ratio, and Omega ratio. It is observed from the results of PPO_RMS that with a combination of all factors, the model shows good performance in terms of Sharpe ratio,

TABLE 4. Results of the proposed RMS models and ^DJI baseline index.

Evaluation metric	PPO_{RMS}	$A2C_{RMS}$	$DDPG_{RMS}$	^DJI Baseline Index
Annual return	**8%**± 0.03098	10%± 0.012871	7%±0.0071	5%
Cumulative returns	**22%**±0.0851	27%± 0.035650	19±0.019	13%
Annual volatility	0.1559± 0.006	0.167± 0.026057	0.1597±0.0094	0.154722
Sharpe ratio	**0.60**±0.177	0.66±0.0552	0.544 ±0.035	0.420725
Calmar ratio	0.43338±0.1790	0.4847±0.0797	0.3684 ±0.0560	0.248275
Stability	0.2647±0.2425	0.3323±0.08543	0.2139±0.0277	0.069647
Max drawdown	-0.20290±0.02	-0.21788±0.062	-0.2111± 0.024	-0.219408
Omega ratio	1.1074±0.03	1.1190±0.0101	1.09646± 0.0067	1.074316
Sortino ratio	0.873957±0.2600	0.96247±0.08345	0.7777±0.0513	0.598321
Tail ratio	1.0074± 0.04880	1.032300±0.0309	1.0100±0.0253	1.028740
Daily value at risk	-0.01927± 0.00083	-0.02065± 0.0032	-0.0197±0.001163	-0.019235

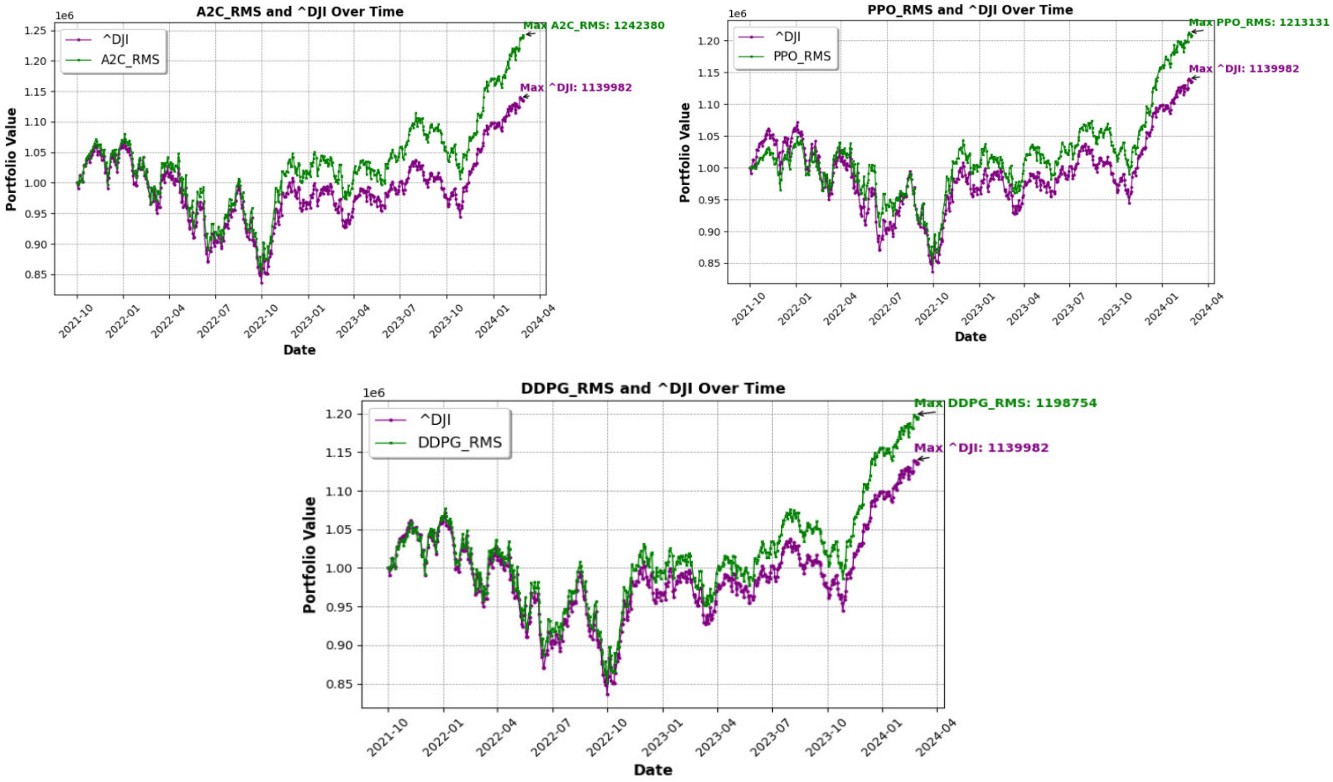


FIGURE 7. (a)Portfolio value of proposed A2C_RMS vs baseline ^DJI index during back testing (b) Portfolio value of proposed PPO_RMS vs baseline ^DJI index during back testing (c) Portfolio value of proposed DDPG_RMS vs baseline ^DJI index during back testing.

Calmar ratio, and returns (both annual and cumulative). The observed findings show that the PPO_RMS model performs better when the state representation includes a wide range of elements, such as daily news sentiment, consumer sentiment, and earnings report data. This shows that PPO_RMS is well-suited to dealing with large and diverse data inputs, using its policy optimization approach to identify correlations and trends across multiple sources of data. Following on, the performance in terms of portfolio value has also been analyzed such as Figure 9 shows the performance of A2C_RMS in optimizing portfolio with different indicators, Figure 10 shows the performance of DDPG_RMS in optimizing portfolio with different indicators and Figure 11 shows the performance of PPO_RMS in optimizing portfolio with different indicators. From the analysis, it can be concluded that the A2C_RMS model (From Figure 9) shows the good value of a portfolio over the backtesting period when only news sentiment has been involved especially during the time period after 2022. Similarly, it is observed that with only data from earning reports, the portfolio values are not as good, but on the other hand, the consumer sentiment shows a good impact. Furthermore, from Figure 10, it is observed that DDPG_RMS shows good performance when data of earning reports and news sentiment has been modeled into state representation. However, performs a little bit worse in cases when only consumer sentiment has been utilized in addition to daily historical stock data and technical indicators into the states of DDPG_RMS. In the last, i.e. from Figure 11, it is observed that the PPO_RMS model shows good results when consumer sentiment is involved. The differences in performance of all models, when different types of data are utilized, are due to their learning mechanisms and trading strategies. In the previous experimental setup as observed in Figure 7, it is clearly depicted that by a combination of all factors i.e. data from earning reports, consumer sentiment index, and daily news sentiments when modeled into the state representation then the results of all models are good. However, during ablation studies when individual factors are considered the performance of A2C_RMS is observed good when only news sentiment is involved, but in the rest of the cases, the performance of individual indicators is averaged.

