

BAB III

METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek yang diteliti pada penelitian ini adalah prediksi harga penutupan saham pada periode waktu harian. Dengan menggunakan dataset historis harga saham, penelitian ini akan menganalisis keefektifan dan keunggulan masing-masing model dalam menangkap pola kompleks dalam data. Hasil dari penelitian ini diharapkan dapat memberikan wawasan mendalam terkait dengan performa relatif kedua model, sehingga dapat memberikan kontribusi signifikan dalam pengembangan metode prediksi harga saham menggunakan *deep learning*.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Terdapat beberapa metodologi yang biasa digunakan dalam melakukan *data mining* seperti CRISP-DM, SEMMA, dan KDD. Masing-masing metodologi memiliki kelebihan dan kekurangan sehingga metodologi terbaik dapat berbeda-beda untuk setiap kasus. Berikut adalah perbandingan untuk masing-masing metodologi [26].

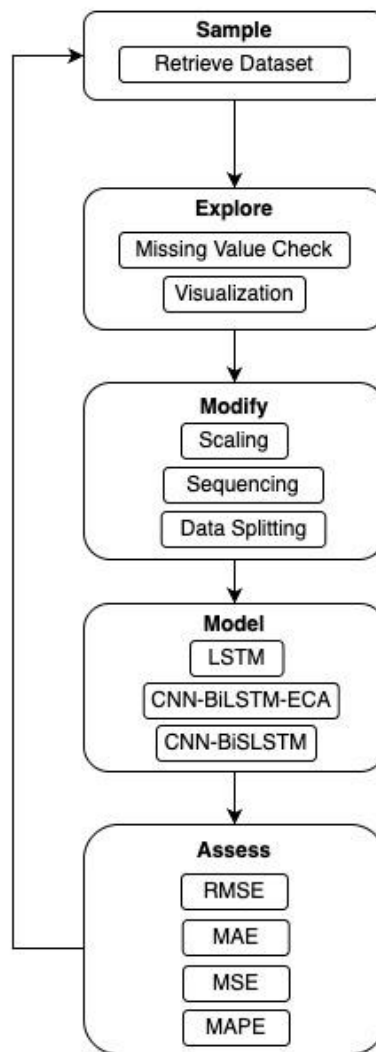
Tabel 3. 1 Kelebihan dan Kekurangan Metodologi

Metodologi	Kelebihan	Kekurangan
CRISP-DM	Mencakup semua aspek dalam melakukan <i>data mining</i>	Membutuhkan waktu yang lama dalam pelaksanaannya
SEMMA	Fokus kepada sisi teknis dari <i>data mining</i>	Tidak ada perspektif dari sisi bisnis

KDD	Metode yang paling tua sehingga paling <i>mature</i>	Tidak ada perspektif dari sisi bisnis
-----	--	---------------------------------------

Penelitian ini berfokus pada perbandingan kinerja dua model dalam memprediksi harga saham berdasarkan data historis, sehingga membutuhkan kerangka kerja yang lebih fokus terhadap aspek teknis. Oleh karena itu, metodologi SEMMA dipilih karena dirancang secara sistematis untuk mendukung proses eksplorasi dan analisis teknis yang mendalam serta memungkinkan eksperimen dan validasi model dilakukan secara efisien dan berulang. Berbeda dengan CRISP-DM yang mencakup seluruh siklus data mining sehingga juga terdapat aspek bisnis yang kurang relevan dalam konteks penelitian ini dan bahkan dapat memperpanjang waktu pelaksanaan. Sementara itu, meskipun KDD merupakan metodologi yang sudah ada sejak lama, ia tidak menyediakan panduan sekomprehensif SEMMA dalam tahapan eksplorasi data dan evaluasi kuantitatif model, yang merupakan inti dari penelitian ini. Dengan struktur yang jelas, fokus pada aspek teknis, dan kemudahan dalam mengevaluasi performa prediksi, SEMMA menjadi metodologi yang paling sesuai untuk mendukung tujuan penelitian ini. Gambar 3. 1 merupakan alur penelitian yang diikuti dalam penelitian ini didasarkan pada metodologi SEMMA.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A



Gambar 3. 1 Diagram metodologi SEMMA

Sebagaimana digambarkan pada gambar 3.1, metodologi ini terdiri dari 5 tahap, yaitu *sample*, *explore*, *modify*, *model*, dan *assess*. Berikut adalah penjelasan selengkapnya untuk masing-masing tahapan [27].

