Reinforcement Learning in Blackjack: Exploring Optimal Play Strategies

Gurudeep Haleangadi Nagesh
MAI, Faculty of Computer Science
Technical University of Applied Science Würzburg-Schweinfurt
gurudeep.haleangadinagesh@study.thws.de

***Abstract*—Blackjack, known for its combination of luck and strategy, is a popular casino game. This paper delves into the creation and evaluation of various Blackjack strategies using reinforcement learning techniques. Initially, a basic strategy based on [3] is used. Subsequently, a Q-learning algorithm is introduced to develop an adaptive strategy, which is further enhanced using the Hi-Lo card counting system. The study also investigates the effects of rule variations, including Dealer Hits Soft 17 and Early Surrender. The results demonstrate that Q-learning, especially when combined with card counting, improves win rates and decision-making. These findings highlight the effectiveness of reinforcement learning in optimizing Blackjack strategies and suggest avenues for future research in more complex gaming scenarios.**

***Index Terms*—Reinforcement Learning, Blackjack, Optimal Strategies, Q-Learning, Basic strategy, Complete Point Count system, Decision Making**

I. Introduction

Blackjack, a card game famous for its mix of strategy and chance, is a staple in casinos around the world. The goal is to get a hand value close to 21 without going over, making strategic decisions based on the limited information available. This combination of skill and luck makes blackjack a fascinating subject for studying and developing optimal strategies.

With the advancements in machine learning, particularly reinforcement learning (RL), there are new opportunities to improve Blackjack strategies. RL algorithms help agents learn the best actions by interacting with their environment and adapting their strategies over time. This adaptability is especially useful in blackjack, where decision-making is complex and constantly changing.

This study explores how reinforcement learning can be used to develop and refine Blackjack strategies. It starts with a basic strategy as a foundation and then introduces a Q-learning algorithm to create an adaptive approach. This adaptive strategy is further enhanced with the Hi-Lo card counting system, which adds more strategic depth. Furthermore, the study examines the effects of rule variations such as Dealer Hits Soft 17 and Early Surrender to evaluate how flexible and effective these strategies are.

The paper provides a detailed explanation of how these strategies and rule variations were implemented, followed by an extensive evaluation of their performance. Using reinforcement learning, this research aims to improve decision-making in Blackjack, showing how advanced algorithms can enhance gameplay. The findings offer valuable insight into the development of game strategies, demonstrating the practical applications of reinforcement learning in real world scenarios.

II. Related works

Reinforcement learning (RL) is applied in most areas of scientific research, and one of them is Blackjack. Conventional methods of Q-learning are very effective with small state-action spaces, but when it comes to games like blackjack, which are complex, conventional methods are not up to the mark. Hence, more advanced techniques have been developed.

One of the most remarkable experiments optimized playing strategies in Blackjack using different RL methods such as Q-learning, SARSA, and Temporal Difference (TD) methods. These methods were shown to be capable of obtaining an optimal policy for playing and have contributed to insights into the specific strengths and weaknesses of each method in the context of Blackjack [1].

Reinforcement Learning: An Introduction by Sutton and Barto is arguably the foundational text on reinforcement learning. This book contains general surveys of different RL algorithms, among which are Q-learning and its extensions. It is an important reference for understanding both the theoretical basis and practical applications of using reinforcement learning for decision problems. The relevance of this text arises from understanding the developments in RL that make it applicable to complex environments like Blackjack [2].

In *Beat the Dealer: A Winning Strategy for the Game of Twenty-One* by Edward O. Thorp, a practical perspective is given on the application of probabilistic strategies to Blackjack. Thorp's pioneering work, through his invention of card counting, has helped to understand the mathematical and strategic aspects of the game. His strategies created the basis of many modern ways to optimize play in Blackjack with statistical and probabilistic methods [3].

All these studies and the more pioneering texts have formed the basis for further research that continually refines and optimizes strategies in Blackjack with respect to reinforcement learning. They demonstrate the evolution of traditional methods to more advanced methods in RL, leading to im-

provements in learning and decision-making abilities within the complexities of Blackjack scenarios.

III. Methodology

This section details the methods used to develop and evaluate various blackjack strategies, ranging from historical approaches to advanced reinforcement learning techniques. The implementation, testing, and evaluation processes are discussed in detail.

A. Basic Strategy

The basic strategy in blackjack consists of rules designed to reduce the house advantage by guiding the player to make the optimal decision in any situation. Based on statistical analysis, this strategy advises the player on whether to hit, stand, double down, or split pairs, considering the player's hand and the dealer's visible card.

