JURNAL FASILKOM
Volume 15 No. 3 | Desember 2025: 523-532

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Perbandingan Metode Q-Learning Dan SARSA Dalam Optimasi Prediksi Tren Saham Pada Indeks Harga Saham Gabungan (IDX)

Muhammad Affarel[1], Fikri Ikhsan Ramadhan[2] , Nugroho Aldi Prayoga[3]
[1,2,3]Teknologi Informasi, Teknik Informatika, Universitas Bina Sarana Informatika
[1]17230674@bsi.ac.id, [2]17230691@bsi.ac.id, [3]17230669@bsi.ac.id

Abstract

This study evaluates the performance of Reinforcement Learning algorithms, namely Q-Learning and SARSA, in generating automated trading strategies for the Indonesian stock market. The research is motivated by the high volatility and uncertainty of stock price movements, which require adaptive decision-making methods. The dataset consists of historical stock price data from five companies listed on the Indonesia Stock Exchange (BBCA, BBRI, TLKM, UNVR, and ASII), obtained using the yfinance library and simulated through 1,000 trading episodes. Performance evaluation was conducted based on reward trends, equity curve patterns, and final performance statistics to assess learning effectiveness under different market conditions. The results indicate that Q-Learning performs better on stocks with strong price momentum due to its more aggressive exploration behavior, while SARSA provides more stable performance in highly volatile markets owing to its conservative on-policy approach. Overall, neither algorithm demonstrates absolute dominance; instead, each offers distinct advantages depending on stock characteristics and risk profiles. These findings highlight the potential of Reinforcement Learning for the development of algorithmic trading strategies in the Indonesian stock market.

Keywords: Reinforcement Learning, Q-Learning, SARSA, Trading, Optimization

Abstrak

Penelitian ini bertujuan mengevaluasi kinerja algoritma Reinforcement Learning, yaitu Q-Learning dan SARSA, dalam membentuk strategi trading otomatis pada pasar saham Indonesia. Latar belakang penelitian ini didasarkan pada tingginya dinamika dan ketidakpastian pergerakan harga saham yang menuntut metode pengambilan keputusan adaptif. Data yang digunakan berupa data historis harga saham lima emiten Bursa Efek Indonesia (BBCA, BBRI, TLKM, UNVR, dan ASII) yang diperoleh melalui pustaka *yfinance*, kemudian disimulasikan dalam proses trading sebanyak 1.000 episode. Evaluasi kinerja dilakukan berdasarkan tren reward, pola equity curve, serta statistik performa akhir untuk menilai efektivitas pembelajaran pada kondisi pasar yang berbeda. Hasil penelitian menunjukkan bahwa Q-Learning cenderung unggul pada saham dengan momentum harga yang kuat karena karakter eksplorasinya yang lebih agresif, sedangkan SARSA memberikan performa yang lebih stabil pada saham dengan volatilitas tinggi berkat pendekatan on-policy yang lebih konservatif. Secara keseluruhan, tidak terdapat algoritma yang mendominasi secara absolut, namun masing-masing metode menunjukkan keunggulan yang bergantung pada karakteristik saham dan profil risiko strategi. Temuan ini menegaskan potensi penerapan Reinforcement Learning dalam pengembangan algorithmic trading di pasar saham Indonesia.

Kata kunci: Reinforcement Learning, Q-Learning, SARSA, Trading, Optimization

1. Pendahuluan

Pasar modal Indonesia dalam satu dekade terakhir telah mengalami peningkatan aktivitas transaksi yang signifikan, ditandai dengan meningkatnya partisipasi investor ritel serta penetrasi teknologi digital dalam perdagangan saham. Fluktuasi harga saham yang dinamis pada Bursa Efek Indonesia (IDX) menunjukkan bahwa proses pengambilan keputusan investasi kini tidak hanya ditentukan oleh intuisi, tetapi juga oleh kemampuan analisis berbasis data yang akurat. Di tengah perubahan tersebut, muncul kebutuhan akan sistem prediksi dan pengambilan keputusan otomatis yang mampu beradaptasi terhadap pola pasar yang kompleks dan nonlinier [1].

Seiring dengan berkembangnya teknologi kecerdasan buatan, pendekatan *machine learning* dan khususnya *reinforcement learning* mulai banyak diterapkan dalam bidang keuangan. Pendekatan ini memungkinkan agen untuk belajar dari pengalaman dengan tujuan memaksimalkan keuntungan kumulatif berdasarkan interaksi terhadap lingkungan pasar [2]. Berbeda dengan metode statistik konvensional seperti regresi linear atau ARIMA yang mengandalkan pemodelan eksplisit, *reinforcement learning* mampu beradaptasi terhadap dinamika pasar yang tidak stasioner dan memiliki tingkat volatilitas tinggi [3]. Hal ini menjadikan pendekatan tersebut relevan untuk menganalisis perilaku saham di IDX yang kerap dipengaruhi oleh faktor makroekonomi dan psikologis investor.

Dua algoritma utama yang banyak digunakan dalam konteks *reinforcement learning* adalah Q-Learning dan SARSA, keduanya termasuk dalam kategori *value-based methods*. Q-Learning beroperasi dengan strategi *off-policy*, yakni pembaruan nilai fungsi aksi dilakukan berdasarkan aksi terbaik yang mungkin diambil, sedangkan SARSA bersifat *on-policy*, memperbarui nilai berdasarkan aksi yang benar-benar dieksekusi.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Perbedaan ini menjadikan Q-Learning lebih agresif dalam eksplorasi strategi baru, sementara SARSA cenderung konservatif dan stabil dalam kondisi pasar yang fluktuatif. Beberapa penelitian menunjukkan bahwa perbedaan kecil dalam mekanisme pembaruan nilai ini dapat menghasilkan perbedaan signifikan terhadap kinerja model dalam jangka panjang [4].

Penelitian-penelitian sebelumnya telah menguji efektivitas algoritma *deep reinforcement learning* dalam perdagangan saham di berbagai pasar global. Misalnya, penelitian dalam [5] berhasil menerapkan *deep Q-learning* untuk perdagangan saham di IDX dan menemukan bahwa algoritma tersebut mampu meningkatkan akurasi prediksi arah harga hingga 7% dibandingkan pendekatan konvensional. Studi lain dari [6] menggabungkan analisis sentimen dengan *reinforcement learning* untuk memperkuat strategi pembelian dan penjualan, menunjukkan peningkatan performa dalam kondisi pasar bergejolak. Hasil-hasil tersebut memberikan dasar yang kuat untuk menerapkan metode serupa dalam konteks saham-saham unggulan di Indonesia.

Namun, sebagian besar penelitian terdahulu lebih berfokus pada model berbasis *deep learning* tanpa mengevaluasi kinerja metode *value-based learning* sederhana seperti Q-Learning dan SARSA secara langsung pada pasar domestik. Padahal, algoritma sederhana sering kali lebih mudah dioptimalkan, membutuhkan sumber daya komputasi lebih rendah, serta lebih transparan dalam interpretasi hasilnya [7]. Kekosongan ini membuka ruang penelitian baru untuk menganalisis sejauh mana Q-Learning dan SARSA dapat memberikan hasil yang kompetitif terhadap pendekatan *deep reinforcement learning* di pasar saham Indonesia yang memiliki karakteristik fluktuatif dan asimetris.