E. MAX DRAWDOWN ANALYSIS OF RL AGENTS ACROSS DIFFERENT DATA SOURCES

Based on the Max Drawdown analysis made on the proposed reinforcement learning (RL) agents, it can be observed from Table 5 that the performance of the agents depends on different data indicators. In terms of Max drawdown analysis on different data indicators, the A2C_RMS model has consistently the least maximum drawdown, which implies that it incurs a less extreme loss. In particular, the DDPG_RMS algorithm shows a maximum drawdown value of -0.202 when all combinations of factors were involved i.e. A2C_RMS_ALL. Similarly, a max drawdown of -0.187 is achieved when data from earning reports are included in the state representation. It is noted that A2C_RMS and PPO_RMS are fairly comparable, however, DDPG_RMS has minor deviations. It has been observed that the suggested RL models have good maximum drawdown values and are more suited for real-world trading since they exhibit controlled risk.

The proposed RMS model efficiently reduces possible portfolio losses under severe market situations, resulting in improved risk management and preservation of capital. In addition, the combined comparative analysis of proposed agents over different conditions is also provided in Figure 12 to show the collective results of all ablation studies.

F. COMPARATIVE ANALYSIS WITH BASELINE METHODS USING BACKTESTING

To verify the effectiveness of the proposed RMS_based models, we compared them to non-reinforcement learning techniques which include the Dow Jones Industrial Average (DJIA), optimization model based on Mean-variance with distinct objective functions [59], machine learning-based Hierarchical Risk Parity method [60], along with Critical Line Method [61]. The results of the proposed RL framework with these baseline algorithms are provided in Figure 13(a) and in Figure 13 (b), Mean Var (Sharpe) demonstrates Sharpe-based objectives, and Mean Var (risk) implies maximizing return for an objective with target risk, along with Mean Var (return) implies minimizing risk for a target return. More precisely, in Figure 13 (b), the comparative analysis has been done on the basis of Cumulative returns and Sharpe ratio, however, in Figure 13 (b), the comparison has been done on the basis of annual returns and the Sortino ratio

The reinforcement learning approach provided in this work yields encouraging results, with improved Sharpe ratio and cumulative returns as shown in Figure 13(b). The suggested trading technique produced high returns while also at the same time risk-adjusted returns, as evidenced by an improved Sharpe ratio.

TABLE 5. Max drawdown analysis of proposed RL agent.

Combinations	A2C_RMS	DDPG_RMS	PPO_RMS
A2C_RMS_News Sentiment	-0.18	-0.27	-0.195
A2C_RMS_Consumer Sentiment	-0.26	-0.23	-0.248
A2C_RMS_Earning Reports	-0.23	-0.25	-0.187
A2C_RMS_ALL	-0.217	-0.202	-0.21

G. DISCUSSIONS AND IMPLICATIONS

Algorithmic stock market trading is one of the critical and most favorable topics among researchers of finance because it helps investors optimize their portfolios by executing wise trading decisions.

These algorithmic trading approaches provide advantages over human traders, such as higher confidence, quicker execution, and lack of emotional biases. In existing studies, different trading strategies have been developed such as mean reversion [10], momentum-based methods [11], and rule-discovery techniques [12]. Nevertheless, these methods fail to cope with the challenging environmental conditions of the stock market and frequently show good performance over specific timeframes. The latest trend in this area is the introduction of RL-based approaches to stock market trading to address such issues. However, the primary challenge in stock trading is to effectively present states and develop agents' understanding of the market, as well as to provide appropriate rewards to encourage agents to learn to create profits while simultaneously managing risk. Therefore, in this research study, the proposed stock market trading model incorporating DRL with the Consumer Sentiment Index (CSI), daily news sentiment, historical stock data, technical indicators, and traditional earnings reports has shown promising