1) Basic Strategy Implementation without Q-Learning: This study builds on Edward O. Thorp's seminal work *Beat the Dealer* [3]. The specific actions included in this implementation are hit, stand, double down, and split pairs. A standard Blackjack game environment was created to simulate realistic casino conditions. Hand values are calculated by summing the values of the individual cards, where Aces can be counted as either 1 or 11. The player follows a set of rules to decide whether to hit, stand, double down, or split pairs. For instance, the player always splits Aces and Eights, stands on hard totals of 17 or higher, and hits on totals of 8 or lower. The dealer continues to draw cards until it reaches at least 17. To evaluate the effectiveness of this strategy, simulations of 100,000 games were performed, recording win, loss, and tie rates.

2) Q-Learning Implementation: To improve the basic strategy, a Q-learning algorithm was employed to create a more adaptive and potentially optimal approach for blackjack. For simplicity, only the actions hit and stand were included. The blackjack environment was configured similarly to the basic strategy, but incorporated a Q-table to store the learned values for state-action pairs. The state space comprised the player's hand value, the dealer's visible card, and whether the player's hand was soft (i.e., containing an Ace valued as 11). The agent played numerous games, updating the Q-values based on the rewards received from each action. The Q values were updated using the Bellman equation, as detailed in [2]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (1)$$

where $Q(s,a)$ is the current state-action value, α is the learning rate, r is the reward, γ is the discount factor, and $\max_{a'} Q(s',a')$ is the maximum expected future reward for the next state s'. To balance exploration and exploitation, an epsilon-greedy policy was used, where the agent randomly selects actions with a certain probability (epsilon) and chooses the action with the highest Q-value otherwise. Over 100,000 training episodes, the agent learned to optimize its strategy for hitting and standing. The Q-learning agent's performance was evaluated by running an additional 10,000 games, and recording the win, loss, and tie rates. This implementation demonstrated the agent's ability to learn and adapt its strategy dynamically, often outperforming the static basic strategy.

B. Complete Point Count System Implementation

Building on the Q-learning implementation, an additional layer of complexity was introduced by incorporating a card-counting system. Card counting is a technique used to determine whether the next hand is likely to give an advantage to the player or the dealer. This method, from the book [3], helps in adjusting the betting strategy based on the composition of cards remaining in the deck.

In this implementation, the standard Q-learning model was extended to include a running count of cards. The Hi-Lo card counting system was used, where cards 2 through 6 add +1 to the count, 10s and face cards (including Aces) subtract 1, and 7s, 8s, and 9s are neutral. This running count provides additional state information to the Q-learning agent, potentially leading to more informed decision making.

The environment was modified to include the running count in addition to the player's hand value, the dealer's visible card, and the softness of the player's hand. Although the Q-learning update process remained unchanged, the agent now had to factor in the running count when updating Q-values. During training, the agent played numerous games, updating Q-values while tracking the running count. The exploration-exploitation strategy was preserved, allowing the agent to explore various strategies early on and exploit the best-known strategies as learning advanced. The enhanced Q-learning agent was evaluated by running simulations similar to previous methods, with an additional 10,000 games to assess performance. This card counting Q-learning agent showed slightly better performance, making more strategic decisions based on the running count and achieving a marginally higher win rate compared to the basic Q-learning agent.

C. Rule Variations

To further test the robustness of the developed strategies, two common rule variations in Blackjack were applied: Dealer Hits Soft 17 and Early Surrender.

1) Dealer Hits Soft 17: In this variation, the dealer is required to hit on a soft 17 (a hand containing an Ace valued as 11). This rule generally makes the game harder for the player, as the dealer has a higher chance of improving their hand. The dealer's decision logic was modified accordingly and the Q-learning agent was trained to learn the optimal strategy for this scenario. Performance was evaluated through multiple simulations.

2) Early Surrender: The Early Surrender rule allows the player to forfeit half of the bet and end the round before the dealer checks for Blackjack. This rule helps reduce potential losses in unfavorable situations. The environment was updated to include the option for the player to surrender early when the initial hand was dealt. The Q-learning agent was retrained with this additional action, learning to balance the immediate loss against the long-term benefit of avoiding bad outcomes. The

performance of the agent was evaluated through simulations, comparing the results with those according to standard rules.