Untuk mengisi celah penelitian tersebut, penelitian ini dirancang dengan fokus pada perbandingan performa algoritma Q-Learning dan SARSA dalam simulasi perdagangan saham pada indeks harga saham gabungan (IDX). Penelitian ini menggunakan data harga penutupan harian dari lima saham utama BBCA, BBRI, TLKM, UNVR, dan ASII pada periode 2020–2024 sebagai representasi stabilitas dan volatilitas pasar. Melalui pendekatan eksperimental, kedua algoritma diuji pada parameter yang identik untuk memastikan evaluasi yang adil dan obyektif terhadap efektivitas strategi pembelajarannya [8].

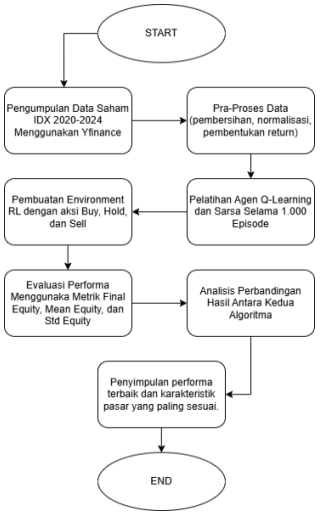
Secara metodologis, penelitian ini berupaya mengukur tiga indikator utama, yaitu *Final Equity*, *Mean Equity*, dan *Standard Deviation of Equity*, yang digunakan untuk menilai performa serta konsistensi masing-masing algoritma. Analisis hasil simulasi juga mencakup laju konvergensi reward antar episode pelatihan sebagai tolok ukur kestabilan pembelajaran. Dengan membandingkan kedua algoritma pada berbagai kondisi saham, diharapkan dapat diketahui model mana yang lebih unggul dalam mengoptimalkan strategi perdagangan berbasis pembelajaran agen.

2. Metode Penelitian

Metode penelitian ini dirancang untuk menggambarkan alur eksperimen secara sistematis, mulai dari pengumpulan data historis saham hingga evaluasi performa model. Setiap tahapan disusun untuk merefleksikan dinamika pasar yang bersifat nonlinier, sehingga algoritma mampu mempelajari pola harga yang terus berubah. Kerangka reinforcement learning yang digunakan mengikuti konsep interaksi agent–environment, di mana agen menentukan aksi berdasarkan estimasi nilai dan menerima umpan balik berupa reward [5]. Pendekatan ini dipilih karena efektif dalam membangun sistem pengambilan keputusan adaptif pada lingkungan keuangan yang memiliki ketidakpastian tinggi [9].

2.1. Alur Penelitian

Alur penelitian disusun dalam tujuh tahapan utama yang merepresentasikan proses eksperimen reinforcement learning dari pengumpulan data hingga evaluasi akhir. Proses dimulai dari pengumpulan dan pra-pemrosesan data harga saham historis untuk membentuk environment pembelajaran yang stabil, dilanjutkan dengan perancangan environment trading melalui pendefinisian state, action, dan fungsi reward. Selanjutnya, algoritma Q-Learning dan SARSA diimplementasikan dan dilatih melalui proses pembelajaran berulang agar agen mampu beradaptasi terhadap pola harga yang fluktuatif. Evaluasi kinerja dilakukan menggunakan tren reward, equity curve, dan statistik performa untuk menilai efektivitas strategi yang dihasilkan. Seluruh tahapan ini mengacu pada kerangka kerja reinforcement learning modern yang umum digunakan dalam sistem pengambilan keputusan otomatis berbasis data pasar [9].



Gambar 1. Research Flowchart

P-ISSN : 2089-3353
E-ISSN : 2808-9162

A. Pengumpulan Data Saham IDX(2020-2024)

Dataset penelitian diperoleh menggunakan pustaka yfinance yang menyediakan data historis saham harian untuk lima emiten Bursa Efek Indonesia, yaitu BBCA, BBRI, TLKM, UNVR, dan ASII. Data mencakup harga pembukaan, penutupan, tertinggi, terendah, serta volume transaksi selama periode Januari 2020 hingga Desember 2024, yang merepresentasikan fase volatilitas tinggi, pemulihan pascapandemi, dan stabilisasi pasar. Seluruh data disusun dalam format time-series yang konsisten dan digunakan sebagai dasar pembentukan environment trading, sehingga agen menerima informasi pasar yang terstruktur untuk mendukung proses pembelajaran sekuensial berbasis reinforcement learning [12].

B. Pengumpulan dan Pra-Proses Data

Tahap pengumpulan dan pra-proses data dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan sebagai masukan bagi model Q-Learning dan SARSA. Proses ini meliputi pembersihan data untuk menghilangkan nilai hilang dan anomali harga, serta normalisasi menggunakan MinMaxScaler agar skala harga seragam dan mempercepat konvergensi pelatihan. Perubahan harga harian kemudian direpresentasikan dalam bentuk return menggunakan persamaan:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Dengan r_t sebagai *state representation* dalam e*nvironment* pelatihan P_t sebagai harga penutupan pada hari ke-t dan P_{t-1} sebagai harga penutupan hari sebelumnya. Representasi berbasis return dipilih karena mampu menangkap dinamika relatif pergerakan harga tanpa dipengaruhi skala absolut antar saham. Untuk meningkatkan kualitas state, dilakukan pula proses windowing, di mana satu state terdiri atas rangkaian return beberapa hari sebelumnya untuk membantu agen mengenali pola jangka pendek yang relevan dalam pengambilan keputusan trading. Tahapan pra-proses ini memungkinkan pembentukan observasi yang stabil, informatif, dan sesuai dengan karakteristik pasar berbasis time-series [7].

C. Pembentukan Environment RL

Environment reinforcement learning dibangun berdasarkan kerangka Markov Decision Process (MDP) yang merepresentasikan keputusan trading melalui komponen state, action, dan reward. State dibentuk dari rangkaian return hasil pra-proses yang dilengkapi indikator tren sederhana, sementara aksi dibatasi pada buy, hold, dan sell. Reward ditentukan berdasarkan perubahan nilai portofolio agar agen belajar memaksimalkan profit kumulatif. Transisi antar-state mengikuti dinamika data historis harian, sehingga lingkungan pembelajaran bersifat stabil, terukur, dan sesuai dengan karakteristik pasar keuangan berbasis time-series [10].

