BAB 3
METODOLOGI PENELITIAN

Pada bagian ini dijabarkan langkah-langkah yang hendak dilakukan dalam menyusun dan mengerjakan penelitian. Langkah-langkah penelitian yang dijabarkan dimulai dari awal hingga akhir selesai (dokumentasi). Beberapa contohnya yang dapat diikuti antara lain studi literatur, pengumpulan data, prapemrosesan data, pembuatan model, evaluasi dan validasi, dan pelaporan.

3.1 Studi Literatur

Pada bagian ini, dilakukan studi literatur dengan tujuan mengumpulkan wawasan serta referensi dari beberapa penelitian yang membahas topik-topik terkait, seperti *deep learning*, *English Premier League* (EPL), metode *Long Short-Term Memory* (LSTM), dan metrik evaluasi yang digunakan dalam penelitian ini. Melalui proses ini, didapatkan pemahaman yang lebih luas untuk mendukung pengembangan penelitian, khususnya dalam merancang pendekatan yang tepat dalam memprediksi klasemen akhir suatu kompetisi sepak bola.

Selain itu, studi literatur ini bertujuan menyusun perencanaan yang matang terkait kebutuhan dalam pelaksanaan penelitian, termasuk penentuan fitur yang relevan, pendekatan pemodelan, teknik evaluasi, dan validasi hasil. Pemahaman terhadap studi-studi terdahulu membantu dalam mengidentifikasi celah penelitian (*research gap*) dan menyusun kontribusi baru yang bersifat orisinal. Proses studi literatur dilakukan dengan menelaah berbagai sumber terpercaya, seperti jurnal ilmiah, artikel konferensi, laporan teknis, buku, serta karya ilmiah lainnya yang berkaitan langsung dengan fokus penelitian ini.

Tabel 3.1. Ringkasan studi terdahulu prediksi klasemen akhir EPL dan *research gap*

No	Peneliti	Tahun	Periode Data	Model/Metode	Target Prediksi	Metrik Evaluasi	Keterbatasan (Research Gap)
1	Alzidan & Kusuma (JUTIF)	2022	2015–2020	*Long Short-Term Memory*, Distribusi Poisson	Posisi klasemen akhir musim EPL	Akurasi klasemen	Tidak ada tuning hiperparameter; fitur terbatas (tidak ada *elo*, *xG*, atau *form*); cakupan musim terbatas.
2	Wahyono & Novitasari (ResearchGate)	2022	2020–2021	Distribusi Poisson klasik	Posisi akhir untuk pertimbangan investasi	Evaluasi visual, klasemen akhir	Tidak menggunakan *time-series*; tidak mempertimbangkan data urutan atau dinamika performa tim sepanjang musim.

Lanjut pada halaman berikutnya

Tabel 3.1 (lanjutan)

No	Peneliti	Tahun	Periode Data	Model/Metode	Target Prediksi	Metrik Evaluasi	Keterbatasan (Research Gap)
3	Penelitian ini	2025	2000–2025	*LSTM + Hyperband Tuning*, *Embedding*, *BatchNormalization*	Prediksi posisi akhir EPL berdasarkan 12 pertandingan terakhir (`FinalRank`)	MAE, MSE, Spearman, Kendall Tau	Menyempurnakan pendekatan sebelumnya dengan sequence-based modeling, tuning performa optimal, dan dataset historis panjang (25 musim).

Penelitian terkait prediksi klasemen akhir English Premier League (EPL) sebelumnya sebagian besar menggunakan pendekatan statistik seperti distribusi Poisson atau metode regresi tradisional. Contohnya, studi dari Wahyono dan Novitasari (2022) menggunakan metode distribusi Poisson untuk memprediksi posisi akhir tim berdasarkan rata-rata gol dan kebobolan, namun belum mampu memanfaatkan data performa tim dalam bentuk deret waktu (*time-series*) secara menyeluruh. Selain itu, penelitian yang dipublikasikan di JUTIF juga memanfaatkan *Long Short-Term Memory* (LSTM), namun hanya terbatas pada subset data dengan cakupan fitur yang belum optimal serta belum melakukan pencarian hiperparameter secara eksploratif.

