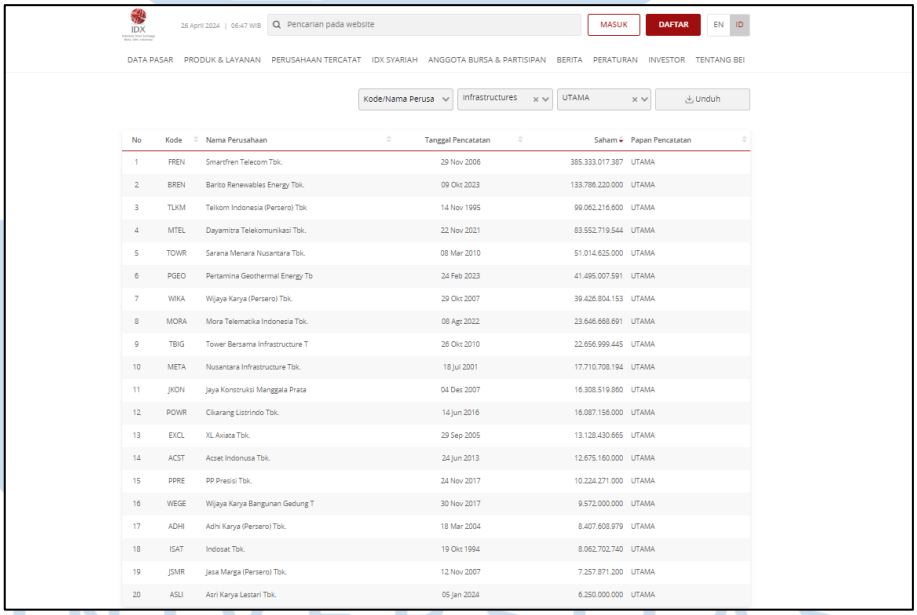
BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini dilakukan untuk melakukan analisis prediksi harga saham pada sektor infrastruktur. Data yang digunakan dalam penelitian ini adalah data historis harga saham sektor infrastruktur yang diambil dari *website* Yahoo Finance. Adapun data yang digunakan adalah data harga saham dari 4 perusahaan sektor infrastruktur di Indonesia, yaitu PT XL Axiata Tbk. (EXCL.JK), PT Jasa Marga (Persero) Tbk (JSMR.JK), PT Telkom Indonesia (Persero) Tbk (TLKM.JK), dan PT Smartfren Telecom Tbk (FREN.JK). Periode data historis yang digunakan adalah 16 tahun dari 1 Januari 2008 hingga 1 Maret 2024. Keempat dataset memiliki 7 variabel, yaitu *date* (indeks data), *open, high, low, close, adj. close,* dan *volume* (OHLC & *Volume*).



Gambar 3. 1 Top 20 Daftar Saham Sektor Infrastruktur Papan Utama

Keempat perusahaan sektor infrastruktur tersebut dipilih dengan alasan keempat kode sahamnya tercatat dalam 20 besar saham sektor infrastruktur dalam papan pencatatan utama dengan jumlah saham terbesar di Bursa Efek Indonesia. Keempat data saham memiliki variasi rentang nilai saham yang

sejenis dan pergerakan yang variatif. Selain itu, data historis saham memiliki periode yang cukup untuk kebutuhan analisis pada penelitian ini.