D. Pelatihan Agen Q-Learning dan SARSA

Parameter pembelajaran memainkan peran fundamental dalam menentukan kualitas keputusan yang dihasilkan oleh agen Reinforcement Learning, karena setiap kombinasi nilai yang digunakan akan memengaruhi stabilitas proses pelatihan, pola eksplorasi, serta kecepatan konvergensi menuju kebijakan optimal. Pada konteks pasar saham yang bersifat non-stasioner, pemilihan parameter yang tepat menjadi semakin krusial untuk memastikan model mampu beradaptasi di bawah dinamika harga yang berubah-ubah

Rumus pembaruan nilai Q pada Q-Learning diberikan oleh:

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

Sedangkan pembaruan nilai Q pada SARSA didefinisikan sebagai:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s', a') - Q(s,a)]$$

Pada kedua rumus tersebut, *s* merepresentasikan state atau kondisi pasar pada saat tertentu, sedangkan *a* adalah action yang dipilih agen, yaitu aksi *buy*, *hold*, atau *sell* dalam lingkungan trading. State berikutnya dilambangkan sebagai *s'*, dan *a'* merupakan aksi yang diambil pada state tersebut. Parameter α (learning rate) mengatur seberapa besar pembaruan nilai dilakukan setiap kali agen menerima reward, γ (discount factor) menentukan sejauh mana reward jangka panjang dipertimbangkan dan r adalah reward langsung dari transisi (s')(s,a)→(s'). Operator $\max_{a'}$ pada Q-Learning menegaskan sifat off-policy, sedangkan SARSA menggunakan aksi aktual *a'* sehingga bersifat on-policy. Selain itu, parameter ε pada mekanisme *epsilon-greedy* mengatur keseimbangan antara eksplorasi dan eksploitasi selama 1.000 episode pelatihan. Pemilihan parameter ini berpengaruh langsung pada efektivitas strategi trading dalam lingkungan pasar yang dinamis dan penuh ketidakpastian [15].

E. Evaluasi Performa

Evaluasi performa dilakukan untuk menilai kualitas strategi trading Q-Learning dan SARSA menggunakan tiga metrik utama, yaitu Final Equity, Mean Equity, dan Standard Deviation of Equity. Final Equity merepresentasikan nilai akhir portofolio, Mean Equity menggambarkan kinerja rata-rata selama simulasi, sedangkan Standard Deviation mengukur stabilitas strategi melalui fluktuasi nilai portofolio. Pendekatan ini mengikuti standar evaluasi pada sistem trading berbasis reinforcement learning yang menekankan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

keseimbangan antara profitabilitas dan risiko dalam lingkungan pasar yang dinamis [14].

F. Analisis Perbandingan

Analisis perbandingan bertujuan mengevaluasi perbedaan karakteristik Q-Learning dan SARSA berdasarkan pola pembelajaran, stabilitas kebijakan, serta respons terhadap dinamika lingkungan pasar yang bersifat stokastik. Evaluasi difokuskan pada tren reward, sensitivitas eksplorasi–eksploitasi, dan kecenderungan konvergensi kebijakan, sehingga memberikan dasar teoretis untuk memahami perbedaan perilaku algoritma off-policy dan on-policy dalam konteks keuangan [13].

G. Penyimpulan Hasil

Penyimpulan hasil dilakukan dengan mengintegrasikan proses pembentukan environment, pelatihan agen, dan evaluasi performa untuk memperoleh gambaran komprehensif mengenai efektivitas algoritma reinforcement learning. Fokus utama tahap ini adalah memahami kecenderungan konvergensi kebijakan, stabilitas strategi, serta sensitivitas model terhadap dinamika pasar, dengan menekankan pengaruh mekanisme update dan sifat on-policy maupun off-policy terhadap robustnes strategi dalam lingkungan finansial yang volatil [13].

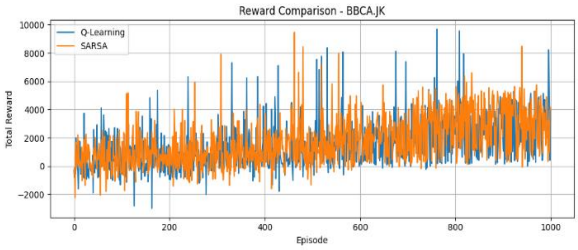
3. Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen dan analisis kinerja algoritma Reinforcement Learning Q-Learning dan SARSA pada pengambilan keputusan trading saham di pasar Indonesia. Pengujian dilakukan menggunakan data historis lima saham Bursa Efek Indonesia (BBCA, BBRI, TLKM, UNVR, dan ASII) periode 2020–2024 dalam lingkungan agen trading dengan aksi buy, hold, dan sell. Kedua algoritma dievaluasi berdasarkan tren reward, perkembangan nilai portofolio melalui grafik equity, serta statistik kinerja akhir yang meliputi Final Equity, Mean Equity, dan Standard Deviation Equity. Analisis difokuskan pada perbedaan karakteristik pembelajaran antara pendekatan off-policy Q-Learning dan on-policy SARSA dalam merespons dinamika dan volatilitas harga saham IDX.

3.1 Hasil Saham BBCAJK (Bank Central Asia)

Saham BBCA.JK dipilih sebagai objek analisis karena merepresentasikan sektor keuangan dengan volatilitas menengah dan pergerakan harga yang relatif stabil dalam jangka panjang, sehingga sesuai sebagai benchmark untuk mengevaluasi kemampuan algoritma reinforcement learning pada kondisi pasar yang lebih terprediksi namun tetap dinamis. Proses pelatihan dilakukan melalui simulasi lingkungan trading yang memungkinkan agen Q-Learning dan SARSA mengambil aksi buy, hold, dan sell berdasarkan sinyal harga harian dari data historis BBCA. Dengan konfigurasi state berbasis window 10 hari dan 1.000 episode pelatihan, kedua algoritma dianalisis untuk menilai tingkat konvergensi, stabilitas reward, serta konsistensi strategi yang terbentuk, sehingga memberikan gambaran awal mengenai perbedaan karakteristik pembelajaran pada pasar saham yang relatif stabil.

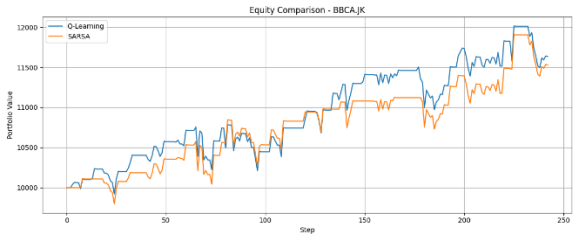
A. Analisis Reward Comparison



Gambar 2. Reward Comparison BBCAJK

Pola reward pada saham BBCA.JK menunjukkan perbedaan karakter pembelajaran antara Q-Learning dan SARSA. Pada fase awal, Q-Learning menghasilkan reward relatif moderat dengan nilai negatif sekitar −802 hingga −866 dan lonjakan positif hingga ±3.700, sedangkan SARSA mengalami fluktuasi lebih ekstrem dengan reward positif di atas 2.500 dan penurunan hingga di bawah −1.700. Pada fase pertengahan, Q-Learning semakin stabil dengan dominasi reward positif yang mencapai lebih dari 7.500, sementara SARSA mampu menghasilkan reward sangat tinggi hingga di atas 9.000 namun masih disertai penurunan tajam pada beberapa episode. Pada fase akhir, Q-Learning menunjukkan konsistensi reward positif pada kisaran 3.700 hingga di atas 6.300, menandakan konvergensi kebijakan yang stabil, sedangkan SARSA tetap memperlihatkan variasi besar dengan lonjakan di atas 8.400 dan reward rendah pada episode tertentu. Secara keseluruhan, Q-Learning lebih unggul dalam stabilitas reward jangka panjang, sementara SARSA cenderung lebih volatil akibat respons yang lebih reaktif terhadap dinamika pasar.