3.2.1.1 Sample

Pada tahap ini, data sampel diambil dari dataset yang lebih besar. Ini melibatkan pemilihan subset data yang akan digunakan dalam analisis. Data yang digunakan sebagai sampel diharapkan bisa menjadi perwakilan untuk keseluruhan data. Penelitian ini akan menggunakan data historis saham dari perusahaan Apple Inc, Microsoft Corporation, dan Alphabet Inc. Data

saham yang digunakan akan mencakup periode mulai dari tanggal saham masing-masing perusahaan melantai di bursa hingga tanggal 24 Januari 2025. Untuk saham AAPL, periode dimulai sejak 12 Desember 1980, saham GOOG sejak 19 Agustus 2004, dan saham MSFT sejak 13 Maret 1986. Dengan rentang waktu ini, saham AAPL terdiri dari 11120 baris data, saham GOOG terdiri dari 5142 baris data, MSFT terdiri dari 9794 baris data, S&P 500 terdiri dari 24382 baris data, DJI terdiri dari 8326 baris data, dan NASDAQ terdiri dari 13608 baris data. Masing-masing dataset akan terdiri dari 6 kolom, yaitu Date, Open, High, Low, Close, dan Volume. Kolom Date berisi informasi tentang tanggal perdagangan saham pada hari tersebut. Kolom Open mencatat harga pembukaan saham pada awal sesi perdagangan. Kolom High menunjukkan harga tertinggi yang dicapai oleh saham selama satu hari perdagangan. Kolom Low berisi harga terendah yang dicapai pada hari yang sama. Kolom Close merepresentasikan harga penutupan saham pada akhir sesi perdagangan. Terakhir, kolom Volume menunjukkan jumlah saham yang diperdagangkan pada hari tersebut.

3.2.1.2 Explore

Pada tahap ini akan dilakukan eksplorasi terhadap data sampel yang sudah diambil pada tahap sebelumnya. Eksplorasi data dapat diperkaya dengan membuat visualisasi data dan penerapan metode statistika. Tujuan dari tahap ini adalah untuk mendapatkan gambaran komprehensif terhadap dataset yang akan digunakan. Grafik garis dari pergerakan masing-masing harga saham dari waktu ke waktu juga dapat dibuat untuk memvisualisasikan data. Selain itu, dilakukan juga pengecekan terhadap *missing value*.

3.2.1.3 Modify

Sesuai namanya, pada tahap ini data akan dimodifikasi agar dapat lebih sesuai untuk dilakukan pemodelan. Salah satu hal yang dilakukan pada tahap ini adalah melakukan skala ulang terhadap nilai-nilai yang ada pada data. MinMaxScaler digunakan dalam penelitian ini untuk memastikan nilai

pada dataset berada di antara 0 hingga 1 berdasarkan nilai minimum dan maksimum pada kolom tersebut. Proses ini dilakukan untuk memastikan keseragaman skala antar fitur sehingga variabel dengan nilai absolut besar tidak secara tidak proporsional mempengaruhi model. Dengan rentang nilai yang seragam, fungsi *loss* pada model dapat beroperasi lebih stabil dan mencapai konvergen lebih cepat. Tidak seperti StandardScaler yang menggeser data ke rata-rata nol dan deviasi baku satu tanpa batas rentang, yang bisa menghasilkan nilai input negatif atau di luar jangkauan ideal aktivasi model, MinMaxScaler mempertahankan distribusi asli data sehingga hubungan relatif antar titik data tidak berubah, sehingga pola naik turun harga saham tetap terjaga [10].

Proses *timestepping* pada penelitian ini merupakan teknik pengelompokan data historis menjadi beberapa window waktu atau interval sejumlah *time step* yang ditentukan. *Time step* yang digunakan pada penelitian kali ini adalah 10. Artinya model akan menggunakan data historis 10 hari kebelakang untuk memprediksi harga saham di hari selanjutnya. Nilai *time step* ini dipilih berdasarkan hasil dari penelitian terdahulu yang membuktikan nilai ini bisa meminimalisir tingkat *error* dari model [5], [12]. Proses ini dilakukan secara *sliding*, sehingga seluruh data historis bisa diubah menjadi banyak pasangan input dan target. Setelah proses *timestepping*, data kemudian dibagi menjadi dua bagian, yaitu 80% sebagai dataset training dan 20% sebagai dataset validasi. Rasio ini digunakan karena hal ini cukup umum pada penelitian-penelitian yang sudah pernah dilakukan dan juga terbukti sebagai rasio yang dapat memberikan hasil terbaik [28], [29].