Penelitian ini menghadirkan pendekatan yang lebih modern dan menyeluruh dengan mengombinasikan teknik *deep learning* berbasis LSTM dan strategi *hyperparameter tuning* menggunakan algoritma *Hyperband*. Dataset yang digunakan juga lebih lengkap, mencakup 25 musim terakhir (2000–2025), dengan fitur tambahan seperti *elo rating*, *expected goals* (xG), *form*, dan progres musim. Dengan pembuatan data dalam bentuk sekuensial dan pelatihan model berbasis tren performa tim, penelitian ini menjawab kekurangan dari studi sebelumnya yang belum mempertimbangkan dinamika temporal performa tim sepanjang musim. Hal ini menjadikan penelitian ini sebagai kontribusi penting dalam memodelkan performa kompetitif EPL secara *data-driven*.

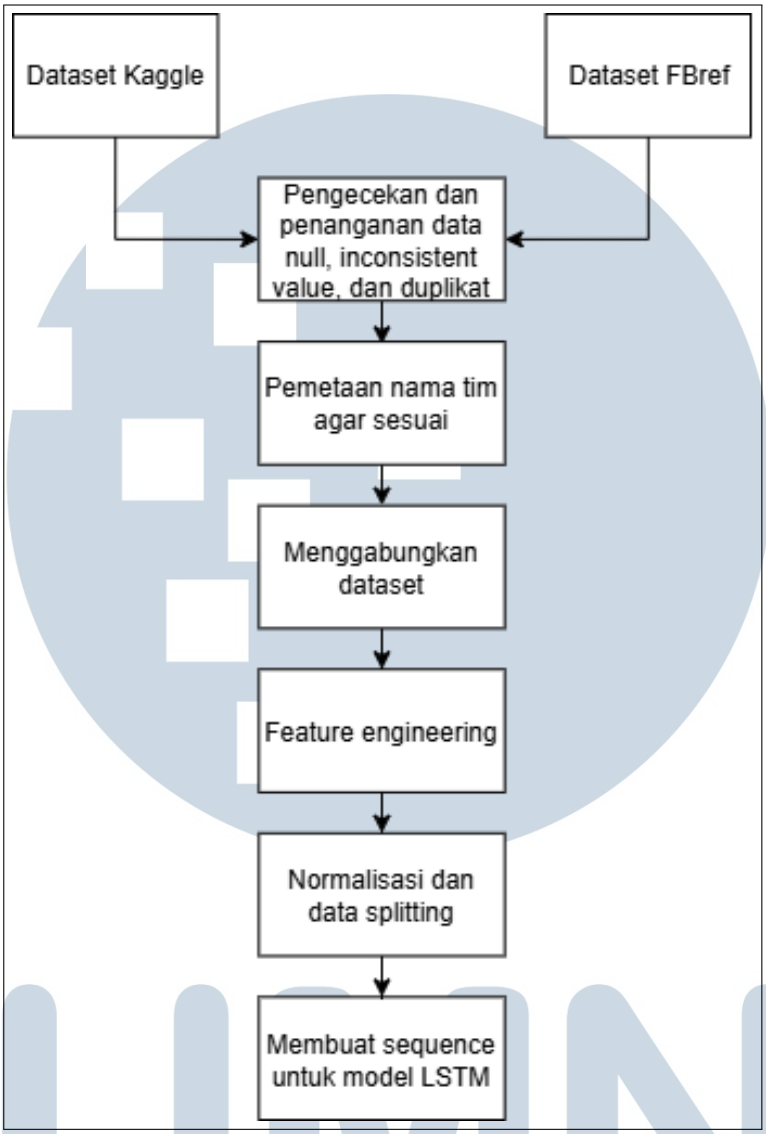
3.2 Pengumpulan Data

Pada tahap pengumpulan data, penelitian ini menjelaskan proses dan sumber data yang digunakan, yang meliputi data sekunder. Data sekunder adalah data yang sudah ada dan disediakan orang lain. Contoh data sekunder dapat berupa data yang diperoleh melalui penelitian orang lain maupun situs penyedia data. Dalam konteks penelitian ini, pengumpulan dataset dilakukan dengan memanfaatkan sumber data sekunder yang tersedia di