B. Analisis Equity Comparison



Gambar 3. Equity Comparison BBCAJK

Grafik equity saham BBCA.JK menunjukkan perbedaan gaya pembelajaran yang jelas antara Q-Learning dan SARSA. Pada fase awal, Q-Learning membangun kenaikan equity yang lebih stabil dan berada sedikit di atas SARSA, menandakan efektivitas awal dalam membentuk keputusan trading. Memasuki fase pertengahan, SARSA sempat menunjukkan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

momentum kenaikan yang lebih dinamis dan mendekati posisi Q-Learning, namun peningkatan tersebut tidak bertahan lama. Setelah itu, Q-Learning kembali menguat dengan tren kenaikan equity yang lebih konsisten hingga akhir simulasi. Pada fase akhir, kedua algoritma memperlihatkan pola pergerakan equity yang serupa, tetapi Q-Learning tetap mempertahankan posisi di atas SARSA. Kondisi ini tercermin pada nilai final equity, di mana Q-Learning mencapai 11.636 dan SARSA berada pada 11.530, dengan standar deviasi yang relatif rendah pada keduanya. Secara keseluruhan, hasil ini menunjukkan bahwa Q-Learning lebih efektif dalam mempertahankan akumulasi nilai portofolio jangka panjang pada saham BBCA, sementara SARSA menghasilkan kinerja yang stabil namun sedikit lebih rendah.

C. Analisis Statistik Kinerja

Tabel 1.Ringkasan Kinerja Model Pada Saham BBCAJK

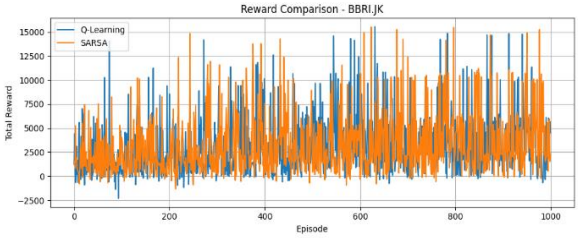
Metode	Final Equity	Mean Equity	Std Equity
Q-Learning	11636.326049	10932.427821	554.640456
SARSA	11530.248483	10780.471108	495.544853

Statistik kinerja saham BBCA.JK menunjukkan bahwa Q-Learning memiliki profitabilitas yang sedikit lebih unggul dibandingkan SARSA, dengan final equity masing-masing sebesar 11.636 dan 11.530. Keunggulan ini juga tercermin pada mean equity, di mana Q-Learning mencatat nilai rata-rata 10.932, lebih tinggi dibandingkan SARSA sebesar 10.780, menandakan bahwa sepanjang proses pelatihan Q-Learning lebih konsisten mempertahankan nilai portofolio yang lebih besar. Dari sisi stabilitas, SARSA menunjukkan performa yang lebih terkendali dengan standard deviation lebih rendah (495.54) dibandingkan Q-Learning (554.64), sesuai dengan karakteristik on-policy yang lebih konservatif. Namun demikian, tingkat stabilitas yang lebih tinggi tersebut tidak mampu mengimbangi keunggulan Q-Learning dalam akumulasi keuntungan akhir. Secara keseluruhan, hasil ini menegaskan bahwa pada saham BBCA, Q-Learning lebih efektif dalam menghasilkan profitabilitas jangka panjang, sementara SARSA unggul tipis dari sisi kestabilan performa.

3.2 Hasil Saham BBRIJK (Bank Rakyat Indonesia)

Saham BBRI.JK merepresentasikan sektor keuangan dengan volatilitas relatif tinggi dan sensitivitas kuat terhadap kondisi ekonomi makro, sehingga relevan untuk menguji kemampuan adaptasi algoritma reinforcement learning dalam menghadapi fluktuasi pasar ekstrem. Melalui simulasi perdagangan berbasis data historis, algoritma Q-Learning dan SARSA mempelajari pengambilan keputusan trading dengan mekanisme reward–penalty yang mencerminkan keuntungan dan kerugian, sehingga efektivitas strategi yang terbentuk dapat dievaluasi secara objektif dalam lingkungan pasar yang dinamis.

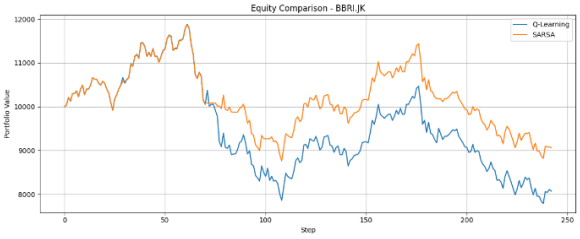
A. Analisis Reward Comparison



Gambar 3. Reward Comparison BBRIJK

Pola reward pada saham BBRI.JK menunjukkan perbedaan karakter pembelajaran yang jelas antara Q-Learning dan SARSA pada lingkungan pasar yang sangat volatil. Q-Learning menghasilkan reward positif tinggi seperti 7.019, 12.610, hingga 14.822 pada fase pembelajaran yang lebih matang, dengan penurunan negatif yang relatif terbatas, menunjukkan proses pembelajaran yang semakin stabil dan terkontrol. Sebaliknya, SARSA mencatat lonjakan reward yang lebih ekstrem, mencapai nilai sangat tinggi seperti 14.251, 14.903, hingga 15.236, namun disertai penurunan tajam hingga −801 atau −513 pada beberapa episode. Pola ini mengindikasikan bahwa SARSA sangat responsif terhadap perubahan harga jangka pendek, menghasilkan potensi reward besar tetapi dengan fluktuasi yang tinggi. Secara keseluruhan, Q-Learning menunjukkan kemampuan membangun reward yang lebih konsisten dan tahan terhadap noise pasar, sementara SARSA menawarkan reward yang lebih agresif namun berisiko pada saham BBRI yang memiliki dinamika harga cepat dan tidak stabil.

B. Analisis Equity Comparison



Gambar 4. Equity Comparison BBRIJK

Pergerakan equity pada saham BBRI.JK memperlihatkan dinamika yang sangat fluktuatif, mencerminkan karakter volatilitas tinggi sepanjang periode pelatihan. Pada kondisi ini, SARSA menunjukkan performa equity yang lebih unggul dibandingkan Q-Learning, dengan pola pertumbuhan yang lebih stabil dan konsisten setelah fase awal simulasi. Pendekatan on-policy memungkinkan SARSA menyesuaikan strategi secara bertahap

P-ISSN : 2089-3353
E-ISSN : 2808-9162

terhadap perubahan harga yang cepat, sehingga mampu menjaga kenaikan nilai portofolio tanpa mengalami penurunan ekstrem. Sebaliknya, Q-Learning menampilkan kurva equity yang lebih tidak stabil, dengan beberapa penurunan tajam akibat eksplorasi agresif yang kurang selaras dengan dinamika pasar BBRI. Meskipun Q-Learning memiliki potensi pertumbuhan yang tinggi, kesulitan dalam mempertahankan kestabilan equity membuat performa jangka panjangnya lebih rendah dibandingkan SARSA. Temuan ini menunjukkan bahwa pada saham dengan volatilitas tinggi seperti BBRI, SARSA lebih efektif dalam menjaga konsistensi dan stabilitas nilai portofolio.

