

BAB 2

LANDASAN TEORI

Bab ini menjelaskan terkait teori-teori yang menjadi dasar dalam penelitian prediksi harga VanEck *Semiconductor* ETF (SMH) menggunakan model *Long Short-Term Memory* (LSTM). Pembahasan difokuskan pada konsep prediksi harga aset finansial berbasis data *time series*, *deep learning*, arsitektur LSTM, teknik *preprocessing* data, serta metrik evaluasi model.

2.1 Konteks Investasi *Exchange Traded Fund* (ETF) Sektor *Semiconductor*

Exchange Traded Fund (ETF) merupakan sebuah instrumen investasi yang diperdagangkan di bursa saham dan dirancang untuk merepresentasikan kinerja suatu indeks, sektor industri, atau kumpulan aset tertentu. ETF memungkinkan investor memperoleh diversifikasi portofolio secara efisien karena satu unit ETF mencerminkan kepemilikan pada berbagai saham penyusun indeks tersebut. Selain itu, ETF memiliki karakteristik likuiditas tinggi, transparansi harga, serta biaya transaksi yang relatif rendah dibandingkan reksa dana tradisional [8].

Dalam beberapa tahun terakhir, ETF berbasis sektor teknologi, khususnya sektor semikonduktor, mengalami pertumbuhan signifikan seiring meningkatnya permintaan global terhadap chip dan komponen elektronik. Industri semikonduktor berperan penting dalam pengembangan teknologi modern seperti kecerdasan buatan, kendaraan listrik, komputasi awan, serta perangkat *Internet of Things* (IoT). Peningkatan kebutuhan terhadap infrastruktur digital menyebabkan saham perusahaan semikonduktor menjadi salah satu sektor dengan volatilitas dan potensi pertumbuhan yang tinggi di pasar keuangan global [9].

Meskipun memiliki potensi *return* yang menarik, harga ETF sektor semikonduktor cenderung mengalami fluktuasi yang signifikan akibat dipengaruhi oleh siklus bisnis teknologi, kondisi ekonomi makro, serta sentimen pasar. Tingginya volatilitas ini menyebabkan ketidakpastian dalam pengambilan keputusan investasi sehingga diperlukan metode analisis prediktif yang mampu memodelkan pola historis harga secara efektif [10].

Dalam konteks penelitian modern, pendekatan berbasis *machine learning* dan *deep learning*, khususnya model *Long Short-Term Memory* (LSTM), semakin banyak digunakan dalam prediksi harga aset finansial. Model ini memiliki

kemampuan untuk mempelajari pola temporal jangka panjang pada data *time series* sehingga dapat meningkatkan akurasi prediksi dibandingkan metode statistik tradisional. Oleh karena itu, penerapan model LSTM dalam prediksi harga ETF sektor semikonduktor menjadi relevan untuk mendukung pengambilan keputusan investasi yang lebih optimal [7].

2.2 *Financial Time Series*

Financial time series merupakan data berurutan yang dicatat berdasarkan urutan waktu dan menggambarkan perubahan nilai suatu aset keuangan, seperti harga saham, indeks pasar, maupun *Exchange Traded Fund* (ETF). Data ini biasanya disajikan dalam bentuk harga penutupan harian, harga pembukaan, harga tertinggi, harga terendah, serta volume transaksi. Analisis *financial time series* bertujuan untuk mengidentifikasi pola historis guna membantu proses prediksi pergerakan harga di masa mendatang [11].

Salah satu karakteristik utama data finansial adalah sifat *non-stationary*, yaitu kondisi ketika nilai statistik seperti rata-rata (mean) dan varians (tingkat penyebaran data) berubah seiring waktu. Perubahan ini dapat disebabkan oleh faktor ekonomi makro, kebijakan pemerintah, kondisi industri, maupun sentimen pasar. Akibatnya, pola historis tidak selalu konsisten sehingga model prediksi harus mampu beradaptasi terhadap dinamika perubahan tersebut [12].

Selain itu, *financial time series* juga mengandung *noise*, yaitu fluktuasi acak yang tidak mencerminkan pola fundamental pergerakan harga. *Noise* dapat muncul akibat aktivitas perdagangan jangka pendek, spekulasi investor, maupun ketidaksempurnaan informasi di pasar. Kehadiran *noise* menyebabkan sinyal pola yang sebenarnya menjadi sulit dikenali oleh model prediktif, sehingga meningkatkan kompleksitas proses analisis [13].