3.2.1.4 Model

Data yang sudah dimodifikasi pada tahap sebelumnya akan digunakan untuk melatih model yang sudah ditentukan. Tujuan dari tahap ini adalah membuat model yang dapat memprediksi dengan baik untuk menyelesaikan permasalahan utama. Dalam penelitian ini, terdapat tiga

model yang akan dilatih, yaitu CNN-BiSLSTM, CNN-BiLSTM-ECA, serta LSTM yang digunakan sebagai model dasar untuk perbandingan kinerja. LSTM dipilih menjadi tolak ukur karena model ini juga diuji pada masing-masing penelitian yang mengusulkan model CNN-BiSLSTM dan CNN-BiLSTM-ECA.

Masing-masing model ini kemudian dilakukan *hyperparameter tuning*. Konfigurasi seperti jumlah unit LSTM, jumlah unit dense, *learning rate*, *loss function*, jumlah filter CNN, dan jumlah *kernel* CNN akan diubah-ubah secara programatis untuk mencari konfigurasi terbaik. Metode yang digunakan untuk melakukan proses ini adalah Bayesian Optimization. Metode ini dipilih karena dapat mencari konfigurasi yang cukup baik dengan jumlah percobaan yang lebih sedikit. Hal ini bisa terjadi karena Bayesian Optimization menggunakan probabilitas untuk memprediksi nilai dari titik yang belum pernah diuji [30]. Masing-masing model memiliki kesempatan untuk mencoba 50 konfigurasi yang berbeda. Nilai ini diambil dengan pertimbangan sumber daya komputasi yang ada serta waktu yang diperlukan untuk menjalankan proses *tuning*. *Epoch* yang digunakan untuk setiap model adalah 50 dengan *batch size* 64 dan juga terdapat *early stopping* dengan nilai *patience* 15 yang memperhatikan nilai *validation loss*. *Early stopping* digunakan untuk menghentikan proses *training* jika nilai *validation loss* tidak kunjung membaik dalam 15 iterasi terakhir.

Untuk memantau apakah model mengalami *overfitting*, *fixed validation set* yang terdiri dari 20% data terakhir dari dataset asli akan digunakan sepanjang proses pelatihan. Setelah pelatihan, performa model akan dibandingkan nilai *loss* dan *validation loss* untuk memeriksa adanya indikasi *overfitting*. Jika terdapat perbedaan signifikan antara nilai *loss* pada data pelatihan dan *validation loss*, ini menunjukkan adanya *overfitting*, di mana model mungkin terlalu menyesuaikan diri dengan data pelatihan dan tidak mampu menggeneralisasi dengan baik pada data yang tidak terlihat sebelumnya.

3.2.1.5 Assess

Model yang sudah dilatih kemudian akan dinilai dan dibandingkan pada tahap ini. Tujuan dari tahap ini adalah untuk mengetahui performa model dalam menjalankan fungsinya. Penilaian dapat dilakukan menggunakan metrik-metrik evaluasi yang sesuai dengan permasalahan yang ingin dipecahkan. Metrik-metrik yang akan digunakan untuk menilai masing-masing model adalah Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), dan Mean Absolute Error (MAE).

3.2.2 Pemilihan Algoritma *Deep Learning*

Terdapat dua algoritma *deep learning* yang digunakan dalam penelitian ini yaitu CNN-BiSLSTM dan CNN-BiLSTM-ECA. CNN-BiSLSTM berhasil mendapatkan skor tertinggi pada studinya dengan nilai RMSE sebesar 162,53164 yang dilatih menggunakan data harga *Shenzen Component Index* [6]. Berikut adalah tabel perbandingan nilai RMSE dengan arsitektur lain.

Tabel 3. 2 Perbandingan nilai RMSE model CNN-BiSLSTM [6]

Model	RMSE
LSTM	181,01318
CNN-BiSLSTM	162,53164

Hal yang serupa juga berlaku untuk CNN-BiLSTM-ECA di mana nilai RMSE secara konsisten lebih baik dibandingkan dengan model LSTM. Pada Shanghai Composite Index, CNN-BiLSTM-ECA mencapai RMSE sebesar 44,227, lebih rendah dibandingkan LSTM yang sebesar 64,978. Untuk China Unicom, CNN-BiLSTM-ECA memiliki RMSE 0,167, sementara LSTM mencapai 0,189. Begitu pula pada CSI 300, nilai RMSE CNN-BiLSTM-ECA adalah 58,604, lebih baik dari LSTM yang sebesar

76,218. Tabel 3.3 menunjukkan perbandingan nilai RMSE CNN-BiLSTM-ECA dengan arsitektur LSTM.