C. Analisis Statistik Kinerja

Tabel 2.Ringkasan Kinerja Model Pada Saham BBRIJK

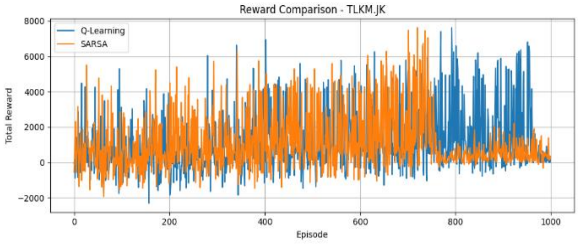
Metode	Final Equity	Mean Equity	Std Equity
Q-Learning	8069.993662	9540.316294	1017.468889
SARSA	9059.393081	10176.685307	700.478829

Statistik kinerja saham BBRI.JK menunjukkan bahwa SARSA unggul dibandingkan Q-Learning baik dari sisi profitabilitas maupun stabilitas. Final equity SARSA mencapai 9.059,39, lebih tinggi daripada Q-Learning sebesar 8.069,99, menandakan efektivitas pendekatan on-policy dalam mengikuti pergerakan harga BBRI yang volatil. Keunggulan ini diperkuat oleh mean equity SARSA yang lebih tinggi (10.176,69) dibandingkan Q-Learning (9.540,32), yang menunjukkan konsistensi performa portofolio sepanjang proses pembelajaran. Dari sisi risiko, SARSA juga memiliki standard deviation yang lebih rendah (700,48) dibandingkan Q-Learning (1.017,47), mencerminkan fluktuasi equity yang lebih terkendali. Secara keseluruhan, hasil statistik ini menegaskan bahwa SARSA lebih sesuai untuk karakteristik pasar BBRI yang dinamis karena mampu menjaga keseimbangan antara keuntungan dan stabilitas strategi.

3.3 Hasil Saham TLKMJK (Telkom Indonesia)

Saham TLKM.JK merepresentasikan sektor telekomunikasi yang bersifat defensif dengan volatilitas menengah dan pola pergerakan harga yang cenderung stabil, sehingga sesuai untuk menguji konsistensi dan adaptabilitas algoritma reinforcement learning pada lingkungan pasar yang tidak agresif. Melalui simulasi trading dengan aksi buy, hold, dan sell, algoritma Q-Learning dan SARSA dievaluasi kemampuannya dalam memanfaatkan peluang pergerakan harga yang terbatas sekaligus menjaga kestabilan strategi pada pasar yang terstruktur.

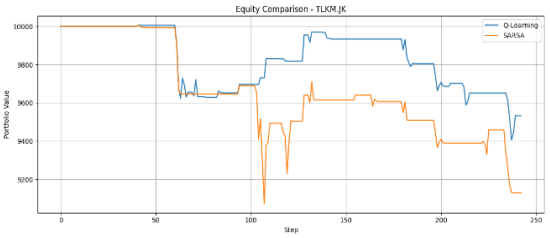
A. Analisis Reward Comparison



Gambar 5. Reward Comparison TLKMJK

Pola reward saham TLKM menunjukkan dinamika pembelajaran yang fluktuatif akibat pergerakan harga yang cenderung sideways. Pada fase awal, Q-Learning mencatat reward negatif hingga −341 namun juga lonjakan positif sekitar 2.860, sedangkan SARSA lebih ekstrem dengan reward di atas 3.100 dan penurunan hingga −1.366. Pada fase pertengahan, Q-Learning mulai stabil dengan dominasi reward positif yang mencapai lebih dari 4.900, sementara SARSA tetap menunjukkan variasi tinggi meskipun mampu menghasilkan reward di atas 4.000. Pada fase akhir, Q-Learning mempertahankan konsistensi reward positif hingga sekitar 6.805, sedangkan SARSA masih berfluktuasi. Secara keseluruhan, Q-Learning menunjukkan pembelajaran yang lebih stabil dan terkontrol pada TLKM, sementara SARSA lebih sensitif terhadap perubahan harga jangka pendek.

B. Analisis Equity Comparison



Gambar 6. Reward Comparison TLKMJK

Grafik equity comparison pada saham TLKM.JK menunjukkan bahwa kedua algoritma menghasilkan kinerja portofolio yang relatif stabil, sejalan dengan karakter pergerakan harga TLKM yang cenderung sideways. Sepanjang proses pelatihan, kurva Q-Learning umumnya berada sedikit di atas SARSA, menandakan akumulasi nilai portofolio yang lebih baik. Pada fase awal, kedua algoritma beradaptasi dengan pola pasar yang serupa sehingga garis equity bergerak berdekatan, namun memasuki fase pertengahan Q-Learning mulai menunjukkan kenaikan yang lebih konsisten, sementara SARSA sesekali mengalami koreksi kecil. Menjelang akhir pelatihan, perbedaan posisi kedua kurva semakin jelas tanpa adanya fluktuasi ekstrem, mengonfirmasi bahwa meskipun kedua metode stabil, Q-Learning lebih optimal dalam mempertahankan dan meningkatkan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

nilai portofolio pada saham TLKM yang berkarakter pergerakan harga moderat dan konsisten.

C. Analisis Statistik Kinerja

Tabel 3.Ringkasan Kinerja Model Pada Saham TLKMJK

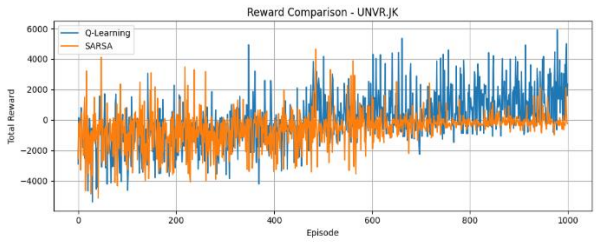
Metode	Final Equity	Mean Equity	Std Equity
Q-Learning	9532.635797	9829.181990	152.009869
SARSA	9129.183359	9648.125545	235.786146

Statistik kinerja saham TLKM.JK menunjukkan bahwa Q-Learning unggul pada seluruh indikator utama. Nilai final equity Q-Learning sebesar 9.532 lebih tinggi dibandingkan SARSA yang mencapai 9.129, diikuti oleh mean equity yang juga lebih besar (9.829 berbanding 9.648), menandakan kemampuan Q-Learning mempertahankan profitabilitas yang lebih konsisten sepanjang pelatihan. Dari sisi stabilitas, Q-Learning mencatat standard deviation yang jauh lebih rendah (152) dibandingkan SARSA (235), mengindikasikan fluktuasi portofolio yang lebih terkendali dan risiko yang lebih kecil. Temuan ini menegaskan bahwa untuk saham dengan karakter pergerakan harga stabil dan cenderung sideways seperti TLKM, Q-Learning lebih efektif dalam menghasilkan strategi trading yang optimal dan konsisten dibandingkan SARSA.

3.4 Hasil Saham UNVRJK (Unilever Indonesia)

Saham UNVR.JK merepresentasikan sektor consumer goods yang bersifat defensif dengan volatilitas rendah dan pergerakan harga relatif stabil, sehingga sesuai untuk menguji respons algoritma reinforcement learning pada kondisi pasar yang minim fluktuasi. Melalui lingkungan trading dengan aksi buy, hold, dan sell serta representasi state berbasis window, kinerja Q-Learning dan SARSA dianalisis menggunakan reward, perkembangan equity, dan statistik kinerja untuk menilai kemampuan masing-masing metode dalam membentuk strategi yang stabil pada karakteristik pasar defensif.