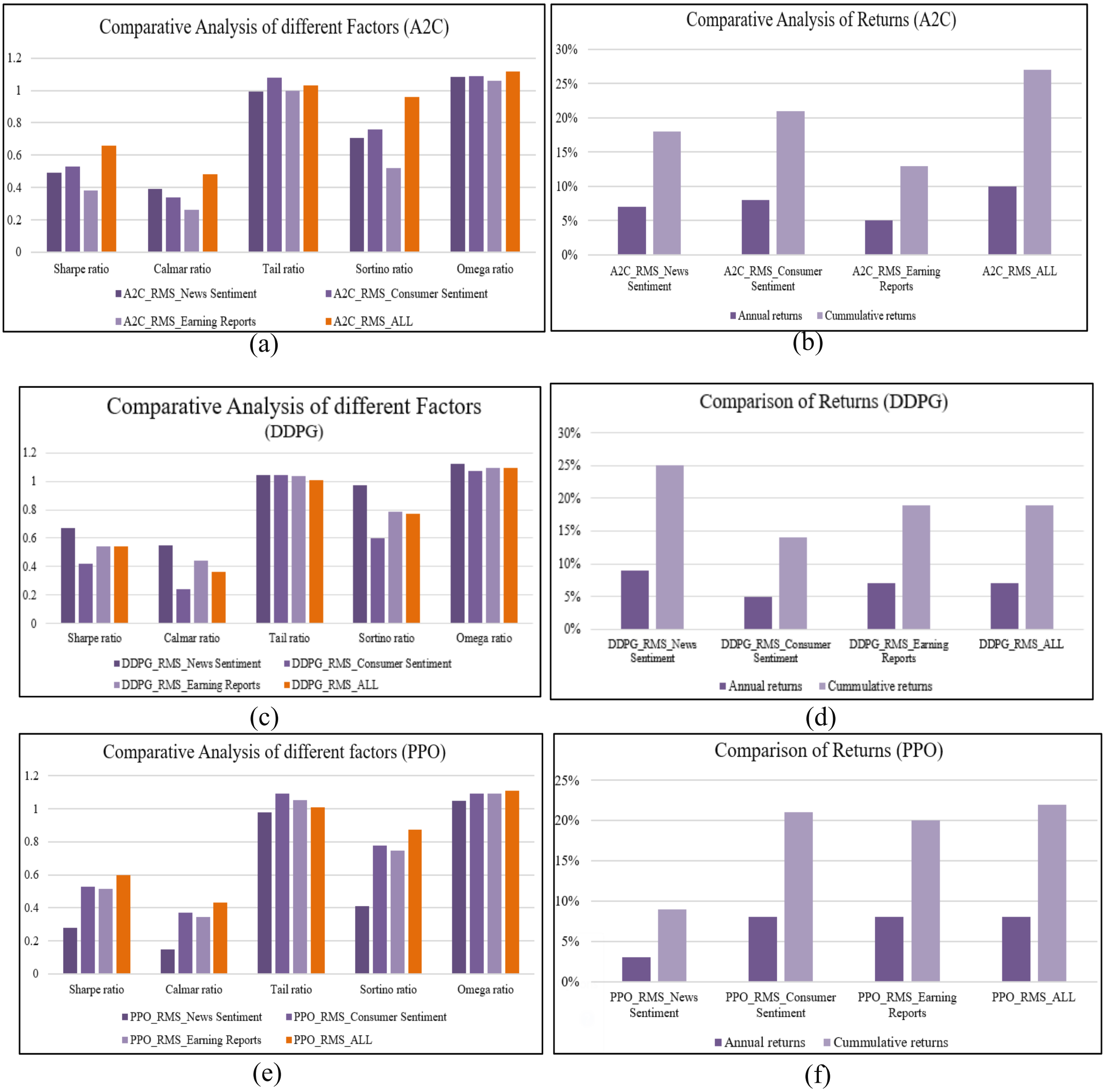


FIGURE 8. **(a)Comparative Analysis of Individual Indicators using proposed A2C based RMS model (b)Comparative Analysis of Individual Indicators using proposed DDPG based RMS model (c)Comparative Analysis of Individual Indicators using proposed PPO based RMS model.**

potential in improving the trading model. The most significant finding from this research is that combining consumer and news sentiment in addition to earning reports considerably improves trading decision-making. This inclusion provides a comprehensive view of the market by combining conventional financial indicators—such as profits and revenue from earnings reports, and with psychological and emotional factors derived from news sentiments that influence investor purchasing and investment choices. The variants of DRL proposed in this work including PPO_RMS, A2C_RMS, and DDPG_RMS underscore the benefits of including sentiment data with risk management approaches such as by including Max Drawdown based penalty in reward function. Both these models give due importance to long-term strategic planning, yet they are able to ensure short-term gains as well, thereby giving an edge over the conventional techniques proven by better risk-adjusted returns and other parameters such as the Sharpe and Calmar ratios. This implies that the proposed models are not only advantageous but that they are more robust in fluctuating markets. The most important implication of this research is the introduction of different factors in the state representation of trading

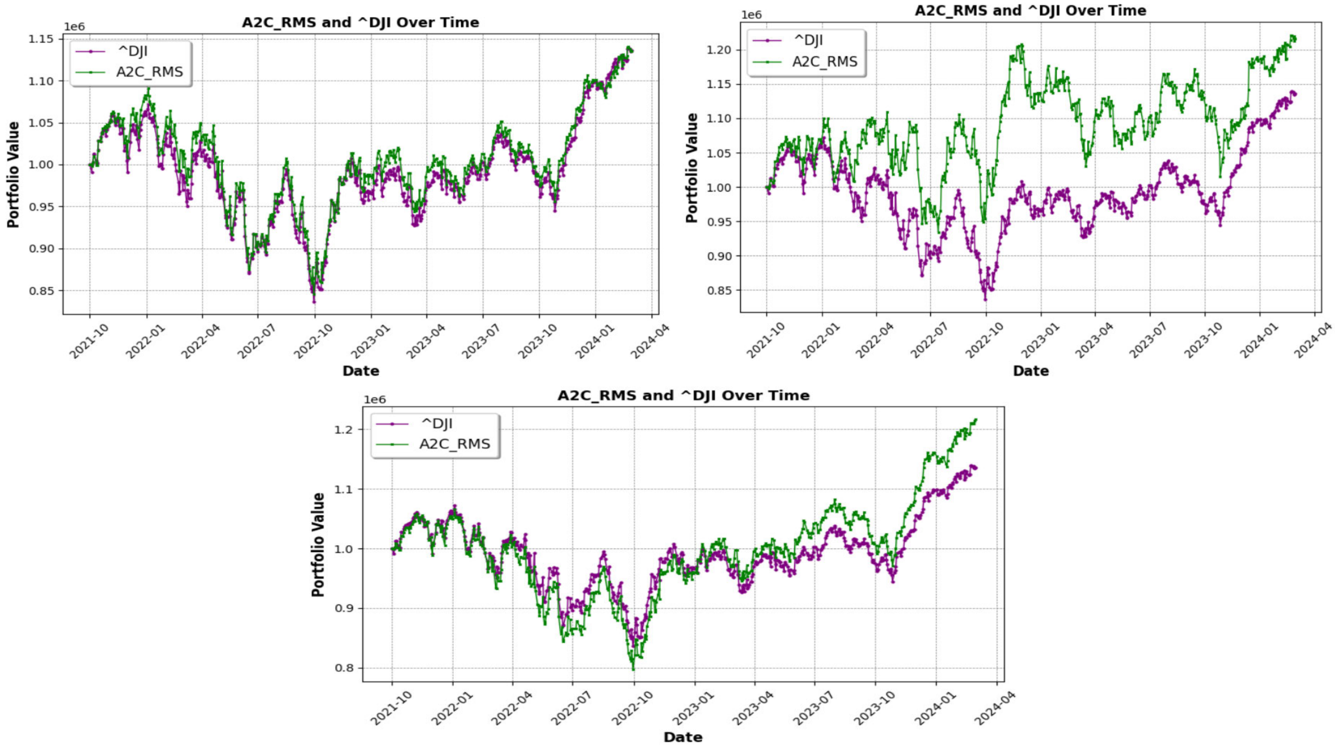


FIGURE 9. Portfolio value of proposed A2C_RMS vs baseline ^DJI index during back testing with data (a) of Earning Reports (b) news sentiment (c) Consumer sentiment.