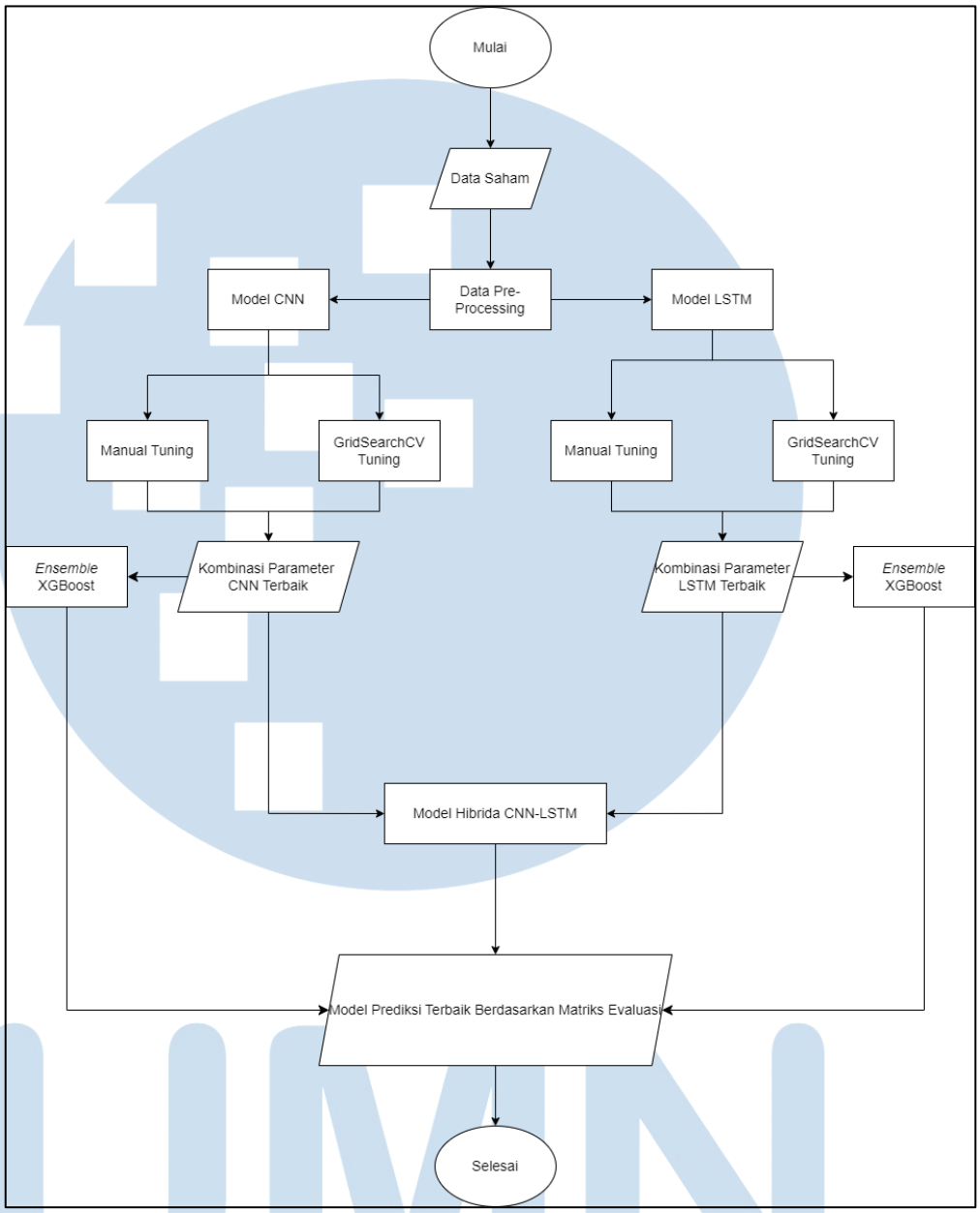
Pada penelitian ini, data saham akan diterapkan pada model prediksi menggunakan model algoritma LSTM dan CNN. Penelitian ini akan menghasilkan perbandingan performa dari ketiga model dengan penerapan *Grid Search* sebagai metode *hyperparameter tuning,* algoritma *ensemble* XGBoost, dan algoritma hibrida CNN-LSTM untuk upaya optimasi model. Pemilihan model terbaik dilakukan dengan melihat tingkat *error* yang paling rendah berdasarkan nilai RMSE, MSE, MAE, dan MAPE.

3.2 Metode Penelitian

Penelitian ini merupakan penelitian kuantitatif di mana penelitian ini menghasilkan data numerik dengan menerapkan analisis statistik dalam pengolahan datanya [101]. Penelitian ini menerapkan dua metode penelitian, yaitu *data mining* dan *solving problem*. Berikut adalah penjelasan untuk alur penelitian dan kedua metode penelitian yang digunakan:

3.2.1 Alur Penelitian

Alur penelitian merupakan suatu susunan rencana penelitian yang akan dilakukan. Susunan alur ini digunakan sebagai acuan dalam penelitian agar proses penelitian dapat dilakukan secara terstruktur. Alur penelitian dijelaskan dalam diagram sebagai berikut:



Gambar 3. 2 Alur Penelitian

Gambar 3.2 menggambarkan alur penelitian yang dimulai dari data saham sebagai *input.* Data dipersiapkan pada tahap *pre-processing* agar memiliki kondisi yang sudah siap untuk dimasukkan ke dalam model. Data diolah dengan menggunakan dua algoritma tunggal, yaitu LSTM dan CNN. Pada setiap model, diberlakukan dua jenis *hyperparameter tuning*, yaitu secara manual dan otomatis menggunakan *GridSearchCV.* Penggunaan dua metode ini dilakukan

dengan alasan untuk memvalidasi hasil kombinasi parameter terbaik. Adapun *tuning* manual menggunakan kombinasi yang ada pada *parameter grid* untuk *GridSearchCV*. Model dengan hasil metode *tuning* terbaik akan diterapkan algoritma *ensemble* XGBoost untuk meningkatkan performa model tunggal. Selain itu, kombinasi parameternya diambil untuk digunakan sebagai parameter model hibrida CNN-LSTM. Adapun tujuan *ensemble* dan hibrida ini adalah untuk mengoptimasi hasil model tunggal. Hasil performa seluruh model dievaluasi menggunakan matriks evaluasi dan model terbaik untuk setiap data saham diambil berdasarkan nilai RMSE, MSE, MAE, dan MAPE terendah.