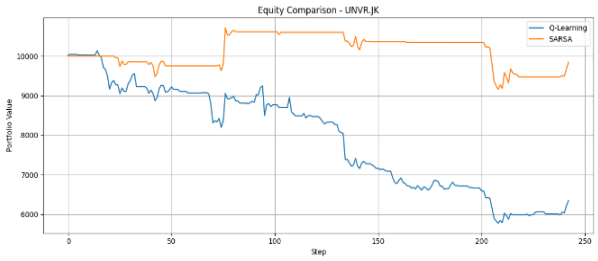
A. Analisis Reward Comparison



Gambar 7.Reward Comparison UNVRJK

Pola reward saham UNVR.JK menunjukkan fluktuasi tinggi pada fase awal, di mana Q-Learning dan SARSA sama-sama menghasilkan reward negatif besar akibat eksplorasi awal yang belum efisien, masing-masing mencapai −2888 dan −2533, meskipun mulai muncul reward positif seperti 801 dan 1497. Pada fase pertengahan, Q-Learning lebih sering mencatat reward positif bernilai tinggi (di atas 2000), sementara SARSA menunjukkan peningkatan yang lebih terbatas. Memasuki fase akhir, Q-Learning semakin dominan dengan lonjakan reward hingga mendekati 6000, sedangkan SARSA menghasilkan reward yang lebih moderat dan sesekali masih negatif. Secara keseluruhan, Q-Learning lebih efektif dalam mengoptimalkan reward jangka panjang pada UNVR yang bersifat defensif, sementara SARSA cenderung lebih konservatif dan stabil.

B. Analisis Equity Comparison



Gambar 8. Equity Comparison UNVRJK

Perbandingan equity pada saham UNVR.JK memperlihatkan perbedaan karakter pembelajaran yang jelas antara Q-Learning dan SARSA. Kurva equity Q-Learning cenderung lebih fluktuatif dengan beberapa kenaikan dan penurunan tajam, mencerminkan sifat eksploratif dan agresif yang berpotensi menghasilkan pertumbuhan besar namun disertai risiko penurunan signifikan. Sebaliknya, SARSA menunjukkan jalur equity yang lebih stabil dan konsisten, dengan kenaikan bertahap tanpa fluktuasi ekstrem akibat pendekatan on-policy yang lebih berhati-hati. Pada fase awal, kedua algoritma masih bergerak relatif berdekatan, namun pada fase pertengahan hingga akhir, SARSA mulai mempertahankan posisi equity yang lebih unggul dan stabil, sementara Q-Learning kerap mengalami koreksi akibat volatilitas internal yang tinggi. Secara keseluruhan, hasil ini menegaskan bahwa SARSA lebih sesuai untuk karakter saham UNVR yang defensif karena mampu menjaga stabilitas dan kontrol risiko yang lebih baik dibandingkan Q-Learning.

C. Analisis Statistik Kinerja

Tabel 4.Ringkasan Kinerja Model Pada Saham UNVRJK

Metode	Final Equity	Mean Equity	Std Equity
Q-Learning	6344.103352	7919.052305	1315.954871
SARSA	9844.293901	10114.715094	407.334764

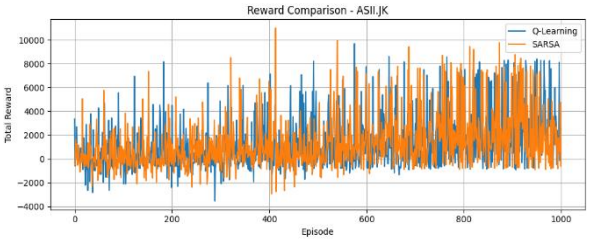
P-ISSN : 2089-3353
E-ISSN : 2808-9162

Statistik kinerja pada saham UNVR.JK menunjukkan perbedaan performa yang sangat jelas antara kedua algoritma. SARSA mencatat final equity sebesar 9844.29, jauh lebih tinggi dibandingkan Q-Learning yang hanya mencapai 6344.10. Keunggulan ini juga terlihat pada mean equity, di mana SARSA berada di angka 10114.72, secara konsisten lebih tinggi dibandingkan Q-Learning yang mencatat 7919.05. Perbedaan ini mengindikasikan bahwa sepanjang proses pembelajaran, SARSA membangun strategi yang lebih stabil dan efektif, sementara Q-Learning lebih sering mengalami penurunan nilai portofolio akibat keputusan agresif yang kurang sesuai dengan karakteristik harga UNVR yang defensif dan minim volatilitas.

3.5 Hasil Saham ASIIJK (Astra International)

Saham ASII.JK dipilih karena merepresentasikan sektor industri dan otomotif yang bersifat siklikal dengan pergerakan harga dinamis dan sensitif terhadap kondisi ekonomi makro. Dibandingkan saham defensif atau relatif stabil, ASII menunjukkan fluktuasi lebih tajam pada periode perubahan ekonomi dan aktivitas manufaktur, sehingga relevan untuk menguji kemampuan adaptasi algoritma reinforcement learning pada pola harga yang agresif dan tidak teratur. Karakter siklikal ini menimbulkan ketidakpastian reward dan dinamika jangka menengah yang menuntut fleksibilitas model. Oleh karena itu, Q-Learning dan SARSA diuji dalam struktur aksi buy–hold–sell yang seragam guna memastikan perbandingan kinerja yang konsisten dengan saham lain.

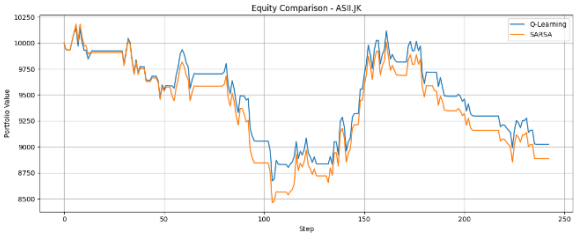
A. Analisis Reward Comparison



Gambar 9.Reward Comparison ASII.JK

Pola reward saham ASII.JK menunjukkan perbedaan karakter pembelajaran yang jelas antara Q-Learning dan SARSA pada saham bersifat siklikal. Pada fase awal, Q-Learning bersifat agresif dengan lonjakan reward seperti +3343 dan +2686, namun juga mengalami penurunan tajam hingga −2845, sedangkan SARSA lebih moderat meski sempat mencatat reward tinggi seperti +5051. Pada fase pertengahan, Q-Learning menghasilkan lonjakan besar hingga +7126 dan +6142 namun masih disertai risiko, sementara SARSA menunjukkan tren yang lebih stabil dengan reward tinggi seperti +10994 dan +3648. Memasuki fase akhir, Q-Learning tetap agresif dengan reward hingga +8390, sedangkan SARSA mempertahankan pola yang lebih konsisten seperti +8748 dan +7139. Secara keseluruhan, Q-Learning unggul dalam menangkap peluang profit besar dengan volatilitas tinggi, sedangkan SARSA lebih stabil dan terkendali pada ASII yang bersifat siklikal.