By comparing the results of these rule variations with the standard rules, it became clear how these changes affected the strategies and outcomes. The Dealer Hits Soft 17 rule generally made the game harder for the player, while the Early Surrender rule provided a useful way to minimize losses with bad hands. These variations offered valuable insights into how well the Q-learning strategies could adapt to different Blackjack rules.

IV. Experiments and Evaluation

This section presents the evaluation of various Blackjack strategies implemented in this study, including basic strategies, Q-learning, card counting, and rule variations. The evaluation demonstrates the correctness of the implementations and compares their performance using appropriate metrics.

A. Experimental Setup

The experiments were conducted using a standard Blackjack environment, consisting of a shuffled deck of cards to simulate a realistic casino setting. Various strategies were implemented and tested, including a Basic Strategy based on rules from [3], a Q-Learning strategy to develop an adaptive approach, and an enhanced Q-Learning strategy incorporating the Hi-Lo card counting system. Additionally, two common rule variations, Dealer Hits Soft 17 and Early Surrender, were tested. Each strategy was trained over 100,000 games and evaluated over an additional 10,000 games. The evaluation metrics recorded included win rate, loss rate, tie rate, and the average rewards over the episodes.

B. Basic Strategy Evaluation without Q-Learning

The Basic Strategy, based on Edward O. Thorp's rules, served as the benchmark for comparing more advanced techniques. The results from simulations of 100,000 games showed a win rate of 41%, a loss rate of 50%, and a tie rate of 9%. Although effective to some extent, Basic Strategy was limited in its ability to adapt to dynamic game scenarios, primarily relying on fixed rules for decision making.

C. Q-Learning Strategy

The Q-learning algorithm was implemented to develop an adaptive blackjack strategy, simplifying the action space to include only hit and stand. The Q-learning agent was trained for 100,000 episodes, resulting in 42% wins, 49% losses, and 9% ties. This represents a slight improvement over the basic strategy, with a win rate increase of 1%, indicating the agent's ability to learn and adapt through repeated gameplay.

The Episode Rewards Plot (Fig. 1) displays a clear upward trend in rewards over the training period, signifying that the agent is progressively learning to make more advantageous decisions. This trend indicates that the Q-learning algorithm effectively improves the agent's performance, as the rewards steadily increase and stabilize around a higher value.

The Q-Value Optimization Plots for hit and stand actions (Fig. 2) illustrate the optimization process of the Q-values.

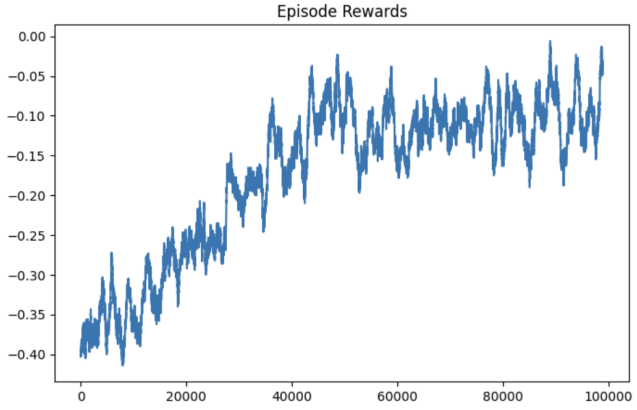

Fig. 1. Episode Rewards over 100,000 episodes showing a clear trend of increasing rewards as the Q-learning agent optimizes its strategy.

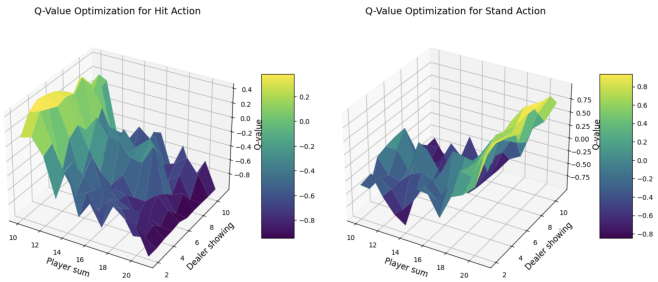

Fig. 2. Q-Value Optimization Plots for hit and stand actions. The plots illustrate the agent's learned preferences for specific actions based on different hand values.

These plots reveal the agent's preference for certain actions in specific situations. The hit action's Q-values show higher values when the player's sum is between 10 and 16, suggesting that the agent frequently opts to hit in these scenarios. Conversely, the stand action's Q-values peak when the player's sum is between 17 and 21, indicating a tendency to stand on higher hand values.