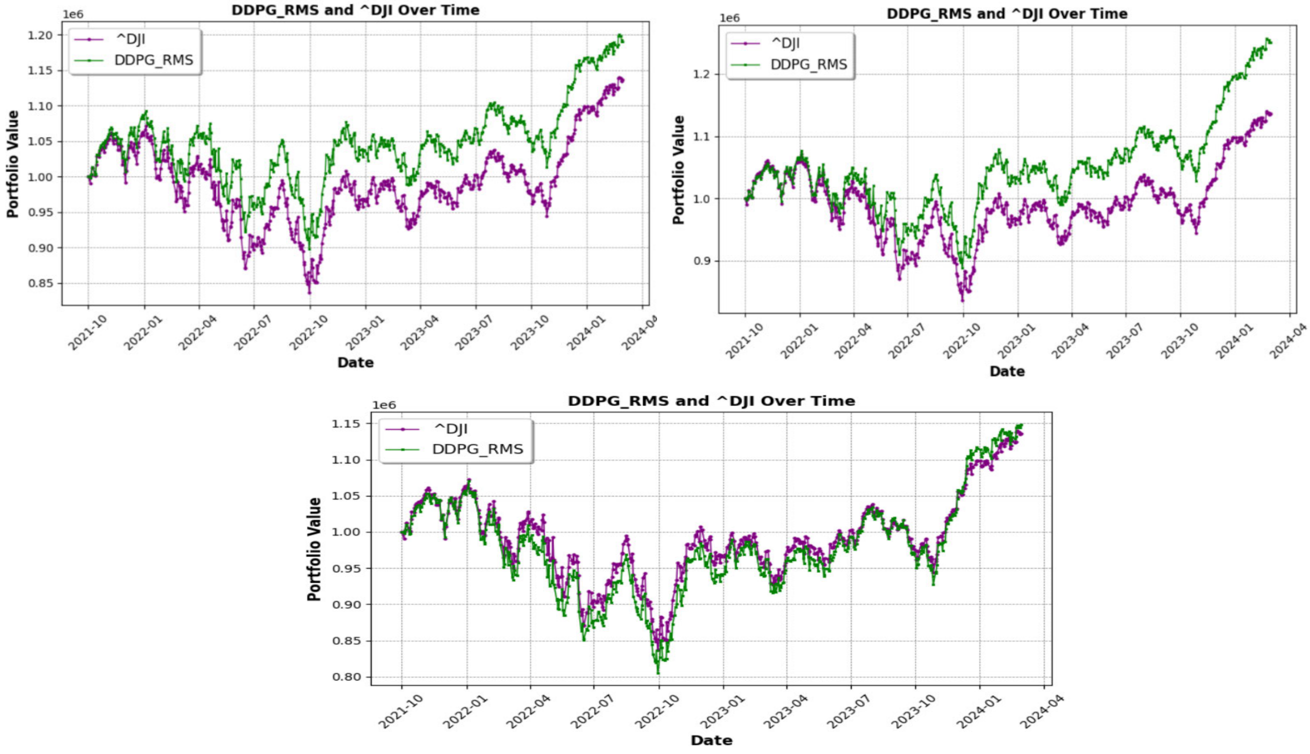


FIGURE 10. Portfolio value of proposed DDPG_RMS vs baseline ^DJI index during back testing with data (a) of Earning Reports (b) news sentiment (d) Consumer sentiment Index.

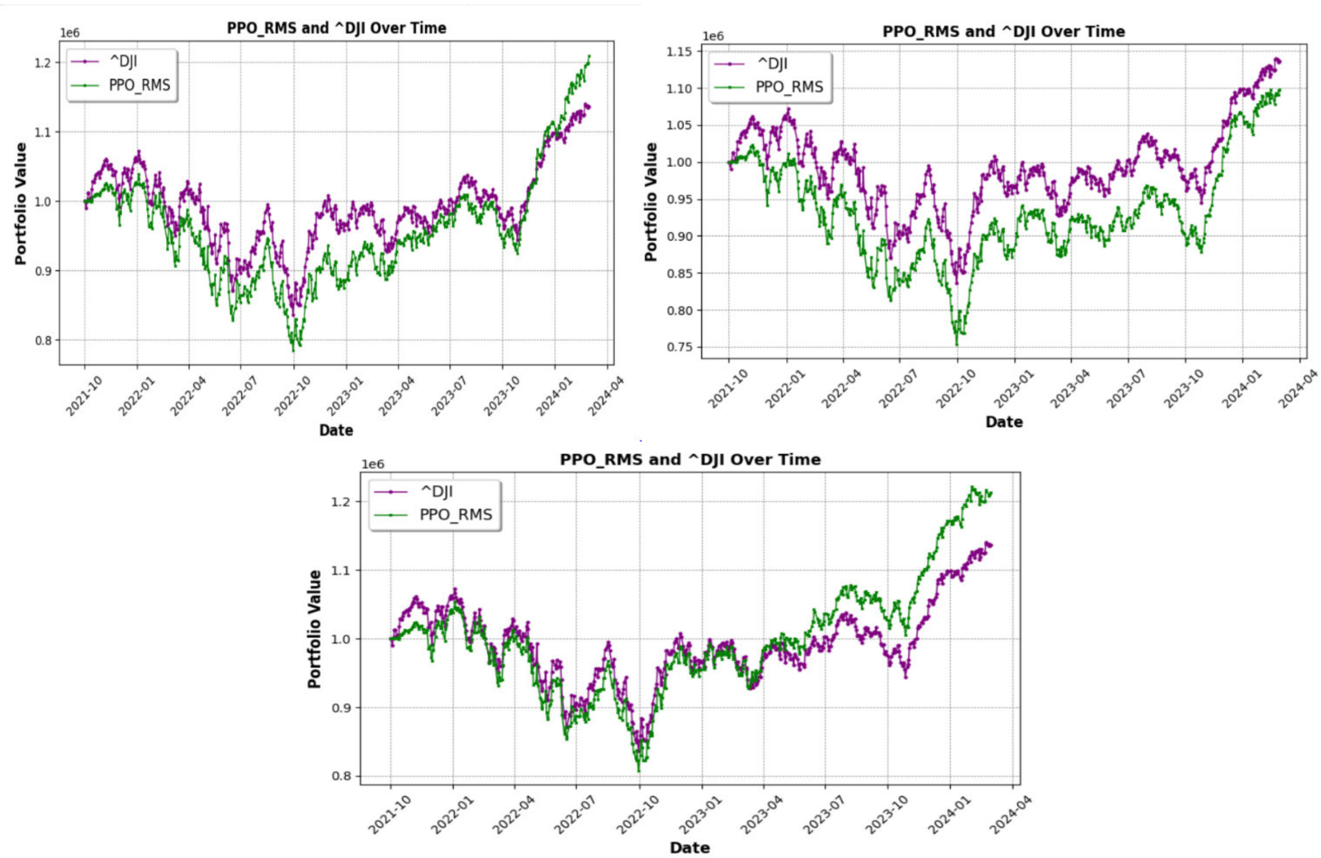


FIGURE 11. Portfolio value of proposed PPO_RMS vs baseline ^DJI index during back testing with data (a) of Earning Reports (b) news sentiment (d) Consumer sentiment.

algorithms. In the past, most of the policies simply depended on historical data, and other factors but neglected consumer sentiment index and data from earning reports. However, this study has demonstrated that there exists a lot of potential in involving these factors which can give an extra layer of insight to the algorithms and help them to predict the market trends better and also to respond to them better could become the foundation for a new generation of trading systems that would function better in today's global financial environment. Furthermore, the study recommends that sentiment indices when combined with the earnings reports may be useful in risk management. Max Drawdown integrated into the reward function of the proposed models captures the need for trading strategies not to always aim at high gains but low losses as well. The combination and/or integration of these two aspects is vital for traders and investors who want to get the most out of their profit and at the same time, minimize the level of risks.