3.2.2 Metode *Data Mining*

Penelitian ini menggunakan metode *data mining Cross-Industry Process for Data Mining* (CRISP-DM). Adapun penggunaan metode ini berdasarkan hasil perbandingan dari metode *data mining* serupa yaitu *Sample, Explore, Modify, Model, and Asses* (SEMMA), *Knowledge Discovery in Databases* (KDD), dan *Cross-Industry Process for Data Mining* (CRISP-DM) [102]. Perbandingan ketiga metode tersebut ditunjukkan pada Tabel 3.1.

Tabel 3. 1 Perbandingan Metode SEMMA, CRISP-DM, dan KDD

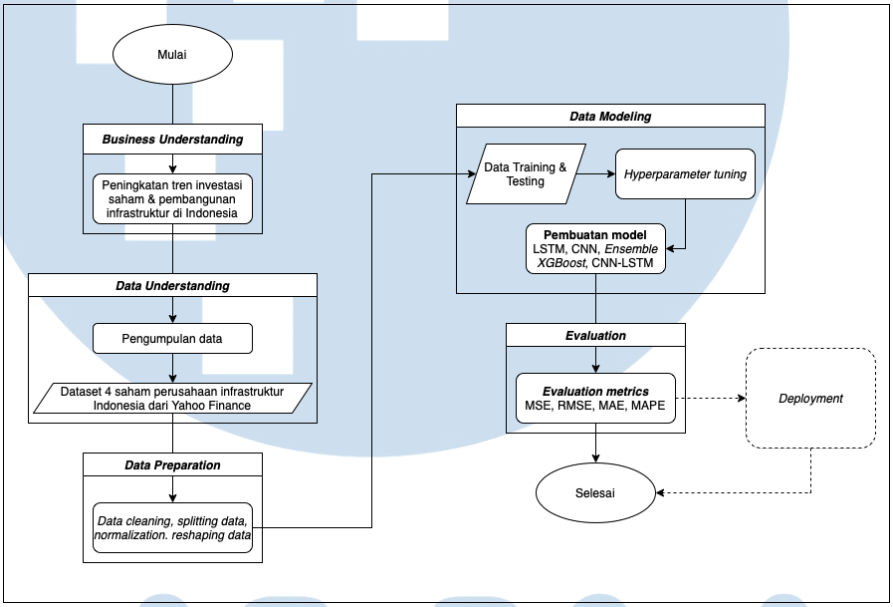
Metode *Data Mining*	**SEMMA**	**CRISP-DM**	**KDD**
Jumlah Fase	5	6	7
Tahap [72], [77]	-	*Business Understanding*	*Pre-KDD*
	Sample	*Data Understanding*	*Selection*
	Explore		*Pre-Processing*
	Modify	*Data Preparation*	*Transformation*
	Model	*Data Modeling*	*Data Mining*
	Assessment	*Evaluation*	*Interpretation/Evaluation*
	-	*Deployment*	*Post-KDD*
Kelebihan[72], [77]	○ Proses iteratif ○ Fleksibel pada *data exploration* dan *modification*	○ Proses iteratif ○ Sudah umum digunakan ○ Berfokus pada	○ Proses iteratif ○ Berfokus pada *data processing* dan *knowledge discovery* ○ *Human-centered approach* ○ Hasil lebih baik sebab

		business understanding ○ Memiliki mekanisme umpan balik ○ Memiliki struktur yang jelas	dapat dilakukan secara repetitif
Kekurangan[72], [77]	○ Didesain untuk bekerja pada SAS Enterprise Miner ○ Tidak terlalu memperhitungkan *business understanding*	○ Fase *data preparation* dan *modeling* berbeda dari *data mining* tradisional ○ Dapat memakan banyak waktu	○ Kompleksitas tinggi ○ Memakan banyak waktu sebab fasenya selalu mencari *return points*
Ringkasan	Fasenya lebih singkat sehingga masih memerlukan fase lain agar lebih sempurna.	Fasenya merupakan ringkasan dari fase-fase pada KDD.	Fasenya kompleks dan lengkap namun masih bisa disederhanakan.

Tabel 3.1 menampilkan perbandingan metode SEMMA, CRISP-DM, dan KDD. Masing-masing *framework* ini mempunyai tahap serta tujuan implementasinya tersendiri. Berdasarkan evaluasi kelebihan dan kekurangan yang tercatat dalam Tabel 3.1, CRISP-DM menawarkan panduan yang terstruktur dan sistematis dalam proses eksplorasi data. Dibandingkan dengan metode lain seperti KDD dan SEMMA, CRISP-DM menonjol dengan penekanan yang signifikan pada tahap evaluasi dan implementasi. Tahapan-tahapan yang ada pada CRISP-DM membuatnya sangat lengkap dan terdokumentasi. Selain itu, tahapannya juga teratur, terstruktur, dan ditentukan sehingga memungkinkan penelitian dapat dipahami atau diperbaiki dengan mudah.