Karakteristik penting lainnya adalah *volatility*, yaitu tingkat perubahan harga yang cepat dan signifikan dalam periode waktu tertentu. Volatilitas tinggi sering terjadi pada sektor teknologi dan aset dengan pertumbuhan cepat, termasuk ETF sektor semikonduktor. Kondisi ini menciptakan peluang keuntungan yang besar, namun juga meningkatkan risiko kerugian sehingga memerlukan pendekatan analisis yang lebih canggih untuk memahami perilaku pasar [14].

Berdasarkan karakteristik tersebut, prediksi harga pada *financial time series* menjadi tantangan yang kompleks karena melibatkan pola *non-linear*, ketergantungan jangka panjang, serta dinamika perubahan pasar yang cepat. Oleh

karena itu, metode berbasis deep learning seperti *Long Short-Term Memory* (LSTM) banyak digunakan karena mampu mempelajari hubungan temporal dalam data sekuensial serta menangani permasalahan *non-stationary* dan *volatility* secara lebih efektif dibandingkan metode statistik tradisional [15].

2.3 Machine Learning untuk Prediksi Harga

Machine Learning merupakan metode komputasi yang memungkinkan model mempelajari pola dari data historis untuk menghasilkan prediksi pada data baru. Dalam pasar keuangan, *machine learning* digunakan untuk menganalisis hubungan kompleks yang memengaruhi pergerakan harga saham maupun *Exchange Traded Fund* (ETF), terutama karena data finansial cenderung bersifat *non-linear* dan dinamis [16].

Pendekatan yang umum digunakan adalah *supervised learning*, yaitu model dilatih menggunakan data *input* dan *target output*. Pada prediksi harga, data historis seperti harga penutupan atau volume transaksi digunakan sebagai fitur, sedangkan harga pada periode berikutnya menjadi *target* prediksi. Proses ini termasuk dalam *forecasting* numerik, yaitu prediksi nilai kontinu berdasarkan urutan waktu [17].

Dibandingkan metode statistik klasik seperti *Autoregressive Integrated Moving Average* (ARIMA), *machine learning* memiliki keunggulan dalam menangkap pola *non-linear* serta beradaptasi terhadap perubahan kondisi pasar. Selain itu, metode ini mampu memanfaatkan data dalam jumlah besar dan berbagai fitur tambahan sehingga berpotensi meningkatkan akurasi prediksi harga aset finansial [18].

Perkembangan lebih lanjut dari *machine learning* adalah penggunaan model *deep learning* seperti *Long Short-Term Memory* (LSTM), yang dirancang untuk mempelajari dependensi jangka panjang pada data *time series* dan banyak digunakan dalam penelitian prediksi harga saham dan ETF [19].

2.4 Deep Learning

Deep Learning merupakan pengembangan dari *machine learning* yang menggunakan arsitektur jaringan saraf tiruan (*neural network*) dengan banyak lapisan untuk mempelajari representasi data secara otomatis. Pendekatan ini memungkinkan model mengekstraksi fitur kompleks dari data dalam jumlah besar tanpa memerlukan perancangan fitur manual, sehingga efektif digunakan dalam

berbagai permasalahan prediksi berbasis data [20].

Struktur *multilayer* pada *neural network* membuat *deep learning* mampu menangkap pola *non-linear*, yaitu hubungan yang tidak sederhana antara variabel *input* dan *output*. Dalam konteks pasar keuangan, kemampuan ini menjadi penting karena pergerakan harga aset dipengaruhi oleh interaksi berbagai faktor ekonomi dan sentimen pasar yang kompleks [20].

Pada *financial time series*, *deep learning* banyak digunakan untuk memodelkan pola temporal dan dependensi jangka panjang pada data historis harga. Model *deep learning* seperti *Recurrent Neural Network* (RNN) dan *Long Short-Term Memory* (LSTM) terbukti efektif dalam meningkatkan performa prediksi dibandingkan metode statistik tradisional karena mampu mempelajari dinamika perubahan harga secara lebih adaptif [15].