Tabel 3. 3 Perbandingan nilai RMSE model CNN-BiLSTM-ECA [5]

Model	RMSE		
	Shanghai Composite Index	China Unicom	CSI 300
LSTM	64.978	0.189	76.218
CNN-BiLSTM-ECA	44.227	0.167	58.604

3.3 Teknik Pengumpulan Data

3.3.1 Sumber Data

Sumber data yang digunakan adalah Yahoo Finance. Yahoo Finance dipilih karena memiliki data saham historis yang lengkap untuk saham AAPL, GOOG, dan MSFT. Yahoo Finance menyediakan data saham dimulai dari hari pertama perusahaan terdaftar di bursa hingga saat ini. Data yang tersedia juga mudah untuk diunduh dengan beberapa pilihan periode waktu yang dapat disesuaikan. Pilihan periode waktu tersebut adalah harian, mingguan, dan bulanan.

Data saham ini didapatkan menggunakan *library* Python bernama *yfinance*. *Library* ini memungkinkan pengguna untuk mengunduh data harga saham secara otomatis langsung dari Yahoo Finance melalui kode program. Dengan menggunakan *yfinance*, data harga saham untuk AAPL, GOOG, dan MSFT diunduh sesuai dengan periode waktu yang dibutuhkan, kemudian disimpan dalam *format* CSV agar dapat digunakan untuk proses analisis lebih lanjut. Proses pengunduhan data dapat dilakukan dengan beberapa baris kode sederhana, sehingga memudahkan proses pengambilan data historis saham.

3.3.2 Data

Seluruh dataset yang digunakan dalam penelitian ini berasal dari sumber data yang sama sehingga format data dan jumlah kolom pada setiap dataset identik. Ketiga dataset memiliki format CSV atau Comma-Separated Values. Tabel 3.4 merupakan penjelasan untuk masing-masing kolom yang terkandung dalam masing-masing dataset.

Tabel 3. 4 Penjelasan Kolom Dataset

Nama Kolom	Penjelasan
Date	Tanggal yang merupakan titik waktu yang mengaitkan data harga saham dengan tanggal tertentu dalam <i>time series</i> .
Open	Harga saham pada saat awal periode waktu.
High	Harga tertinggi saham selama periode waktu.
Low	Harga terendah saham selama periode waktu.
Close	Harga saham pada saat akhir periode waktu.
Volume	Jumlah total saham yang diperdagangkan pada tanggal tersebut.

Adapun perbedaan untuk masing-masing dataset dimana terdapat perbedaan jumlah baris untuk masing-masing dataset dikarenakan perbedaannya waktu saat perusahaan-perusahaan ini mendaftar ke bursa efek setempat. Perusahaan Apple Inc dengan dataset AAPL memiliki data harga saham dari tanggal 12 Desember 1980 hingga 24 Januari 2025 dengan total 11120 baris data. Pada dataset GOOG terdapat 5142 baris data yang dimulai dari tanggal 19 Agustus 2004 hingga 24 Januari 2025. Adapun Microsoft Corporation memiliki 9794 baris data yang dimulai dari tanggal 13 Maret 1986 hingga 24 Januari 2025.

3.4 Teknik Pengambilan Sampel

Ketiga data saham yang akan digunakan pada penelitian ini didapatkan melalui *library* Python bernama *yfinance*. Library ini mempermudah pengambilan

data saham yang disediakan oleh Yahoo Finance [31]. Data yang sudah diambil kemudian akan disimpan ke dalam komputer dengan format CSV. Konfigurasi dari data yang diambil memiliki interval waktu harian.

3.5 Teknik Analisis Data

Model-model yang sudah dilatih kemudian akan dievaluasi menggunakan metrik evaluasi seperti RMSE, MSE, MAPE, dan MAE. Selain itu, hasil prediksi masing-masing model juga akan divisualisasikan bersamaan dengan harga aktual dari masing-masing saham. Untuk menjalankan seluruh proses penelitian ini akan digunakan Python sebagai bahasa pemrograman utama bersama Jupyter Notebook yang ditulis menggunakan Microsoft Visual Studio Code sebagai *text editor*.