Penelitian ini menggunakan *framework* CRISP-DM sebab keenam tahap tersebut tidak kaku, yaitu tahapnya bisa dilakukan secara maju

dan mundur. Hasil dari setiap tahap akan menentukan tahap apa yang harus dilakukan selanjutnya [74]. CRISP-DM telah umum digunakan sebab kepraktisannya dan fleksibilitasnya dalam pengembangan solusi *machine learning,* serta berfokus pada *business* understanding [69], [103]. Penelitian ini memiliki tujuan dari segi bisnis, sehingga pemilihan *framework* CRISP-DM menjadi opsi yang paling sesuai. Gambar 3.3 menampilkan *diagram flowchart* detail untuk alur penelitian dengan CRISP-DM.



Gambar 3. 3 Diagram Alur CRISP-DM

Berikut adalah penjelasan alur penelitian CRISP-DM berdasarkan Gambar 3.3:

a. *Business Understanding*

Pada tahap ini, penelitian dimulai dengan mendefinisikan tujuan proyek melalui perspektif bisnis guna mengidentifikasi masalah *data mining* [104]. *Business understanding* pada penelitian ini adalah adanya tren investasi saham dan pembangunan infrastruktur di Indonesia. Adapun tujuan yang ditentukan adalah pembangunan model prediksi menggunakan data harga saham sektor infrastruktur perusahaan Indonesia. Data historis harga saham digunakan untuk memprediksi tren harga saham di masa depan sebagai *financial*

time series analysis [105]. Hasil model ini akan membantu pemegang saham untuk mengambil keputusan analisis saham agar meminimalisir risiko kerugian. Penelitian ini akan menghasilkan model terbaik dengan nilai *error* terkecil dalam memprediksi nilai saham.

b. *Data Understanding*

Tahap ini merupakan proses pengumpulan data dan familiarisasi dengan data tersebut [104]. Tahap ini mencakup pengambilan data, penyimpanan data, pemilihan atribut, dan eksplorasi data. Pengumpulan empat dataset historis harga saham perusahaan infrastruktur Indonesia didapatkan melalui Yahoo! Finance. Setiap dataset memiliki periode 16 tahun dari 1 Januari 2008 hingga 1 Maret 2024 dengan data sebanyak 3.986 baris data serta kolom *Date, Open, High, Low, Close*, Adj. Close**,* dan *Volume.* Adapun total data yang digunakan adalah sebanyak 15.944 baris data. Setelah itu, dilakukan eksplorasi data untuk memahami karakteristik data.

c. *Data Preparation*

Pada tahap ini, data harga saham disiapkan menjadi bentuk dataset final sebelum masuk ke tahap pemodelan [104]. Secara berurutan, tahap ini melibatkan beberapa proses sebagai berikut:

1. Pembersihan data dengan menangani missing values. Pencarian *missing values* dilakukan menggunakan fungsi *isnull()*. Apabila ditemukan *missing values,* penghapusannya dilakukan menggunakan fungsi *dropna()* maupun imputasi data untuk mengganti nilai yang hilang [47]. Apabila tidak ditemukan, maka penghapusan maupun imputasi tidak akan dilakukan. Pada penelitian ini, apabila jumlah *missing value* tidak signifikan maka akan di-*drop*, apabila sebaliknya maka akan dilakukan imputasi data [106]. Penanganan *missing values* bertujuan untuk menghindari bias yang mengakibatkan penelitian menjadi kurang akurat [32], [47], [107].

2. Mengekstrak variabel target '*Close*' [23], [32] ke dalam *data frame closing_prices* sebagai data yang akan digunakan dan membagi dataset menjadi set pelatihan dan pengujian. Pada penelitian ini digunakan set 80% data pelatihan dan 20% data pengujian sebagaimana telah berhasil diterapkan pada penelitian [20], [23], [27], [108], [109].
3. *Reshape* data agar sesuai dengan bentuk i*nput* lalu dilakukan normalisasi data dengan alasan untuk menormalkan data agar seluruh fitur memiliki skala yang sama, yaitu dalam rentang 0 hingga 1. Normalisasi data dilakukan dengan fungsi *MinMaxScaler* sebagaimana telah berhasil diterapkan pada penelitian [23], [47], [108], [109].
4. *Reshape* data yang telah dinormalisasi ke dalam bentuk tiga dimensi untuk menyesuaikan bentuk *input* model.

d. *Data Modeling*

Tahap ini merupakan pembuatan model dan penyetelan parameter untuk diterapkan pada dataset [104]. Pembangunan model prediksi pada penelitian ini menggunakan algoritma LSTM dan CNN. Guna mengoptimasi hasil model, penentuan arsitektur model mencakup *hyperparameter tuning* seperti konfigurasi *learning rate*, *batch size, optimizer, activation,* jumlah *unit* LSTM, jumlah *filter* CNN, jumlah *unit* lapisan *dense*, ukuran *kernel,* dan ukuran *pool.* Adapun penentuan *hyperparameter* terbaik akan dilakukan dengan membandingkan hasil *tuning* manual dengan *tuning GridSearchCV*.