Dengan demikian, penggunaan *deep learning* menjadi pendekatan penting dalam pengembangan model prediksi harga saham maupun *Exchange Traded Fund* (ETF), terutama pada data dengan karakteristik *non-stationary* dan volatilitas tinggi.

2.5 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) merupakan salah satu pengembangan dari *Recurrent Neural Network* (RNN) yang dirancang agar dapat memproses data berurutan atau sekuensial, seperti data *time series*. Berbeda dengan RNN konvensional yang mengalami kesulitan dalam mengingat informasi dari periode yang jauh sebelumnya, LSTM memiliki mekanisme memori internal yang memungkinkan model mempertahankan informasi penting dalam jangka waktu yang lebih panjang.

Pada prediksi harga aset keuangan, informasi dari beberapa periode sebelumnya sering kali masih memengaruhi pergerakan harga pada periode berikutnya. Oleh karena itu, kemampuan LSTM dalam menyimpan hubungan jangka panjang menjadikan model ini sesuai digunakan untuk memprediksi harga VanEck Semiconductor ETF (SMH) berdasarkan data historis.

LSTM memiliki komponen utama yang disebut *cell state*. *Cell state* merupakan jalur memori yang membawa informasi sepanjang urutan waktu. Informasi yang disimpan dalam *cell state* dikendalikan oleh tiga komponen utama yang disebut sebagai *gate*, yaitu *forget gate*, *input gate*, dan *output gate*. Ketiga *gate* tersebut bekerja untuk menentukan informasi mana yang perlu dipertahankan, diperbarui, atau dikeluarkan dari memori model [20].

Forget gate memiliki fungsi untuk menentukan apakah informasi dari periode sebelumnya masih relevan atau perlu dihapus. Jika suatu informasi dianggap tidak lagi berpengaruh terhadap prediksi berikutnya, maka informasi tersebut tidak akan diteruskan ke proses selanjutnya.

Input gate memiliki fungsi untuk menentukan informasi baru yang akan dimasukkan ke dalam memori model. Pada tahap ini, LSTM memilih informasi dari data saat ini yang dianggap penting untuk membantu proses prediksi.

Setelah itu, *cell state* diperbarui dengan menggabungkan informasi lama yang dipertahankan dan informasi baru yang telah dipilih. Proses ini memungkinkan LSTM menyimpan pola historis yang relevan, seperti tren kenaikan, penurunan, maupun fluktuasi harga yang terjadi dalam beberapa periode sebelumnya.

Terakhir, *output gate* menentukan informasi yang akan digunakan sebagai keluaran model pada periode tersebut. Keluaran ini kemudian diteruskan ke langkah berikutnya dan digunakan untuk menghasilkan prediksi harga.

Melalui mekanisme tersebut, LSTM mampu mempelajari pola *non-linear* dan hubungan temporal jangka panjang pada data *financial time series*. Kemampuan ini membuat LSTM lebih efektif dibandingkan metode statistik tradisional dalam menangani data dengan karakteristik *non-stationary*, volatilitas tinggi, dan perubahan pola yang dinamis [21].

Dengan kemampuan tersebut, LSTM dipilih dalam penelitian ini karena dinilai mampu menangkap pola historis harga VanEck Semiconductor ETF (SMH) secara lebih baik untuk menghasilkan prediksi harga pada periode berikutnya.

2.6 Penelitian Terdahulu

Penelitian-penelitian terdahulu menunjukkan bahwa pendekatan *deep learning*, khususnya *Long Short-Term Memory* (LSTM), telah banyak digunakan dalam prediksi pada sektor keuangan dan industri semikonduktor. Chen (2025) menggunakan pendekatan *deep learning* untuk memprediksi ETF semikonduktor dengan fokus pada klasifikasi arah pergerakan harga. Sementara itu, Yen dan Chen (2025) memanfaatkan LSTM untuk melakukan *forecasting* tren pasar semikonduktor dengan mempertimbangkan variabel tambahan. Meskipun demikian, kedua penelitian tersebut belum secara spesifik memprediksi *return H+1* pada VanEck Semiconductor ETF (SMH) sebagai nilai kontinu yang kemudian dapat direkonstruksi menjadi harga prediksi.