Moreover, the proposed study is based on an intraday trading strategy in which daily stock data (i.e., end-of-day prices and relevant indicators) is employed. The rationale is that daily data achieves an ideal balance among granularity and stability. In comparison to high-frequency data (e.g., tick or minute-level), daily data includes substantially less noise and better represents real market trends. In addition, high-frequency data (e.g., minute-by-minute or tick-level data) is computationally extensive to process and requires more training time. Overfitting to short-term market signals is risky because high-frequency models might become too dependent on microstructure noise. As a result, daily data enables the RL agent to generalize more accurately while remaining computationally feasible. The RL-based methodology developed in this study is feasible for managers who want to maximize daily trading choices. The day-by-day approach provides individual investors with an automatic decision-making technique that assists in balancing returns and risk without requiring ongoing market monitoring. The model is transformed into an evolving automated decision-making system that changes with the evolving market conditions and helps managers in their portfolio decisions. The proposed RL-based strategy is able to adapt to continual updates using new data with online learning. Furthermore, the proposed model highlights crucial criteria for modeling the problem of portfolio optimization, such as the importance of consumer sentiment, the news sentiment index, and data from stock earnings reports.

Furthermore, the proposed RL model has good advantages over financial studies that achieve the same task with some

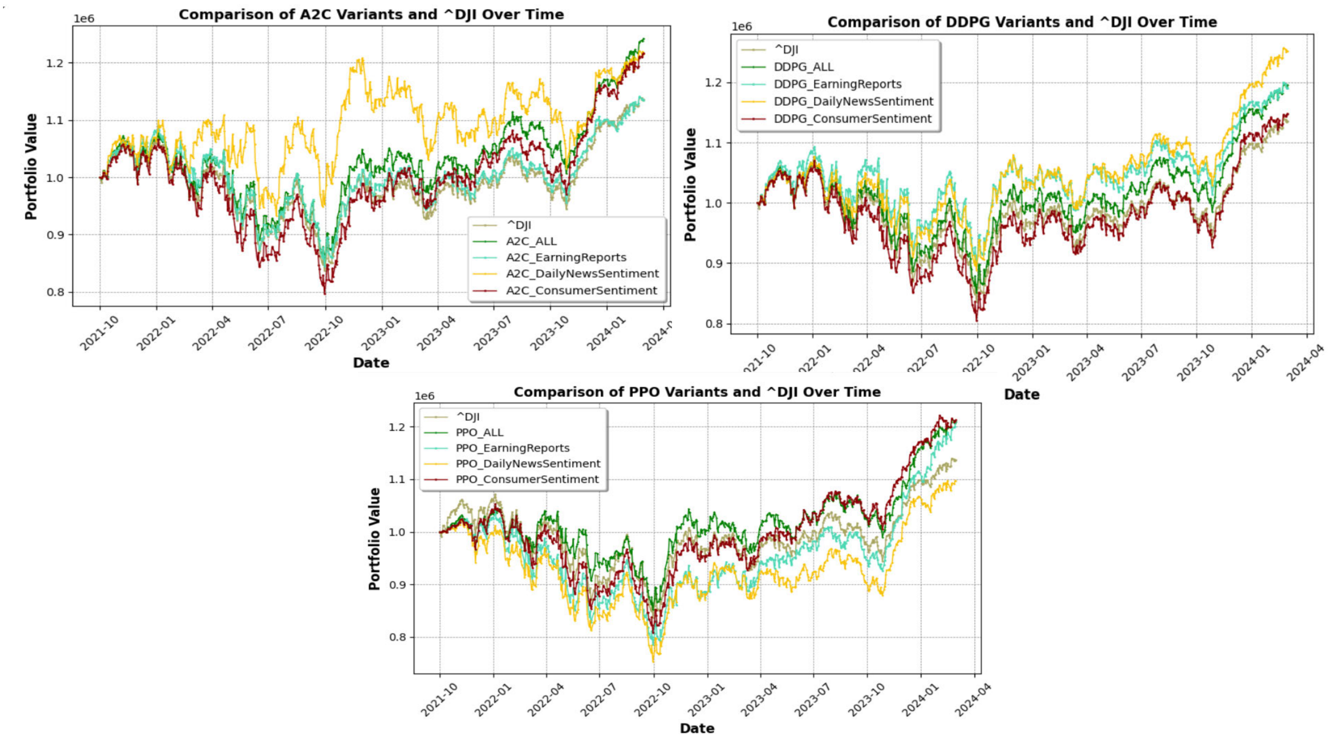


FIGURE 12. Combined comparison of proposed agents (A2C, DDPG, and PPO) on different indicators namely daily news sentiment, consumer sentiment index, and earning reports.

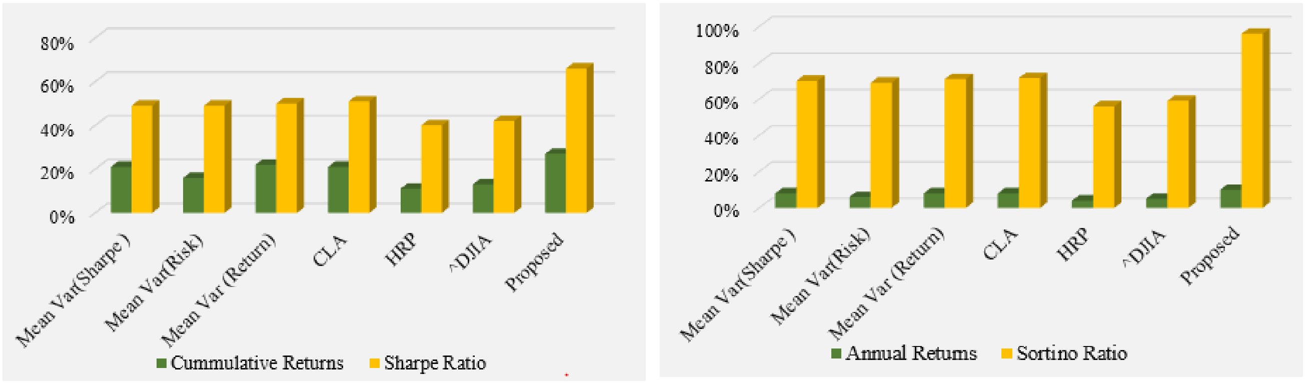


FIGURE 13. Comparison of proposed agents (A2C, DDPG, and PPO) with baseline methods (a) Cumulative returns and Sharpe ratio (b) Annual returns and Sortino ratio.