B. Analisis Equity Comparison



Gambar 10. Equity Comparison ASII.JK

Pergerakan equity saham ASII.JK menunjukkan bahwa Q-Learning dan SARSA memiliki pola yang relatif serupa pada fase awal hingga pertengahan simulasi, dengan fluktuasi naik–turun yang aktif sesuai karakter ASII sebagai saham siklikal. Pada tahap ini, kedua algoritma merespons dinamika harga dengan pola yang hampir sama dan belum menunjukkan perbedaan yang signifikan. Memasuki pertengahan episode, kedua kurva mengalami tren penurunan, namun SARSA cenderung mencatat koreksi yang lebih sering dan lebih dalam, sementara Q-Learning mengalami penurunan dengan amplitudo yang lebih terbatas sehingga pemulihan nilai portofolio berlangsung lebih cepat. Pada fase akhir simulasi, perbedaan stabilitas semakin terlihat, di mana SARSA masih mengalami beberapa penurunan tambahan, sedangkan Q-Learning mampu mempertahankan tren equity yang lebih konsisten hingga akhir episode. Meskipun bentuk kurva keduanya tetap mirip, posisi akhir equity Q-Learning berada sedikit lebih tinggi, menunjukkan bahwa pendekatan off-policy lebih efektif menjaga nilai portofolio pada fase akhir perdagangan saham ASII

C. Analisis Statistik Kinerja

Tabel 5.Ringkasan Kinerja Model Pada Saham ASIIJK

Metode	Final Equity	Mean Equity	Std Equity
Q-Learning	9025.629222	9487.715466	381.549383
SARSA	8887.409907	9377.617994	427.404194

Statistik kinerja saham ASII.JK menunjukkan bahwa Q-Learning memiliki keunggulan tipis dibandingkan SARSA baik dari sisi profitabilitas maupun stabilitas. Final equity Q-Learning tercatat sebesar 9.025,63, sedikit lebih tinggi dibandingkan SARSA yang mencapai 8.887,41, sejalan dengan kemampuannya menjaga nilai portofolio lebih stabil menjelang akhir simulasi. Keunggulan ini juga tercermin pada mean equity, di mana Q-Learning mencatat nilai rata-rata

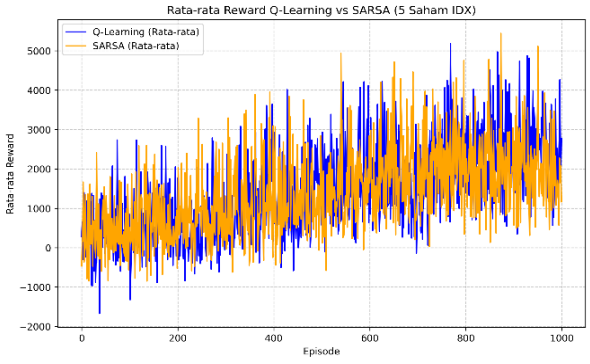
P-ISSN : 2089-3353
E-ISSN : 2808-9162

9.487,72, melampaui SARSA yang berada pada 9.377,62, menandakan konsistensi kinerja yang lebih baik sepanjang proses pelatihan pada saham ASII yang bersifat siklikal. Dari sisi risiko, Q-Learning menunjukkan fluktuasi yang lebih terkendali dengan standar deviasi equity sebesar 381,55, lebih rendah dibandingkan SARSA yang mencapai 427,40. Secara keseluruhan, ketiga indikator tersebut menegaskan bahwa Q-Learning memberikan kombinasi stabilitas dan profitabilitas yang sedikit lebih unggul dibandingkan SARSA, meskipun pola kinerja kedua algoritma relatif serupa pada saham ASII.JK.

3.6 Analisis Perbandingan Umum dan Tren Reward Keseluruhan

Bagian ini menyajikan analisis menyeluruh terhadap kinerja Q-Learning dan SARSA berdasarkan lima saham yang telah diuji, yaitu BBCA, BBRI, TLKM, UNVR, dan ASII. Setelah seluruh proses pelatihan dan evaluasi dilakukan secara terpisah pada tiap saham, langkah selanjutnya adalah mengamati tren final reward, final equity, serta stabilitas performa masing-masing algoritma dalam skala agregat. Pendekatan ini memungkinkan pengamatan pola umum yang mungkin tidak terlihat ketika setiap saham dianalisis secara individual, terutama karena karakteristik volatilitas dan dinamika harga dari tiap emiten berbeda-beda. Melalui perbandingan lintas-saham ini, dapat diidentifikasi kecenderungan algoritma yang lebih konsisten, strategi yang lebih adaptif terhadap berbagai kondisi pasar, serta keunggulan relatif masing-masing metode dalam menghasilkan keuntungan dan menjaga stabilitas portofolio selama proses pembelajaran berlangsung.

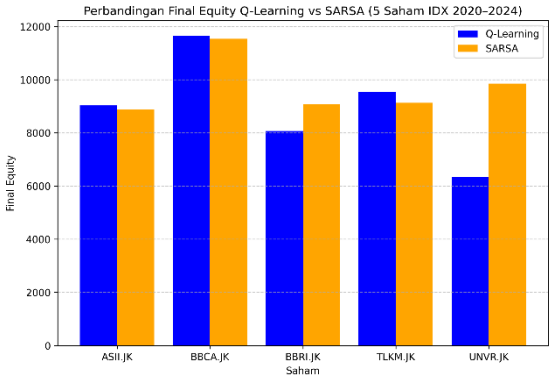
A. Analisis Final Reward Comparison



Gambar 11.Final Reward Comparison

Analisis final reward menunjukkan bahwa kinerja Q-Learning dan SARSA sangat bergantung pada karakteristik masing-masing saham, tanpa pola keunggulan yang bersifat universal. Pada ASII.JK, SARSA sedikit unggul dengan reward 1350,41 dibandingkan Q-Learning sebesar 1276,49, sementara pada BBCA.JK selisihnya lebih jelas dengan SARSA mencapai 1753,00 dan Q-Learning 1488,13, menandakan efektivitas pendekatan on-policy pada pasar yang relatif stabil. Pola serupa juga terlihat pada BBRI.JK, di mana SARSA kembali mencatat reward lebih tinggi (3351,68) dibandingkan Q-Learning (3187,20) pada saham yang sangat volatil. Sebaliknya, pada TLKM.JK, Q-Learning lebih unggul dengan reward 1204,56 dibandingkan SARSA sebesar 1100,57, menunjukkan kelebihannya pada pasar sideways. UNVR.JK menjadi pengecualian ekstrem karena kedua algoritma menghasilkan reward negatif, meskipun Q-Learning (–13,92) jauh lebih baik dibandingkan SARSA (–565,14). Secara agregat, Q-Learning memiliki rata-rata reward keseluruhan yang sedikit lebih tinggi (1428,49) dibandingkan SARSA (1398,10), namun SARSA unggul pada tiga dari lima saham, menegaskan bahwa efektivitas algoritma lebih ditentukan oleh karakter volatilitas aset daripada dominasi satu metode secara umum.