Selain itu, penelitian terdahulu umumnya masih berfokus pada prediksi arah pergerakan atau tren pasar secara umum, sehingga belum memberikan analisis yang lebih spesifik terhadap prediksi nilai *return* harian pada SMH. Oleh karena itu, penelitian ini berupaya mengisi kekosongan tersebut dengan membangun model LSTM untuk memprediksi *return* H+1 SMH menggunakan enam fitur historis, yaitu *Open*, *High*, *Low*, *Close*, *Volume*, dan *Return*.



Tabel 2.1. Ringkasan Penelitian Terdahulu, Metode Evaluasi, dan Kesenjangan Penelitian

No	Peneliti (Tahun)	Fokus Penelitian	Metode, Objek, dan Data	Metode Evaluasi dan Hasil	Kesenjangan Penelitian
1	Chen (2025)	Prediksi arah perubahan harga VanEck <i>Semiconductor</i> ETF (SMH) pada horizon lima hari perdagangan (H+5).	Klasifikasi menggunakan Logistic Regression, SVM, Random Forest, LightGBM, XGBoost, CatBoost, BPN, dan <i>soft voting</i> . Data berupa harga dan volume harian SMH periode 2 Januari 2014–31 Desember 2024, serta indikator VIX, VXN, NDXT, dan indikator teknikal.	Evaluasi menggunakan <i>test accuracy</i> , F1- <i>score</i> , AUC, dan <i>cross-validation accuracy</i> . <i>Random Forest</i> memperoleh <i>test accuracy</i> terbaik sebesar 56,64%, sedangkan model <i>voting</i> memperoleh AUC terbaik sebesar 58,76%.	Berfokus pada klasifikasi arah pergerakan H+5, sehingga belum menghasilkan prediksi <i>return</i> kontinu H+1 maupun harga prediksi yang dapat dievaluasi dengan metrik galat regresi.
2	Yen dan Chen (2025)	Prediksi tren industri semikonduktor Taiwan melalui indikator operasional dan keuangan TSMC.	LSTM dengan data kuartalan TSMC periode 1998Q1–2023Q4 (104 observasi): <i>Net Sales</i> , <i>Cost of Sales</i> , <i>Gross Profit</i> , <i>Net Income</i> , EPS, <i>Wafer Shipment</i> , <i>Income from Operations</i> , dan <i>Operating Expenses</i> . Analisis sentimen laporan perusahaan diperkuat dengan intervensi peristiwa internal dan eksternal.	Evaluasi terutama bersifat kualitatif, yaitu kesesuaian pola prediksi dengan berita produk dan peristiwa industri. Studi tidak melaporkan MAE, RMSE, MAPE, atau metrik kuantitatif lain pada data uji.	Objek penelitian berupa tren fundamental TSMC dengan frekuensi kuartalan, bukan harga ETF harian. Selain itu, belum ada evaluasi galat regresi yang terukur pada data pengujian.
3	Fei dan Li (2026)	Prediksi harga ETF menggunakan model hibrida Bi-LSTM–Bi-GRU.	Regresi harga tiga ETF: China Asset Nasdaq 100 ETF, China Asset S&P 500 ETF, dan China Asset Hang Seng Biotechnology ETF. Data periode 19 Oktober 2021–24 Oktober 2025; fitur harga dan volume. Data dibagi secara kronologis menjadi 60% pelatihan, 20% validasi, dan 20% pengujian.	Evaluasi menggunakan MAE, RMSE, dan R^2 , serta dibandingkan dengan ARIMA, XGBoost, SVR, LSTM, GRU, Transformer, dan TCN. Model Bi-LSTM–Bi-GRU memberikan galat terendah pada ketiga ETF.	Menggunakan arsitektur hibrida dan ETF non-SMH; belum menguji LSTM standar untuk prediksi <i>return</i> H+1 SMH serta belum mengevaluasi harga hasil rekonstruksi dari <i>return</i> .
4	Penelitian ini	Prediksi harga VanEck <i>Semiconductor</i> ETF (SMH) pada horizon satu hari perdagangan (H+1).	Regresi multivariat <i>return</i> H+1 menggunakan model <i>stacked LSTM</i> . Data harian SMH dari Yahoo <i>Finance</i> periode 8 Juni 2021–17 April 2026 (1.221 observasi), dengan fitur <i>Open</i> , <i>High</i> , <i>Low</i> , <i>Close</i> , Volume, dan <i>Return</i> . Data dibagi kronologis 90% pelatihan dan 10% pengujian.	Prediksi <i>return</i> dikonversi menjadi harga prediksi menggunakan harga penutupan hari sebelumnya. Evaluasi pada harga aktual dan harga hasil rekonstruksi menggunakan MAE, RMSE, dan MAPE.	Mengisi celah melalui prediksi <i>return</i> kontinu H+1 pada SMH dan evaluasi kuantitatif harga hasil rekonstruksi.