other simpler methods. Those simpler methods are although show good results but they have some limitations which is why the latest studies adopt the use of RL [62]. Several simple methods such as mean reversion [10], momentum-based methods [11], and rule-discovery techniques [12] exist in the literature, but they have poor generalization and do not show good results in all market conditions [62]. Some methods are based on two-step strategies such as the prediction of stock prices using supervised ML algorithms and later on trading decisions are made based on predictions, however, there exists a major gap between both stages [62]. A comparative analysis of baseline strategies both in existing studies [1], [41], [63], [64], [65], [66] as well as in this research (e.g. Figure 13) reveals the importance of RL over basic simple methods. These RL models modify their learned strategy in response to environmental conditions in comparison with fixed static rule-based simple methods. We acknowledge that these RL-based methods require more computational complexity in training in contrast to simple methods and are less interpretable but show good performance over these methods in the presence of complex, and non-linear dependencies. Moreover, the problem in this study is made simpler and is in line with that of actual portfolio managers by employing intra-day (i.e. daily) data instead of high-frequency data. The primary focus of this study is on algorithmically improving RL-based models instead of making problem complex for stock trading by considering and exploiting the usage of different factors such as earning reports, and consumer and news sentiments in addition to RL agent feedback using the proposed max-drawdown penalty-based reward function. Since these are overlooked in the existing literature of financial RL and this study aims to fill these research gaps. Hence, it is an important step forward in furthering and improving the use of RL in stock trading applications and thus showing good results over baseline simple methods as observed in Figure 13.

However, the study has some limitations that should be taken into consideration for future work. First of all, backtesting was conducted on a single stock market of DOW Jones, which in some way might not reflect the results in other markets such as SP500 having different levels of volatility in their prices. Future work also includes validating the model such as how it will perform when applied to real-time trading with live data and unanticipated events. Likewise, the findings of this study are mainly based on the U.S. stock markets and the results might not necessarily generalize to other country's markets characterized by different conditions and environments. One possible future work entail following the development of the given model and applying it to different international markets, considering local sentiment indicators and the market environment. Finally, although this study incorporates sentiment analysis and earnings reports, other variables including geopolitics, macros factors, and technology disruptions can also be considered to further improve the performance. Further research could continue the development of this model by incorporating these other data sources with the aim of improving the generalizability and accuracy of the model.

V. CONCLUSION

Stock market trading continues to be an important part of the worldwide economy, providing the potential for financial accumulation and diversification in a portfolio. In general, traders use financial domain knowledge and fundamental analysis to arrive at informed trading decisions with the goal of maximizing gains in the markets. Nevertheless, trading in the stock market is always complicated and challenging for traders because of the unanticipated characteristics of market behavior, high-dimensional information, and the interaction of numerous elements including news sentiments, consumer sentiment, macroeconomic variables, and stock micro-factors such as earnings. To address this, a novel strategy for stock market trading based on RMS factors i.e., enriching states of the DRL model with news sentiment, consumer sentiment, earning reports in addition to daily historical data, technical indicators, and Max drawdown rewards has been proposed.

The findings indicate that these variables helped in learning the DRL model for analyzing market movements and constructing better trading strategies. It can be concluded that the suggested RL models (PPO_RMS, A2C_RMS, and DDPG_RMS) outperform the DJI index as well as baseline methods of trading e.g., Mean Variance Optimization. The suggested reward function, with its Max Drawdown component, efficiently balances profitability and risk, resulting in effective trading strategies. Compared to all models, DDPG_RMS demonstrates good outcomes by demonstrating yearly returns of 10%, cumulative returns of 27%, Sharpe ratio of 0.66, and Sortino ratio of 0.96. In general, the addition of sentiment indices with traditional earnings enhances the DRL-based trading models and can be termed a major step forward in trading science by offering improved market insights and consistently high performance. Future studies will include leveraging additional factors to improve the DRL's state representation, such as macroeconomic indicators, event sentiments (such as sentiments centered on political events such as elections or natural disasters events), and correlation relationships across stocks. Furthermore, designing a better reward function, such

as utilizing the Sortino ratio, is a promising future study direction. Similarly, investigating hybrid RL models or adding ensemble learning approaches can improve stability as well as performance.

ACKNOWLEDGMENT

The authors would like to thank Prince Sultan University for their support.

REFERENCES

[1] Y. Ansari, S. Gillani, M. Bukhari, B. Lee, M. Maqsood, and S. Rho, "A multifaceted approach to stock market trading using reinforcement learning," *IEEE Access*, vol. 12, pp. 90041–90060, 2024.

[2] L. D. Oyeniyi, C. E. Ugochukwu, and N. Z. Mhlongo, "Analyzing the impact of algorithmic trading on stock market behavior: A comprehensive review," *World J. Adv. Eng. Technol. Sci.*, vol. 11, no. 2, pp. 437–453, Apr. 2024.

[3] O. Bustos and A. Pomares-Quimbaya, "Stock market movement forecast: A systematic review," *Expert Syst. Appl.*, vol. 156, Oct. 2020, Art. no. 113464.

[4] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 3007–3057, Apr. 2020.

[5] B. Dhingra, S. Batra, V. Aggarwal, M. Yadav, and P. Kumar, "Stock market volatility: A systematic review," *J. Model. Manage.*, vol. 19, no. 3, pp. 925–952, Nov. 2023.

[6] G. Cohen, "Technical analysis in investing," *Rev. Pacific Basin Financial Markets Policies*, vol. 26, no. 2, Apr. 2023, Art. no. 2350013.

[7] G. Chen, K. A. Kim, J. R. Nofsinger, and O. M. Rui, "Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors," *J. Behav. Decis. Making*, vol. 20, no. 4, pp. 425–451, Oct. 2007.

[8] X. Wu, H. Chen, J. Wang, L. Troiano, V. Loia, and H. Fujita, "Adaptive stock trading strategies with deep reinforcement learning methods," *Inf. Sci.*, vol. 538, pp. 142–158, Oct. 2020.

[9] F. G. D. C. Ferreira, A. H. Gandomi, and R. T. N. Cardoso, "Artificial intelligence applied to stock market trading: A review," *IEEE Access*, vol. 9, pp. 30898–30917, 2021.

[10] J. M. Poterba and L. H. Summers, "Mean reversion in stock prices: Evidence and implications," *J. Financial Econ.*, vol. 22, pp. 27–59, Aug. 1987.

[11] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *J. Finance*, vol. 48, no. 1, p. 65, Mar. 1993.

[12] J.-L. Wang and S.-H. Chan, "Stock market trading rule discovery using pattern recognition and technical analysis," *Expert Syst. Appl.*, vol. 33, no. 2, pp. 304–315, Aug. 2007.

[13] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications," *Int. J. Financial Stud.*, vol. 11, no. 3, p. 94, Jul. 2023.