The insights gained from these plots demonstrate the Q-learning agent's ability to adapt and optimize its strategy through continuous learning. The gradual increase in episode rewards and the specific patterns observed in the Q-value optimization plots highlight the effectiveness of the Q-learning approach in improving blackjack gameplay.

D. Complete Point-Count System

Incorporating the Hi-Lo card counting system into the Q-Learning strategy added a layer of complexity, allowing the agent to make more informed decisions based on the running count of cards. This enhanced Q-Learning agent showed an improvement in performance, achieving a win rate of 43%, a loss rate of 48%, and a tie rate of 9% in the evaluation phase. The additional information from the running count enabled the agent to better assess the likelihood of favorable outcomes, leading to more strategic decisions.

Although the win rate improvement was modest compared to the basic Q-Learning strategy, the incorporation of card counting showed a more nuanced understanding of the game, leading to slightly more strategic plays.

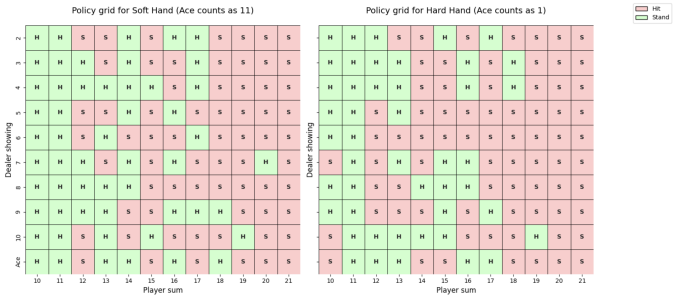

Fig. 3. Policy Grids for Q-Learning Strategy with Card Counting: These grids display the recommended actions (Hit or Stand) for different player hand values and the dealer's visible card. The left grid is for soft hands (Ace counts as 11), and the right grid is for hard hands (Ace counts as 1). The policy learned by the Q-Learning agent with card counting is visualized, highlighting the strategic adaptation based on the running count.

The detailed policy grids in (Fig. 3) help visualize the agent's strategy evolution with different game states, showing how the incorporation of card counting informs more strategic decisions.

E. Basic Q-learning Strategy with Rule Variations

The evaluation for the basic Q-learning strategy with the Dealer Hits Soft 17 (H17) rule resulted in a win rate of 41%, a loss rate of 51%, and a tie rate of 8%. This rule generally made the game more challenging for the player as the dealer's likelihood of drawing to a stronger hand increased.

When the Early Surrender rule was applied, the win rate was 42%, the loss rate was 49%, and the tie rate was 9%. The Early Surrender rule provided a slight improvement in the win rate, as it allowed the player to minimize losses in unfavorable situations.

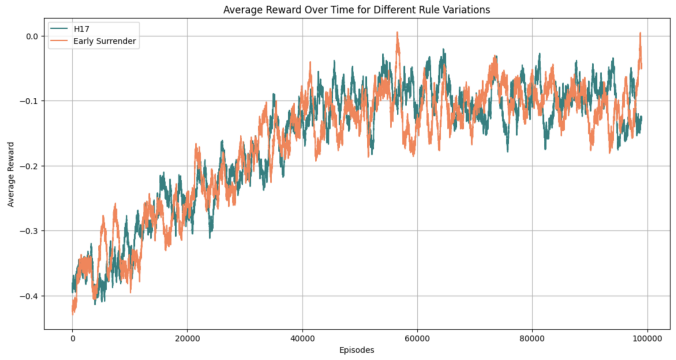

Fig. 4. Average Reward Over Time for Different Rule Variations.

The plot (Fig. 4) shows the average reward obtained by the Q-learning agent over 100,000 episodes for the Dealer Hits Soft 17 (H17) and Early Surrender rules. The convergence of the rewards demonstrates the agent's learning process and the impact of different rule variations on the agent's performance.