website Kaggle (https://www.kaggle.com/code/mahmoudredagamail/club-football-match-data-2000-2025) dan FBref (https://fbref.com/en/comps/9/Premier-League-Stats). Dataset statistik pertandingan EPL 2000/2025 diambil dari website Kaggle yang merupakan sumber data sekunder yang menyediakan informasi lengkap tentang statistik pertandingan tiap *match* dan *elo rating* tiap tim kandang dan tandang. Data tersebut mencakup rentang waktu dari tahun 2000 hingga 2025 beberapa liga sepak bola dunia. Data yang digunakan hanya terkait liga 1 Inggris, kemudian diperoleh total 9300 data yang mencakup periode dua puluh lima tahun terakhir dalam format *home-away*. Selain itu, data xG (*expected goal*) tiap pertandingan diperoleh dari website FBref yang merupakan sumber data sekunder yang menawarkan data historis mengenai peluang gol dari tiap laga, tim, dan musim. Kedua dataset bertujuan digunakan sebagai prediktor, kemudian data dari fitur yang diprediksi diambil melalui situs resmi EPL.

3.3 Prapemrosesan Data

Pada tahap ini, dilakukan prapemorsesan data dengan mencakup beberapa bagian seperti pembersihan data, penggabungan data, seleksi fitur, normalisasi data, pembuatan sequence. Hal ini bertujuan untuk mempersiapkan data yang relevan dengan kebutuhan model nantinya. Detail dari tahapan pemrosesan data, dapat dilihat pada Gambar 3.1.



Gambar 3.1. Kerangka Berpikir Pemrosesan Data

Langkah-langkah perancangan model yang ditunjukkan pada Gambar 3.1 dapat dijelaskan sebagai berikut:

- **Pengecekan dan Penanganan Data:** Tahap awal melibatkan pembersihan data dari masing-masing sumber, yaitu dataset dari Kaggle dan FBref. Proses ini mencakup pengecekan nilai kosong (*null values*), penanganan nilai yang tidak konsisten (*inconsistent values*), serta penghapusan data duplikat. Langkah ini bertujuan memastikan integritas dan kualitas data sebelum proses penggabungan dilakukan.

- **Pemetaan Nama Tim:** Karena kedua dataset berasal dari sumber yang

berbeda, terdapat variasi penamaan klub/tim yang perlu diseragamkan. Proses pemetaan dilakukan agar nama-nama tim memiliki format yang konsisten antar dataset, sehingga dapat digabungkan tanpa konflik atau duplikasi entitas.

- **Penggabungan Dataset:** Setelah data dibersihkan dan nama tim dipetakan, kedua dataset digabung menjadi satu kesatuan data utama. Hasil penggabungan ini akan digunakan sebagai basis untuk tahapan rekayasa fitur dan pemodelan. Penggabungan dilakukan secara horizontal atau vertikal tergantung struktur masing-masing dataset.

- **Rekayasa Fitur (*Feature Engineering*):** Tahap ini melibatkan penciptaan dan transformasi fitur-fitur yang dianggap relevan dalam memprediksi performa tim. Beberapa fitur diturunkan dari statistik pertandingan sebelumnya, seperti rata-rata jumlah tembakan, nilai *expected goals* (xG), jumlah pelanggaran, form tim dalam beberapa laga terakhir, serta *elo rating* sebagai indikator kekuatan relatif tim yang bersifat dinamis. Fitur-fitur ini dikonstruksi secara terpisah untuk tim dan lawannya agar model dapat memahami interaksi kompetitif di setiap pertandingan. Untuk memastikan fitur yang digunakan benar-benar berkontribusi terhadap prediksi, dilakukan analisis *feature importance* menggunakan algoritma *Random Forest*, yang mengevaluasi kontribusi masing-masing fitur terhadap target akhir (`FinalRank`). Hasil dari analisis ini menjadi acuan dalam menyaring fitur yang informatif dan mengeliminasi fitur dengan pengaruh rendah, sehingga struktur data yang dihasilkan lebih optimal dan efisien dalam proses pelatihan model.