Berdasarkan Tabel 2.1, penelitian terdahulu telah menunjukkan bahwa metode *deep learning* dan LSTM memiliki potensi yang baik dalam prediksi pada sektor semikonduktor dan pasar keuangan. Namun, masih terdapat kesenjangan penelitian yang dapat diisi melalui penelitian ini, yaitu belum banyak penelitian yang secara spesifik membangun model LSTM untuk memprediksi *return* H+1 pada

VanEck *Semiconductor* ETF (SMH) sebagai nilai kontinu, menggunakan enam fitur historis, dan kemudian mengonversi hasil prediksi tersebut menjadi harga prediksi.

Kesenjangan penelitian tersebut menjadi dasar dilakukannya penelitian ini. Dengan demikian, kontribusi utama penelitian ini adalah membangun model prediksi *return* SMH berbasis LSTM, mengonversi hasil prediksi menjadi harga, serta mengevaluasi performa model menggunakan MAE, RMSE, dan MAPE agar hasil prediksi dapat dibandingkan secara objektif.

Berdasarkan penelitian terdahulu, masih terdapat beberapa kesenjangan. Pertama, Chen (2025) berfokus pada klasifikasi arah harga SMH H+5, bukan prediksi *return* kontinu. Kedua, Yen dan Chen (2025) memprediksi tren fundamental TSMC dengan data kuartalan dan tidak melaporkan evaluasi *error* regresi standar pada data uji. Ketiga, Fei dan Li (2026) menggunakan model Bi-LSTM–Bi-GRU pada beberapa ETF non-SMH dengan pembagian data 60:20:20.

Belum ditemukan penelitian pada referensi yang digunakan yang secara bersamaan menguji LSTM standar untuk prediksi *return* kontinu SMH, membandingkan beberapa panjang *window*, menguji beberapa rasio pembagian data kronologis, serta menganalisis horizon H+1, H+5, dan H+10. Penelitian ini berupaya mengisi kesenjangan tersebut melalui pendekatan regresi multivariat berbasis LSTM, rekonstruksi *return* menjadi harga, evaluasi MAE, RMSE, dan MAPE, serta analisis perilaku pergerakan *return*.

2.7 *Windowing pada Time Series*

Dalam prediksi *financial time series*, data historis biasanya diubah menjadi bentuk sekuens menggunakan teknik *sliding window*. *Sliding window* merupakan metode pembentukan *dataset* dengan mengambil sejumlah observasi sebelumnya sebagai *input* untuk memprediksi nilai pada periode berikutnya. Teknik ini memungkinkan model *deep learning* seperti *Long Short-Term Memory* (LSTM) mempelajari pola temporal dari pergerakan harga secara berurutan [15].

Penggunaan *historical window* bertujuan memberikan konteks informasi masa lalu kepada model. Pada prediksi harga saham atau *Exchange Traded Fund* (ETF), model tidak hanya melihat satu titik data, tetapi rangkaian harga dalam beberapa hari sebelumnya. Pendekatan ini membantu model menangkap tren baik jangka pendek maupun pola volatilitas yang sering muncul dalam *financial time series* [20].

Dalam banyak penelitian, panjang *window* yang digunakan berkisar antara

30 hingga 90 hari perdagangan. Dalam penelitian ini, panjang *window* tidak langsung ditetapkan hanya berdasarkan pertimbangan teoritis. Tiga variasi *window*, yaitu 30, 60, dan 90 hari perdagangan, digunakan untuk menguji pengaruh jumlah informasi historis terhadap performa model. *Window* 30 hari merepresentasikan konteks jangka pendek, *window* 60 hari merepresentasikan konteks jangka menengah, dan *window* 90 hari merepresentasikan konteks historis yang lebih panjang. *Window* terbaik dipilih berdasarkan hasil evaluasi MAE, RMSE, dan MAPE. [17].