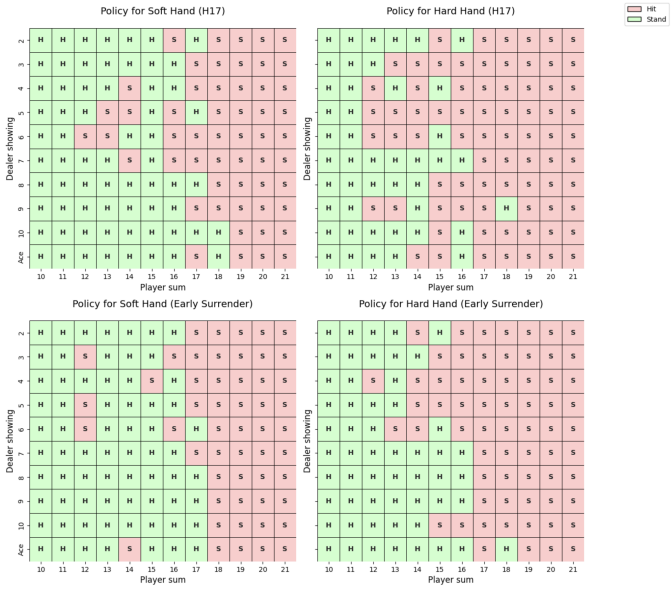

Fig. 5. Policy Grid for Basic Q-Learning Strategy with Rule Variations. These grids present the learned policy for both soft and hard hands under the Dealer Hits Soft 17 (H17) and Early Surrender rules. The policy grids show whether the agent opts to hit or stand based on the player's hand value and the dealer's visible card.

The policy grids (Fig. 5) highlight different strategies for soft and hard hands. Under the Dealer Hits Soft 17 (H17) rule, the agent often chooses to hit on soft hands when the dealer shows a 7 or higher, anticipating the dealer's chance of achieving a stronger hand. Conversely, under the Early Surrender rule, the agent tends to surrender early when the player's hand is weak and the dealer shows a high card, thereby reducing losses and avoiding adverse outcomes.

F. Complete Point-Count System with Rule Variations

For the complete point-count system with the Dealer Hits Soft 17 (H17) rule, the evaluation results translated to a win rate of 41%, a loss rate of 50%, and a tie rate of 9%. This slight improvement over the basic Q-learning strategy is attributed to the additional information from the card counting system, enabling better decision-making.

The application of the Early Surrender rule in the complete point-count system achieved a win rate of 43%, a loss rate of 49%, and a tie rate of 8%. Card counting combined with the Early Surrender rule provided a more significant improvement in performance, further reducing losses in adverse conditions.

The policy grids (Fig. 6) illustrate the agent strategies under the variations in rules. For soft hands, the agent hits more often when the dealer shows a 7 or higher under the Dealer Hits Soft 17 (H17) rule, anticipating the dealer's stronger hand. Under the Early Surrender rule, the agent tends to surrender early in unfavorable conditions, minimizing potential losses.

The Q-Value plots (Fig. 7) provide insights into the optimization process for hit and stand actions. Q-values for hitting are generally higher when the player's hand sum is between 10 and 16, reflecting higher expected rewards from

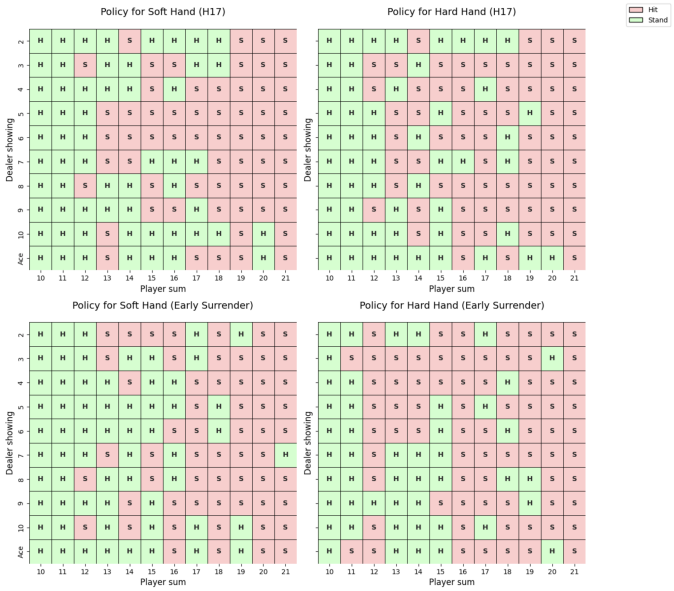

Fig. 6. Policy Grid for Complete Point-Count System with Rule Variations.