[14] T. Kabbani and E. Duman, "Deep reinforcement learning approach for trading automation in the stock market," *IEEE Access*, vol. 10, pp. 93564–93574, 2022.

[15] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," *IEEE Access*, vol. 8, pp. 150199–150212, 2020.

[16] D. Jothimani and S. S. Yadav, "Stock trading decisions using ensemble-based forecasting models: A study of the Indian stock market," *J. Banking Financial Technol.*, vol. 3, no. 2, pp. 113–129, Oct. 2019.

[17] Z. Zhang, S. Zohren, and S. Roberts, "Deep reinforcement learning for trading," 2019, *arXiv:1911.10107*.

[18] N. Bhavatarini, S. T. Ahmed, and S. M. Basha, *Reinforcement Learning-Principles, Concepts and Applications*. Valbrembo, Italy: MileStone Research Publications, 2024.

[19] Y. Shi, W. Li, L. Zhu, K. Guo, and E. Cambria, "Stock trading rule discovery with double deep Q-network," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107320.

[20] B. Kommey, O. J. Isaac, E. Tamakloe, and D. Opoku, "Reinforcement learning review: Past acts, present facts and future prospects," *IT J. Res. Develop.*, vol. 8, no. 2, pp. 120–142, Feb. 2024.

[21] Y. Yu, "A survey of deep reinforcement learning in financial markets," in *Proc. 3rd Int. Academic Conf. Blockchain, Inf. Technol. Smart Finance (ICBIS)*, Jan. 2024, pp. 188–194.

[22] L. Chen and Q. Gao, "Application of deep reinforcement learning on automated stock trading," in *Proc. IEEE 10th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Oct. 2019, pp. 29–33.

[23] A. R. Azhikodan, A. G. K. Bhat, and M. V. Jadhav, "Stock trading bot using deep reinforcement learning," in *Proc. 5th ICICSE*, May 2018, pp. 41–49.

[24] P. Liu, Y. Zhang, F. Bao, X. Yao, and C. Zhang, "Multi-type data fusion framework based on deep reinforcement learning for algorithmic trading," *Appl. Intell.*, vol. 53, no. 2, pp. 1683–1706, Jan. 2023.

[25] Y. Liu, D. Mikriukov, O. C. Tjahyadi, G. Li, T. R. Payne, Y. Yue, K. Siddique, and K. L. Man, "Revolutionising financial portfolio management: The non-stationary transformer's fusion of macroeconomic indicators and sentiment analysis in a deep reinforcement learning framework," *Appl. Sci.*, vol. 14, no. 1, p. 274, Dec. 2023.

[26] W. Wang, C. Su, and D. Duxbury, "Investor sentiment and stock returns: Global evidence," *J. Empirical Finance*, vol. 63, pp. 365–391, Sep. 2021.

[27] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. Soman, "NSE stock market prediction using deep-learning models," *Proc. Comput. Sci.*, vol. 132, pp. 1351–1362, Jan. 2018.

[28] Y. Ansari, "Multi-cluster graph (MCG): A novel clustering-based multi-relation graph neural networks for stock price forecasting," *IEEE Access*, vol. 12, pp. 154482–154502, 2024.

[29] J.-H. Park, J.-H. Kim, and J.-H. Huh, "Deep reinforcement learning robots for algorithmic trading: Considering stock market conditions and U.S. interest rates," *IEEE Access*, vol. 12, pp. 20705–20725, 2024.

[30] M. Metghalchi, N. Durmaz, P. Cloninger, and K. Farahbod, "Trading rules and excess returns: Evidence from Turkey," *Int. J. Islamic Middle Eastern Finance Manage.*, vol. 14, no. 4, pp. 713–731, Jul. 2021.

[31] K. Chourmouziadis and P. D. Chatzoglou, "An intelligent short term stock trading fuzzy system for assisting investors in portfolio management," *Expert Syst. Appl.*, vol. 43, pp. 298–311, Jan. 2016.

[32] A. Z. Khan, P. Gupta, and M. K. Mehlawat, "A fuzzy rule-based system for portfolio selection using technical analysis," *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 9, pp. 4861–4875, Sep. 2024.

[33] Y. Kim, W. Ahn, K. J. Oh, and D. Enke, "An intelligent hybrid trading system for discovering trading rules for the futures market using rough sets and genetic algorithms," *Appl. Soft Comput.*, vol. 55, pp. 127–140, Jun. 2017.

[34] M. Bukhari, K. B. Bajwa, S. Gillani, M. Maqsood, M. Y. Durrani, I. Mehmood, H. Ugail, and S. Rho, "An efficient gait recognition method for known and unknown covariate conditions," *IEEE Access*, vol. 9, pp. 6465–6477, 2021.

[35] J. Shin, Y. Kaneko, A. S. M. Miah, N. Hassan, and S. Nishimura, "Anomaly detection in weakly supervised videos using multistage graphs and general deep learning based spatial–temporal feature enhancement," *IEEE Access*, vol. 12, pp. 65213–65227, 2024.

[36] S. Banik, N. Sharma, M. Mangla, S. N. Mohanty, and S. Shitharth, "LSTM based decision support system for swing trading in stock market," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107994.

[37] A. Shah, M. Gor, M. Sagar, and M. Shah, "A stock market trading framework based on deep learning architectures," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 14153–14171, Apr. 2022.

[38] M. Thakur and D. Kumar, "A hybrid financial trading support system using multi-category classifiers and random forest," *Appl. Soft Comput.*, vol. 67, pp. 337–349, Jun. 2018.

[39] Y. Huang, X. Wan, L. Zhang, and X. Lu, "A novel deep reinforcement learning framework with BiLSTM-attention networks for algorithmic trading," *Expert Syst. Appl.*, vol. 240, Apr. 2024, Art. no. 122581.

[40] B. Yang, T. Liang, J. Xiong, and C. Zhong, "Deep reinforcement learning based on transformer and U-Net framework for stock trading," *Knowl.-Based Syst.*, vol. 262, Feb. 2023, Art. no. 110211.

[41] Y. Huang, C. Zhou, K. Cui, and X. Lu, "A multi-agent reinforcement learning framework for optimizing financial trading strategies based on TimesNet," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121502.

[42] D. Kwak, S. Choi, and W. Chang, "Self-attention based deep direct recurrent reinforcement learning with hybrid loss for trading signal generation," *Inf. Sci.*, vol. 623, pp. 592–606, Apr. 2023.

[43] J. Zou, J. Lou, B. Wang, and S. Liu, "A novel deep reinforcement learning based automated stock trading system using cascaded LSTM networks," *Expert Syst. Appl.*, vol. 242, May 2024, Art. no. 122801.