B. Analisis Final Equity Comparison



Gambar 12.Final Equity Comparison

Analisis final equity memberikan gambaran kumulatif mengenai efektivitas strategi Q-Learning dan SARSA pada masing-masing saham. Pada ASII.JK dan BBCA.JK, performa kedua algoritma relatif berdekatan, dengan Q-Learning sedikit unggul pada ASII (9025,63 vs 8887,41) dan BBCA (11636,33 vs 11530,25), menunjukkan kemampuannya memanfaatkan momentum tertentu pada saham yang stabil namun tetap dinamis. Perbedaan yang lebih tegas terlihat pada BBRI.JK dan TLKM.JK, di mana SARSA mendominasi BBRI dengan final equity 9059,39 dibandingkan Q-Learning 8069,99 akibat keunggulan stabilitasnya pada volatilitas tinggi, sementara pada TLKM Q-Learning unggul dengan 9532,64 dibandingkan SARSA 9129,18 karena efektivitas strategi eksploratif pada pasar sideways. Perbedaan paling signifikan terjadi pada UNVR.JK, di mana SARSA mencatat final equity 9844,29, jauh melampaui Q-Learning yang hanya mencapai 6344,10, menegaskan bahwa pendekatan on-policy lebih sesuai untuk saham defensif dengan volatilitas rendah. Secara rata-rata, SARSA menghasilkan final equity tertinggi sebesar 9690,11, mengungguli Q-Learning yang berada pada 8921,74, sehingga menegaskan keunggulan SARSA dalam stabilitas jangka panjang meskipun Q-Learning tetap kompetitif pada saham tertentu.

P-ISSN : 2089-3353
E-ISSN : 2808-9162

C. Analisis Final Statistik Kinerja

Tabel 6.Ringkasan Kinerja Model Pada 5 Saham IDX

No	Saham	Final Equity (Q)	Final Equity (S)	Rata-rata Reward (Q)	Rata-rata Reward (S)
1	ASII.JK	9025.63	8887.41	1276.49	1350.41
2	BBCA.JK	11636.33	11530.25	1488.13	1752.99
3	BBRI.JK	8069.99	9059.39	3187.20	3351.68
4	TLKM.JK	9532.64	9129.18	1204.56	1100.57
5	UNVR.JK	6344.10	9844.29	-13.92	-565.14
Rata-rata		8921.74	9690.11	1428.49	1398.10

Analisis statistik kinerja menunjukkan bahwa efektivitas Q-Learning dan SARSA sangat bergantung pada karakteristik saham. Pada ASII.JK dan BBCA.JK, kedua algoritma bersaing ketat, dengan Q-Learning sedikit unggul pada final equity, sementara SARSA lebih stabil dengan risiko lebih rendah. Pada BBRI.JK, SARSA mendominasi seluruh metrik, menegaskan keunggulannya dalam menghadapi volatilitas tinggi melalui kontrol risiko yang lebih baik. Sebaliknya, TLKM.JK menunjukkan keunggulan Q-Learning pada final dan mean equity meskipun dengan deviasi standar lebih besar, mencerminkan efektivitas strategi eksploratif pada pasar sideways. Perbedaan paling signifikan terjadi pada UNVR.JK, di mana SARSA menghasilkan final dan mean equity yang jauh lebih tinggi dengan volatilitas rendah, sementara Q-Learning mengalami fluktuasi besar. Secara keseluruhan, SARSA mencatat rata-rata final equity tertinggi sebesar 9690,11 dengan stabilitas risiko terbaik, sedangkan Q-Learning lebih sesuai untuk saham dengan peluang momentum yang mendukung strategi agresif

4. Kesimpulan

Penelitian ini menunjukkan bahwa Q-Learning dan SARSA memiliki karakteristik kinerja yang berbeda dalam penerapan strategi trading otomatis pada saham LQ45. Q-Learning cenderung unggul pada saham dengan momentum kuat karena sifat off-policy yang lebih agresif dan mampu memaksimalkan peluang profit, sedangkan SARSA lebih stabil dan konsisten pada saham dengan volatilitas tinggi atau pola defensif melalui kontrol risiko yang lebih baik. Secara keseluruhan, SARSA menunjukkan stabilitas performa yang sedikit lebih unggul, sementara Q-Learning menawarkan potensi keuntungan lebih tinggi pada kondisi pasar tertentu, sehingga pemilihan algoritma optimal sangat bergantung pada karakteristik saham dan toleransi risiko dalam strategi investasi.

Daftar Rujukan

[1] O. V. Zaporozhets, O. V. Okhrimenko, and O. V. Levchenko, "Digital methods of technical analysis for diagnosis of crisis phenomena in the financial market," International Journal of Technology, vol. 13, no. 7, pp. 1527–1536, 2022, doi: 10.14716/ijtech.v13i7.6187.

[2] X. Chen, B. Xu, Y. Li, and Y. Gao, "A stock prediction method based on deep reinforcement learning and sentiment analysis," Applied Sciences, vol. 14, no. 19, p. 8747, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/19/8747

[3] R. J. Elliott and R. Mamon, "Portfolio management system in equity market neutral using reinforcement learning," Applied Intelligence, vol. 52, pp. 1–17, 2021, doi: 10.1007/s10489-021-02262-0.

[4] D. Antic, "Application of Q-learning in financial markets: Modelling and experimental results," International Education and Research Journal, vol. 7, no. 9, pp. 10–14, 2021. [Online]. Available: https://ierj.in/journal/index.php/ierj/article/view/4529/5337

[5] A. R. T. C. Mandala, I. M. A. Pura, and G. Mahardika, "Application of deep reinforcement learning for stock trading on the Indonesia Stock Exchange," JANAPATI, vol. 12, no. 1, pp. 55–67, 2023. [Online]. Available: https://ejournal.undiksha.ac.id/index.php/janapati/article/view/83775

[6] A. Brim and N. S. Flann, "Deep reinforcement learning stock market trading utilizing a CNN with candlestick images," PLOS ONE, vol. 17, no. 2, 2022. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/pone.0263181

[7] Y. Jiang, C. Xu, X. Ji, and Y. Li, "A multi-scaling reinforcement learning trading system based on multi-scaling convolutional neural networks," Mathematics, vol. 11, no. 11, p. 2467, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/11/2467

[8] M. Ghoreishi and S. Jafari, "Deep reinforcement learning for Tehran stock trading," Journal of New Engineering Science and Technology, vol. 2, no. 3, pp. 55–66, 2022. [Online]. Available: https://journal.iistr.org/index.php/JNEST/article/view/171/125

[9] T. Kabbani and E. Duman, "Deep reinforcement learning approach for trading automation in the stock market," IEEE Access, vol. 10, pp. 96841–96854, 2022, doi: 10.1109/ACCESS.2022.3203697.

[10] A. Rahman and B. Sitohang, "Reinforcement learning for bitcoin trading: A comparative study of PPO and DQN," Jurnal Mandiri, vol. 6, no. 2, pp. 501–512, 2023. [Online]. Available: https://ejournal.isha.or.id/index.php/Mandiri/article/view/455/457

[11] A. S. Perera, A. A. Perera, and D. Dias, "Deep reinforcement learning for trading—A critical survey," Data, vol. 6, no. 11, p. 119, 2021. [Online]. Available: https://www.mdpi.com/2306-5729/6/11/119

[12] M. López de Prado, Financial Machine Learning. SSRN, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501707

[13] A. Kolm and G. Ritter, Foundations of Reinforcement Learning with Applications in Finance. Stanford University, 2021. [Online]. Available: https://stanford.edu/~ashlearn/RLForFinanceBook/book.pdf

[14] Y. Wang, Y. Wu, and Z. Zhang, "Stock trading strategies based on deep reinforcement learning," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/4698656.

[15] M. Liu, Z. Wang, and J. Zhang, "A deep Q-learning portfolio management framework for the cryptocurrency market," Neural Computing and Applications, vol. 34, pp. 1–15, 2022, doi: 10.1007/s00521-020-05359-8.