Penggunaan *Grid Search* sebagai upaya optimasi model dalam *hyperparameter tuning* telah dibuktikan pada beberapa penelitian yang telah dilakukan sebelumnya. Penelitian [53], [110], [111] menemukan bahwa penggunaan *Grid Search* meningkatkan hasil akurasi model dibandingkan menggunakan *Bayesian Search, Random Search,* maupun tanpa menerapkan *tuning*. Selain itu, upaya optimasi model juga dilakukan dengan menggunakan

algoritma *ensemble XGBoost* sebagaimana diterapkan pada penelitian [24], [25] dengan tujuan untuk menggali data secara lebih menyeluruh. Upaya optimasi lainnya juga dilakukan dengan melakukan hibrida kedua model tunggal menjadi model CNN-LSTM untuk menggabungkan kekuatan kedua model untuk menghasilkan performa yang lebih optimal.

e. *Evaluation*

Tahap ini merupakan tahap penilaian atau evaluasi hasil model berdasarkan tujuan bisnis yang ada [104]. Pada tahap ini model dilatih pada training set dan dievaluasi pada set pengujian menggunakan matriks evaluasi RMSE (*Root Mean Square Error*), MSE (*Mean Square Error*), MAE (*Mean Absolute Error*), dan MAPE sebagaimana diterapkan pada penelitian (*Mean Absolute Percentage Error*) [17], [20], [21], [22], [27].

Selain telah umum digunakan pada penelitian terdahulu, keempat matriks evaluasi ini digunakan dengan alasan keempatnya memiliki kelebihan dan kekurangannya masing-masing sehingga bereksperimen dapat membantu pemahaman data yang lebih baik [112]. MSE akan mengukur jarak kuadrat antara nilai asli dan prediksi, RMSE mengukur deviasi standar residu, MAE mengukur perbedaan absolut antara nilai aktual dan prediksi, dan MAPE mengukur persentase kesalahan rata-rata absolut. Hasil dari model CNN, LSTM, *ensemble* XGBoost, dan CNN-LSTM dibandingkan untuk mengevaluasi kinerjanya dan memilih model yang memiliki tingkat *error* paling kecil.

3.2.3 Metode *Solving Problem*

Pada metode *solving problem*-nya, penelitian ini akan menggunakan algoritma CNN, LSTM, *ensemble* XGBoost, dan algoritma hibrida CNN-LSTM. Berikut adalah tabel komparasi kelebihan dan kekurangan algoritma yang dipilih terhadap algoritma pembandingnya:

a. Perbandingan algoritma *Long Short-Term Memory* (LSTM) dengan algoritma *Recurrent Neural Network* (RNN), dan *Artificial Neural Network* (ANN) [108], [113], [114]

Tabel 3. 2 Perbandingan LSTM, RNN, ANN

Algoritma	Cara Kerja	Kelebihan	Kekurangan
LSTM	- *Recurrent Neural Network* (RNN) yang memiliki sel memori khusus. - Menangani masalah *gradient disappearance* pada RNN. - Memanfaatkan *input gate, output gate,* dan *forget gate*.	- Mampu menangani ketergantungan temporal yang kompleks. - Lebih adaptif terhadap perubahan dalam data.	- Memerlukan lebih banyak data pelatihan. - Kompleksitas yang lebih tinggi dan memerlukan sumber daya komputasi yang lebih besar.
RNN	- Mencakup sambungan siklik untuk menangani dependensi temporal.	- Efektif dalam memodelkan dependensi temporal. - Cocok untuk data berurutan seperti *time series*. - Menangani urutan data yang panjang.	- Rentan terhadap masalah *vanishing* atau *exploding gradient*. - Membutuhkan sumber daya komputasi yang lebih besar.
ANN	- Model adaptif berbasis data tanpa asumsi awal. - Digunakan sebagai model prediktif di berbagai bidang.	- Dapat menangani masalah nonlinear dan kompleks. - Tidak memerlukan asumsi awal yang kuat.	- Kompleks dan sulit diinterpretasi. - Memerlukan lebih banyak data pelatihan untuk menghindari *overfitting*.

Berdasarkan perbandingan pada Tabel 3.2, pilihan memilih LSTM dibandingkan RNN dan ANN didasarkan pada kemampuan LSTM menangani ketergantungan temporal kompleks, yang lebih unggul dalam memprediksi harga saham dengan pola jangka panjang. Kelebihan LSTM dalam hal ini

menjadi pertimbangan utama, terutama faktor perubahan harga saham dipengaruhi oleh faktor eksternal yang kompleks [113].