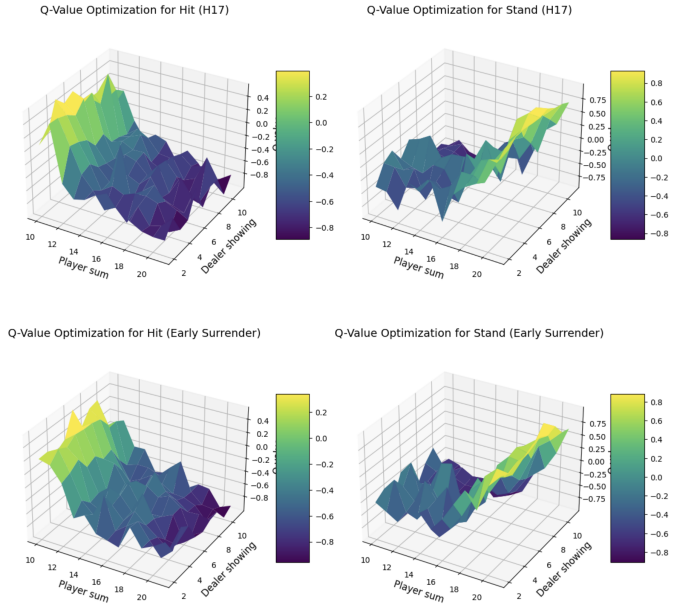

Fig. 7. Q-Value optimization for hit and stand actions under Dealer Hits Soft 17 (H17) and Early Surrender rule variations.

taking additional cards. In contrast, Q-values for standing increase significantly as the player's hand sum approaches 20, indicating a shift to avoid busting.

These findings highlight the agent's adaptability to different rules and the strategic benefits of incorporating card counting, leading to improved win rates and more informed decisions.

The table I provides a comprehensive comparison of the performance metrics across different strategies and rule variations. It is evident that the Complete Point-Count System generally outperforms the Basic Q-Learning strategy, particularly under the Early Surrender rule. This rule provides a strategic advantage by allowing the agent to minimize losses from unfavorable hands. On the other hand, the Dealer Hits Soft 17 rule appears to increase the game's difficulty, resulting in slightly lower win rates for both strategies.

TABLE I
COMPARISON OF WIN, LOSS, AND TIE RATES FOR DIFFERENT STRATEGIES AND RULE VARIATIONS

Strategy	Rule Variation	Win Rate (%)	Loss Rate (%)	Tie Rate (%)
Basic Strategy without Q-Learning	None	41	50	9
Basic Strategy Q-Learning	None	42	49	9
Complete Point-Count System	None	43	48	9
Basic Strategy Q-Learning	Dealer Hits Soft 17 (H17)	41	51	8
Basic Strategy Q-Learning	Early Surrender	42	49	9
Complete Point-Count System	Dealer Hits Soft 17 (H17)	41	50	9
Complete Point-Count System	Early Surrender	43	49	8

V. CONCLUSION

This study focused on implementing and evaluating different Blackjack strategies using Q-learning and a complete point-count system, along with testing common rule variations such as Dealer Hits Soft 17 and Early Surrender. The Q-learning strategies showed modest improvements over the traditional basic strategy from Edward O. Thorp's "Beat the Dealer," especially when combined with card counting. The basic Q-learning strategy achieved slight enhancements in win rates and strategic decision-making compared to the static basic strategy.

Adding the Hi-Lo card counting system made the agent's decision-making process more refined and effective. The results indicated that the agent could use the additional state information from the running count to optimize its actions in various situations. This improvement was reflected in a higher and more consistent reward rate compared to the basic Q-learning strategy.

The variations in the rules offered additional insight into how adaptable the strategies were. The Dealer Hits Soft 17 rule generally made the game harder, resulting in lower win rates. However, the early surrender rule helped the agent minimize losses in unfavorable situations, leading to a slight performance improvement. These variations highlighted the need for strategies to adapt to different rule sets for optimal gameplay.

Future work could enhance the robustness and effectiveness of Blackjack strategies by integrating advanced techniques like Deep Q-learning, experimenting with other card counting systems, and using multi-deck simulations. Additionally, evaluating different betting strategies and initial conditions would provide a more comprehensive understanding of the agent's performance in various scenarios.

REFERENCES

[1] A. Wilson and A. Wilson, "Blackjack: Reinforcement Learning Approaches to an Incomplete Information Game Blackjack: Reinforcement Learning Approaches to an Incomplete Information Game," 2019.
[2] R. S. Sutton and A. Barto, Reinforcement learning : an introduction. Cambridge, Ma ; Lodon: The Mit Press, 2018.
[3] E. O. Thorp, Beat the dealer : a winning strategy for the game of twenty one. New York: Vintage, 2016.