Dengan demikian, teknik *windowing* menjadi tahap penting dalam *preprocessing* data *financial time series* sebelum diterapkan pada model LSTM, karena menentukan kualitas representasi pola historis yang dipelajari model dalam proses *forecasting*.

2.8 Optimizer Adam

Optimizer Adam (Adaptive Moment Estimation) merupakan sebuah algoritma optimisasi yang digunakan untuk memperbarui parameter model selama proses pelatihan *deep learning*. Tujuan utama *optimizer* adalah meminimalkan nilai fungsi kesalahan (*loss function*) sehingga hasil prediksi model menjadi semakin mendekati data aktual.

Adam dipilih karena memiliki kemampuan menyesuaikan besar pembaruan parameter secara otomatis pada setiap iterasi. Dibandingkan metode optimisasi dasar seperti *gradient descent*, Adam mampu menghasilkan proses *training* yang lebih cepat, stabil, dan efisien. Selain itu, Adam dapat mengurangi perubahan parameter yang terlalu besar sehingga membantu model mencapai kondisi konvergen, yaitu keadaan ketika nilai kesalahan model sudah semakin kecil dan relatif stabil [15].

Dalam model *Long Short-Term Memory (LSTM)*, *optimizer Adam* banyak digunakan karena efektif dalam menangani data *time series* yang memiliki pola kompleks dan fluktuatif. Karakteristik ini sesuai dengan data harga VanEck Semiconductor ETF (SMH), yang memiliki perubahan harga tidak stabil dan dipengaruhi oleh berbagai faktor pasar.

Pada penelitian ini, *optimizer Adam* digunakan selama proses pelatihan model LSTM untuk membantu model mempelajari pola historis harga secara lebih optimal dan meningkatkan akurasi prediksi pada periode berikutnya [22].

2.9 Loss Function Mean Squared Error (MSE)

Mean Squared Error (MSE) merupakan sebuah fungsi *loss* yang umum digunakan pada permasalahan regresi dan *forecasting* numerik. Fungsi *loss* adalah ukuran yang biasa digunakan untuk mengetahui seberapa jauh hasil prediksi model berbeda dari data aktual.

Pada model *Long Short-Term Memory* (LSTM), MSE digunakan selama proses *training* untuk mengukur tingkat kesalahan prediksi pada setiap iterasi. Nilai MSE diperoleh dari rata-rata selisih antara nilai prediksi dan nilai aktual, di mana selisih yang lebih besar akan memberikan nilai kesalahan yang lebih besar pula.

MSE dipilih dalam penelitian ini karena sesuai digunakan untuk prediksi harga pada data *financial time series*, termasuk harga VanEck Semiconductor ETF (SMH). Selain itu, MSE mampu memberikan penalti yang lebih besar terhadap kesalahan prediksi yang tinggi, sehingga model terdorong untuk menghasilkan prediksi yang semakin mendekati data aktual [22].

Semakin kecil nilai MSE yang dihasilkan, maka semakin baik performa model dalam memprediksi harga pada periode berikutnya [20].

2.10 Evaluasi Model Prediksi

Evaluasi model bertujuan untuk menilai tingkat kemampuan model dalam menghasilkan prediksi yang sesuai dengan data aktual. Pada prediksi *financial time series*, evaluasi perlu dilakukan menggunakan beberapa metrik karena setiap metrik memiliki karakteristik yang berbeda dalam mengukur distribusi error [16, 10]. Pada penelitian ini, performa model prediksi harga VanEck Semiconductor ETF (SMH) dievaluasi dengan membandingkan hasil prediksi dengan nilai sebenarnya menggunakan tiga metrik evaluasi, yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE).