Selain itu, kompleksitas tinggi dan interpretasi sulit dari ANN menjadi hambatan. Sementara itu, RNN memiliki kekurangan dalam kerentanannya terhadap masalah *vanishing* atau *exploding gradient,* hal ini menyebabkan RNN tidak mahir menangkap *long-term dependencies* [108]. Fokus penelitian ini terletak pada prediksi harga saham, sehingga keunggulan LSTM dalam menangkap pola jangka panjang dan hubungan abstrak dalam deret waktu menjadi faktor utama algoritma ini dipilih. Meskipun memerlukan sumber daya komputasi lebih besar, manfaatnya dianggap lebih signifikan dalam meningkatkan akurasi prediksi [108], [113].

b. Pemilihan algoritma *Convolutional Neural Network* (CNN) [114]

Tabel 3. 3 Pemilihan Algoritma CNN

Algoritma	**Cara Kerja**	**Kelebihan**	**Kekurangan**
CNN	Melibatkan lapisan konvolusi untuk mengekstrak fitur spasial.	- Mampu menangkap pola spasial kompleks. - Baik untuk data dengan pola spasial. - Dapat mengurangi parameter.	- Kurang efektif untuk data temporal. - Lebih rentan terhadap *overfitting*.

Berdasarkan pada Tabel 3.3, CNN dipilih sebagai pelengkap model hibrida dengan LSTM dalam model prediksi harga saham karena CNN efektif dalam menangani pola spasial dan ekstraksi fitur kompleks. Dalam konteks harga saham, CNN lebih cocok untuk mengatasi variasi harga saham yang kompleks. Selain itu, CNN dapat mengurangi penggunaan parameter dan meningkatkan efisiensi model dengan *local perception* dan *weight sharing* [28], [29]. Kelebihan CNN juga terdapat pada

mengekstrak fitur abstrak dan mempelajari data dengan baik [30], [31]. Pada penelitian ini, CNN digunakan untuk mengambil fitur waktu pada data dan kemudian akan dilanjutkan LSTM dalam melakukan prediksi data. Kombinasi kedua algoritma ini memungkinkan pemanfaatan data yang bersifat *time sequence* sehingga dapat mengoptimalkan hasil prediksi harga saham [30]. Oleh karena itu, penggunaan LSTM diperkuat dengan integrasi CNN dalam model CNN-LSTM untuk memanfaatkan kekuatan kedua arsitektur tersebut [114].

c. Perbandingan metode *Grid Search* dengan *Random Search* [52], [53]

Tabel 3. 4 Perbandingan *Grid Search* dengan *Random Search*

Metode	Cara Kerja	Kelebihan	Kekurangan
Grid Search	Mencoba seluruh posibilitas kombinasi untuk kemudian dievaluasi seluruh hasilnya dan diambil hasil kombinasi terbaik.	- Memberikan kombinasi hasil paling optimal. - Dapat mengeksekusi data dengan akurasi yang lebih tinggi.	- Membutuhkan waktu yang relatif lama.
Random Search	Mencoba kombinasi secara acak, kemudian *hyperparameter* akan dievaluasi dan diambil hasil kombinasi terbaik.	- Bekerja lebih cepat. - Lebih efisien. - Dapat bekerja dengan data berdimensi besar.	- Tidak menjamin memberikan hasil paling optimal. - Terbatas pada distribusi pencarian.

Berdasarkan perbandingan pada Tabel 3.4, pilihan metode jatuh pada metode *Grid Search* dibandingkan dengan metode *Random Search.* Berdasarkan tujuan penelitian ini, *Grid Search* dipilih sebab tujuannya adalah melakukan optimalisasi hasil model. Selain itu, data yang digunakan pun tidak berdimensi terlalu besar. Metode *Grid Search* dipilih berdasarkan pada kelebihannya dalam memberikan kombinasi dengan hasil paling optimal. Selain itu, *Grid Search* dapat memberikan hasil akurasi

yang lebih tinggi sebab mencoba seluruh kemungkinan kombinasi yang ada [53].

d. Perbandingan metode algoritma *ensemble* XGBoost dengan LightGBM [25], [115], [116], [117]

Tabel 3. 5 Perbandingan XGBoost dengan LightGBM.

Metode	Cara Kerja	Kelebihan	Kekurangan
XGBoost	Menggabungkan *classifier* yang bersifat lemah ke dalam *classifier* yang kuat untuk menangkap pola yang lebih kompleks.	- Memberikan stabilitas dan generalisasi model. - Memberikan performa prediksi yang tinggi pada data berdimensi tinggi. - Telah umum digunakan sebagai ensemble prediksi *time series*.	- Membutuhkan memori yang besar. - Membutuhkan penyetelan *hyperparameter yang baik.*
LightGBM	Menggunakan pendekatan *leaf-wise.*	- Bekerja lebih cepat. - Lebih efisien. - Penggunaan memori yang lebih sedikit.	- Performa kurang baik apabila data tidak seimbang. - Performa kurang baik pada data berdimensi tinggi.