[44] Y. Kwon and Z. Lee, "A hybrid decision support system for adaptive trading strategies: Combining a rule-based expert system with a deep reinforcement learning strategy," *Decis. Support Syst.*, vol. 177, Feb. 2024, Art. no. 114100.

[45] M. Shin, J. Kim, and M. Kim, "Human learning from artificial intelligence: Evidence from human go players' decisions after AlphaGo," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, 2021, pp. 1795–1801.

[46] T. Chong, W.-K. Ng, and V. Liew, "Revisiting the performance of MACD and RSI oscillators," *J. Risk Financial Manage.*, vol. 7, no. 1, pp. 1–12, Feb. 2014.

[47] M. Maitah, P. Procházka, M. Čermák, and K. Šrédl, "Commodity channel index: Evaluation of trading rule of agricultural commodities," *Int. J. Econ. Financial Issues*, vol. 6, no. 1, pp. 176–178, Jan. 2016.

[48] I. Gurrib, "Performance of the average directional index as a market timing tool for the most actively traded USD based currency pairs," *Banks Bank Syst.*, vol. 13, no. 3, pp. 58–70, Aug. 2018.

[49] S.-K. Bormann, "Sentiment indices on financial markets: What do they measure?" Economics Discussion Papers, Germany, Tech. Rep. 2013-58, 2013.

[50] D. E. Allen, M. McAleer, and A. K. Singh, "Daily market news sentiment and stock prices," *Appl. Econ.*, vol. 51, no. 30, pp. 3212–3235, Jun. 2019.

[51] *Daily News Sentiment Index*. Accessed: Aug. 30, 2024. [Online]. Available: https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/

[52] *University of Michigan, University of Michigan: Consumer Sentiment [UMCSENT]*. Federal Reserve Bank of St. Louis. Accessed: Feb. 28, 2025. [Online]. Available: https://fred.stlouisfed.org/series/UMCSENT

[53] D. E. Hirst, L. Koonce, and S. Venkataraman, "Management earnings forecasts: A review and framework," *Accounting Horizons*, vol. 22, no. 3, pp. 315–338, Sep. 2008.

[54] *Alpha Vantage API*. Accessed: Aug. 30, 2024. [Online]. Available: https://www.alphavantage.co/documentation/

[55] M. Magdon-Ismail and A. F. Atiya, "Maximum drawdown," *Risk Mag.*, vol. 17, pp. 99–102, Oct. 2004.

[56] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 387–395.

[57] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[58] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, Nov. 1999, pp. 1057–1063.

[59] A. Ang, "Mean-variance investing," Columbia Bus. School Res. PaperSer., New York, NY, USA, Tech. Rep. 12/49, Aug. 2012, doi: 10.2139/ssrn.2131932.

[60] M. L. de Prado, "Building diversified portfolios that outperform out-of-sample," *J. Portfolio Manag.*, May 2016. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2708678

[61] D. Bailey and M. L. de Prado, "An open-source implementation of the critical-line algorithm for portfolio optimization," *Algorithms*, vol. 6, no. 1, pp. 169–196, Mar. 2013.

[62] S. Sun, R. Wang, and B. An, "Reinforcement learning for quantitative trading," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 3, pp. 1–29, Jan. 2023.

[63] X.-Y. Liu, H. Yang, J. Gao, and C. Wang, "FinRL: Deep reinforcement learning framework to automate trading in quantitative finance," in *Proc. 2nd ACM Int. Conf. AI finance*, Jan. 2021, pp. 1–9.

[64] T. Théate and D. Ernst, "An application of deep reinforcement learning to algorithmic trading," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114632.

[65] S. Carta, A. Corriga, A. Ferreira, A. S. Podda, and D. R. Recupero, "A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning," *Appl. Intell.*, vol. 51, no. 2, pp. 889–905, Feb. 2021.

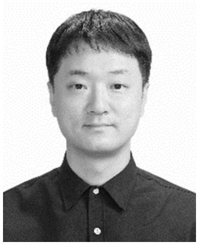
[66] B. Jin, "A mean-VaR based deep reinforcement learning framework for practical algorithmic trading," *IEEE Access*, vol. 11, pp. 28920–28933, 2023.

ASMA SATTAR received the Ph.D. degree in computer science from the University of Pisa, Italy. She focused on deep learning approaches for graphs in context-aware recommendation systems with the University of Pisa. She is currently an Assistant Professor with the Software Engineering Department, Prince Sultan University, Riyadh, Saudi Arabia. She was a Data Scientist with the H&M Research Group, Stockholm, Sweden, where she contributed to deep learning algorithms for prediction tasks involving dynamic multi-relation graphs. Her research interests include machine learning, deep learning for graphs, and their applications in recommender systems.

AMNA SARWAR received the M.Sc. degree in software engineering, with a strong focus on image processing and machine learning-based systems. She is currently a Lecturer with the Department of Computer Science, WAH University, Pakistan. Her research interests include image processing, natural language processing, computer vision, deep learning, and machine learning.

SAIRA GILLANI received the Ph.D. degree in information sciences from the Corvinus University of Budapest, Hungary. She joined the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2016. She was an Assistant Professor with Saudi Electronic University, Jeddah, Saudi Arabia. She is currently a Professor with the University of Central Punjab, Lahore, Pakistan. Previously, she was a Research Scholar with the Corvinno, Technology Transfer Center of Information Technology and Services, Budapest, Hungary, and a Research Associate with the Center of Research in Networks and Telecom (CoReNet), CUST, Pakistan. Her research interests include data sciences, text mining, data mining, machine learning, vehicular networks, mobile edge computing, and the Internet of Things.

MARYAM BUKHARI is currently pursuing the Ph.D. degree in computer science with COMSATS University Islamabad, Attock, Pakistan. Her research interests include computer vision, deep learning and machine learning.



SEUNGMIN RHO is currently an Associate Professor with the Department of Industrial Security, Chung-Ang University. His current research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management, and computational intelligence. He has published 300 papers in refereed journals and conference proceedings in these areas. He has been involved in more than 20 conferences and workshops as various chairs and more than 30 conferences/workshops as a program committee member. He has edited a number of international journal special issues as a guest editor, such as *Multimedia Systems*, *Information Fusion*, and *Engineering Applications of Artificial Intelligence*.

MUHAMMAD FASEEH received the B.S. and M.S. degrees in computer science from COMSATS University Islamabad, Pakistan. He is currently pursuing the Ph.D. degree with Jeju National University, Republic of Korea, showcasing his commitment to academic excellence. He is currently a Dedicated Computer Scientist. He brings a wealth of experience from both the software development industry and academia, highlighting his versatility. His research interests include cutting-edge fields, such as AI-based intelligent systems, computer vision, data science, big data analytics, machine learning, and deep learning, reflecting his passion for advancing technology and knowledge in these domains.