2.10.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) adalah metrik yang digunakan untuk menghitung rata-rata nilai kesalahan absolut antara hasil prediksi dan data aktual. Nilai MAE menggambarkan besarnya deviasi rata-rata prediksi model terhadap nilai sebenarnya dalam satuan yang sama dengan data yang diprediksi. Semakin rendah nilai MAE yang diperoleh, semakin kecil rata-rata kesalahan prediksi yang

dihasilkan, sehingga menunjukkan bahwa performa model semakin baik dalam mendekati nilai aktual [10, 17].

2.10.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) adalah metrik evaluasi yang diperoleh dengan menghitung akar kuadrat dari nilai *Mean Squared Error* (MSE), sehingga hasil akhirnya kembali berada pada satuan yang sama dengan data yang diprediksi. Oleh karena itu, RMSE lebih mudah dipahami dan diinterpretasikan dibandingkan MSE, terutama pada penelitian yang berfokus pada prediksi harga. Nilai RMSE yang semakin kecil menunjukkan bahwa selisih antara hasil prediksi dan data aktual semakin rendah, sehingga mengindikasikan bahwa model memiliki tingkat akurasi yang lebih baik [16, 15].

2.10.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) adalah metrik evaluasi yang digunakan untuk menghitung rata-rata tingkat kesalahan prediksi dalam bentuk persentase terhadap nilai aktual. Penyajian hasil dalam satuan persen membuat metrik ini lebih mudah dipahami dan memudahkan proses evaluasi performa model. Sebagai ilustrasi, apabila model menghasilkan nilai MAPE sebesar 5%, maka dapat diartikan bahwa rata-rata selisih antara hasil prediksi dan nilai aktual mencapai 5% dari nilai sebenarnya [10, 20].

Semakin kecil nilai MAE, RMSE, dan MAPE, semakin baik kemampuan model dalam memprediksi harga VanEck *Semiconductor* ETF (SMH).

2.11 Data Preprocessing

Data *preprocessing* merupakan tahap persiapan data sebelum digunakan dalam proses pelatihan model *machine learning*. Tahap ini meliputi pemeriksaan data, penanganan nilai yang tidak lengkap, normalisasi, pembagian data, dan pembentukan struktur data yang sesuai dengan kebutuhan model. *Preprocessing* diperlukan agar data memiliki format dan skala yang sesuai sehingga proses pembelajaran model dapat berjalan lebih stabil [15, 17, 20]. Tahapan *preprocessing* dalam penelitian ini meliputi:

2.11.1 Penanganan Missing Value

Data yang hilang (*missing value*) diisi menggunakan metode *forward-fill*, yaitu mengisi nilai yang kosong dengan nilai terakhir yang tersedia [16, 17].

2.11.2 Normalisasi

Data dinormalisasi agar berada pada skala yang sama. Normalisasi membuat proses pembelajaran lebih stabil dan cepat. [15, 20] Penelitian ini menggunakan *MinMaxScaler*, dengan rumus:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2.1)$$

di mana:

- x = nilai asli
- $x_{\min}, x_{\max}$ = nilai minimum dan maksimum *dataset*
- x' = nilai setelah dinormalisasi

2.11.3 Train-Test Split

Train-test split merupakan metode yang digunakan untuk membagi *dataset* menjadi dua bagian, yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Data pelatihan dimanfaatkan untuk membangun serta melatih model agar mampu mempelajari pola dari data yang tersedia. Sementara itu, data pengujian digunakan untuk mengukur kemampuan model dalam menghasilkan prediksi terhadap data yang tidak pernah digunakan selama proses pelatihan [16, 10].

Pada data *time series*, pembagian data dilakukan secara kronologis tanpa pengacakan untuk memastikan data masa lalu digunakan sebagai data *training* dan data masa depan digunakan sebagai data *testing*. Selain penggunaan satu rasio pembagian data, pengujian beberapa rasio dapat dilakukan untuk menilai kestabilan kemampuan generalisasi model pada panjang periode *testing* yang berbeda.

2.11.4 Reshaping untuk LSTM

Model LSTM membutuhkan *input* berbentuk 3 dimensi (*samples, timesteps, features*):

- ***samples***: jumlah observasi
- ***timesteps***: jumlah data sebelumnya yang digunakan untuk prediksi
- ***features***: jumlah variabel *input Reshape* ini membuat data menjadi “barisan” agar LSTM bisa membaca pola urutan harga harian dan memprediksi harga berikutnya[15, 17].