Tabel 3.5 menunjukkan perbandingan antara model *ensemble* XGBoost dan LightGBM. Kedua algoritma memiliki keunggulannya tersendiri namun memiliki tujuan yang sama, yaitu meningkatkan performa prediktif pada berbagai kasus regresi maupun klasifikasi. Pada penelitian ini, diterapkan XGBoost sebagai algoritma *ensemble* sebab kemampuannya dalam memberikan stabilitas dan generalisasi model yang diharapkan dapat mengoptimalisasi hasil performa model. Selain itu, penelitian terdahulu [24], [25] yang telah berhasil mengoptimalkan hasil model CNN-LSTM juga menjadi alasan terpilihnya algoritma ini.

3.3 Teknik Pengumpulan Data

Data yang digunakan pada penelitian ini bersifat data sekunder yang diambil langsung dari Yahoo! Finance sebagai platform keuangan yang menyediakan data saham. Yahoo! Finance dipilih dengan alasan portal tersebut memiliki reputasi yang dapat diandalkan dan memiliki ketersediaan data yang kaya. Adapun beberapa penelitian telah menggunakan data saham yang bersumber dari Yahoo! Finance [14], [15], [78], [106]. Penelitian ini menggunakan metode penelitian kuantitatif sebab berfokus pada pengumpulan dan analisis data numerik. Jumlah data yang diambil untuk analisis adalah sebanyak 15.944 baris data saham harian.

3.3.1 Populasi dan Sampel

Jumlah total perusahaan sektor infrastruktur yang tercatat pada BEI adalah 70 perushaan. Penelitian ini menetapkan populasi hanya kepada perusahaan yang tercatat pada papan pencatatan utama dengan total 36 perusahaan. Berdasarkan Peraturan Nomor I.11. tentang Ketentuan Khusus Pencatatan Saham di Papan Akselerasi, papan utama mencatat saham perusahaan berskala besar dengan waktu operasional yang cukup lama [118]. Dari total 36 perusahaan, penelitian ini mengerucutkan populasi menjadi 20 perusahaan berdasarkan jumlah saham terbanyak. Oleh karena itu, populasi pada penelitian ini adalah 20 perusahaan infrastruktur dengan jumlah saham terbanyak yang tercatat pada papan utama di Bursa Efek Indonesia.

Penelitian ini menggunakan metode *purposive sampling* dalam menetapkan 4 perusahaan sektor infrastruktur di Indonesia sebagai sampel, yaitu PT XL Axiata Tbk. (EXCL.JK), PT Jasa Marga (Persero) Tbk (JSMR.JK), PT Telkom Indonesia (Persero) Tbk (TLKM.JK), dan PT Smartfren Telecom Tbk (FREN.JK) yang datanya diperoleh dari Yahoo! Finance. *Purposive sampling* adalah metode pengambilan sampel di mana karakteristik sampel dipilih secara sengaja untuk mencapai tujuan penelitian. Semakin banyak kriteria inklusi dan eksklusi yang ditetapkan untuk setiap tujuan

tertentu, semakin spesifik pemilihan sampel tersebut. Keuntungan dari metode ini antara lain memungkinkan penelitian fokus pada kelompok tertentu, membuat sampel lebih seragam, dan menghindari subjek dengan risiko kejadian merugikan [119]. Adapun kriteria sampel pada data ini adalah memiliki periode data minimal 16 tahun dari 1 Januari 2008 hingga 1 Maret 2024, memiliki *range* harga saham yang sejenis, memiliki pola pergerakan saham yang bervariasi, dan tidak ada dalam posisi *delisting, go-private,* maupun suspensi.

3.3.2 Periode Pengambilan Data

Data yang diambil adalah data historis harga saham dari 4 perusahaan sektor infrastruktur di Indonesia dengan periode 16 tahun dari 1 Januari 2008 hingga 1 Maret 2024. Berdasarkan penelitian terdahulu, penelitian [23] menggunakan periode 5 bulan, penelitian [120] menggunakan periode 3 tahun, dan penelitian [31] menggunakan periode 10 tahun. Selain itu, terdapat keterbatasan periode data pada data saham lain yang ingin dianalisis. Adapun kode saham perusahaan infrastruktur yang diambil beserta sumbernya:

Tabel 3. 6 Kode Saham Perusahaan Infrastruktur

Nama Perusahaan	Kode Saham	Sumber
PT XL Axiata Tbk	EXCL.JK	https://finance.yahoo.com/quote/EXCL.JK/history?p=EXCL.JK
PT Jasa Marga (Persero) Tbk	JSMR.JK	https://finance.yahoo.com/quote/JSMR.JK/history?p=JSMR.JK
PT Telkom Indonesia (Persero) Tbk	TLKM.JK	https://finance.yahoo.com/quote/TLKM.JK/history?p=TLKM.JK
PT Smartfren Telecom Tbk	FREN.JK	https://finance.yahoo.com/quote/FREN.JK/history?p=FREN.JK

3.4 Variabel Penelitian

Penelitian ini menggunakan dua tipe variabel yang digunakan pada data penelitian, yaitu variabel dependen dan variabel independen.

3.4.1 Variabel Independen

Variabel independen (variabel bebas) merupakan faktor yang menyebabkan suatu perubahan atau memberikan pengaruh pada suatu keadaan atau nilai [121]. Variabel independen digunakan sebagai prediktor untuk membuat prediksi terhadap variabel dependen. Variabel independen yang digunakan dalam penelitian ini merupakan atribut *close* dari setiap data historis harga saham dari 4 perusahaan sektor infrastruktur di Indonesia, yaitu PT XL Axiata Tbk. (EXCL.JK), PT Jasa Marga (Persero) Tbk (JSMR.JK), PT Telkom Indonesia (Persero) Tbk, dan PT Smartfren Telecom Tbk (FREN.JK).

3.4.2 Variabel Dependen

Variabel dependen (variabel terikat) merupakan hasil dari perubahan variabel lain, menjadi fokus utama penelitian, dan bergantung pada variabel independen. Perubahan dalam variabel independen akan mempengaruhi sejauh mana variabel dependen berubah [121]. Variabel dependen merupakan variabel yang nilainya akan diprediksi. Variabel dependen pada penelitian ini adalah prediksi harga penutupan saham dari 4 perusahaan sektor infrastruktur di Indonesia, yaitu PT XL Axiata Tbk. (EXCL.JK), PT Jasa Marga (Persero) Tbk (JSMR.JK), PT Telkom Indonesia (Persero) Tbk, dan PT Smartfren Telecom Tbk (FREN.JK).

3.5 Teknik Analisis Data

Teknik analisis data yang digunakan pada penelitian ini adalah dengan menerapkan *framework* CRISP-DM dan menggunakan model algoritma *Long Short-Term Memory* (LSTM), *Convolutional Neural Network* (CNN), XGBoost, serta model hibrida CNN-LSTM. Hasil kinerja model akan dievaluasi menggunakan metrik evaluasi RMSE (*Root Mean Square Error*), MSE (*Mean Square Error*), MAE (*Mean Absolute Error*), dan MAPE (*Mean Absolute Percentage Error*). Pada penelitian ini, akan dibandingkan performa dari model CNN, LSTM, *ensemble* XGBoost, dan model hibrida

CNN-LSTM dengan penerapan *hyperparameter tuning* untuk mencari model mana yang paling baik dalam memberikan hasil prediksi harga saham pada sektor infrastruktur. Adapun bahasa pemrograman yang akan digunakan adalah Python dan Jupyter Notebook, serta TensorFlow untuk membentuk model prediksi bersifat *time-series.* Adapun alasan pemilihan bahasa pemrograman Python dibandingkan bahasa pemrograman lainnya yang juga umum digunakan, yaitu R, dibandingkan sebagai berikut: [98]

Tabel 3. 7 Perbandingan Bahasa Pemrograman R dan Python

Keterangan	R	Python
Deskripsi	Bahasa pemrograman *open-source* yang berfokus pada analisis statistik serta visualisasi data	Bahasa pemrograman *object-oriented* yang berfokus pada kemudahan pembacaan kode
Bahasa Dasar	Bahasa pemrograman S	Bahasa pemrograman C
Syntax	Lebih kompleks	Lebih mudah dibaca
Lisensi	*Open source*	*Open source*
Kegunaan	*Statistical purpose*, cocok untuk pengolahan model statistik yang rumit	*General purpose*, dapat disesuaikan dan diterapkan sesuai kebutuhan, seperti pemrograman, analisis data, dan pembuatan situs web.
Kemudahan	Sulit digunakan pemula	Lebih mudah digunakan pemula
Popularitas	Umum digunakan oleh ahli statistika	Umum digunakan oleh *developer* dan *programmer*
Kecepatan & Kompleksitas	Lebih lambat dan kompleks	Lebih sederhana dan cepat
Keunggulan	Lebih mudah dipahami dan digunakan, implementasinya lebih fleksibel, dapat diimplementasikan dalam skala besar	Dapat menyelesaikan permasalahan spesifik statistika, proses *data science* lebih cepat, dapat menuliskan kode dan dokumentasi sekaligus
Kekurangan	Lebih sulit dipahami, terdapat dependensi setiap *package*	Tidak cocok untuk *multithreaded code*, terdapat dependensi *library* dan *package*

Berdasarkan Tabel 3.7, terlihat keunggulan pada bahasa pemrograman Python dibandingkan dengan bahasa pemrograman R. Bahasa pemrograman Python merupakan bahasa pemrograman yang popular dan umum digunakan sehingga dokumentasinya dapat diakses dengan lebih mudah dan leluasa. Selain itu, *syntax*-nya yang cenderung lebih sederhana dan mudah dibaca, penggunaannya yang cocok untuk analisis data, serta proses

pengolahan datanya cepat dan lebih sederhana akan memudahkan proses analisis data pada penelitian ini. Oleh karena itu, pada penelitian ini akan menggunakan *tools* berupa bahasa pemrograman Python dalam proses analisis datanya [98].