- **Normalisasi dan Pembagian Data:** Semua fitur numerik kemudian dinormalisasi agar berada dalam skala yang seragam dan menghindari dominasi satu fitur terhadap yang lain dalam proses pelatihan model. Setelah itu, data dibagi ke dalam dua bagian: data pelatihan dan data uji.

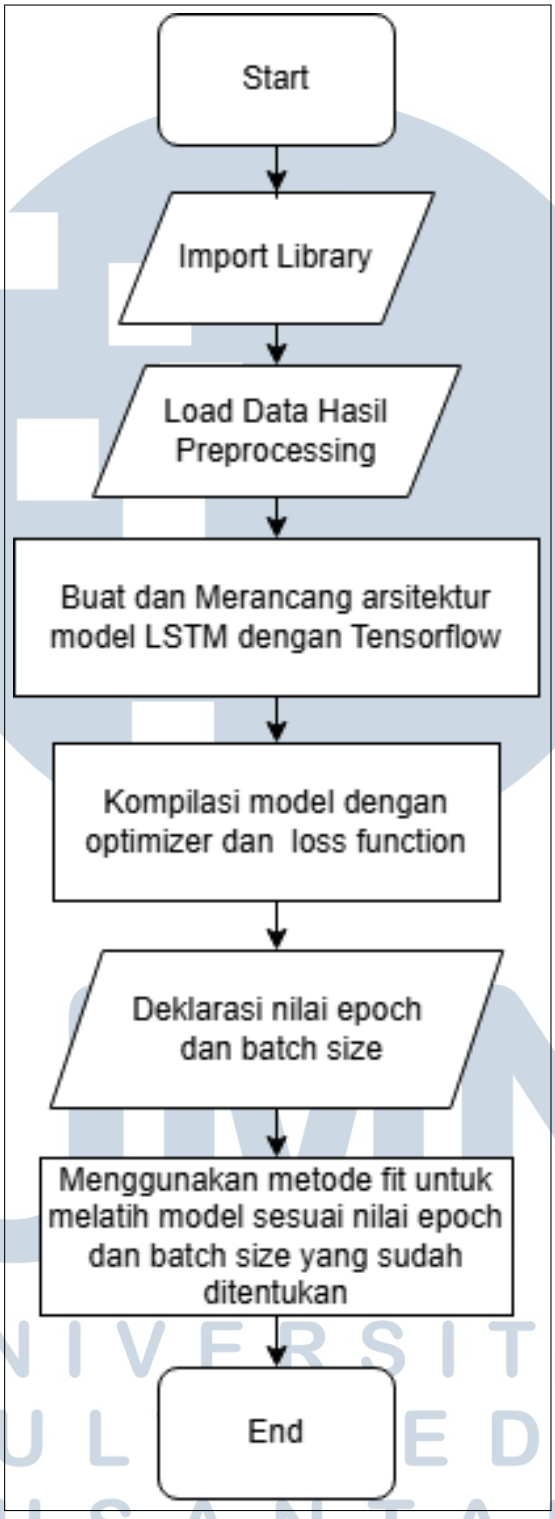
- **Pembuatan *Sequence* untuk Model LSTM:** Dataset yang telah diproses kemudian dikonversi ke dalam format deret waktu atau *sequence* sepanjang 12 *timesteps* untuk setiap tim. Format ini diperlukan agar dapat digunakan sebagai input ke dalam model *Long Short-Term Memory* (LSTM), yang mampu menangkap pola temporal antar pertandingan.

3.4 Perancangan Model

Setelah melakukan preprocessing data, langkah selanjutnya adalah mengembangkan model LSTM. Tujuan dari proses ini adalah untuk mendapatkan model yang mampu memprediksi akhir klasemen EPL dengan baik nantinya. Selanjutnya, model *Long Short-Term Memory* (LSTM) akan dilatih dengan data yang sudah diproses. Proses ini dibagi dalam dua bagian: pertama adalah latihan model dengan dataset latih, kedua model yang sudah dilatih akan dilakukan *hyperparameter tuning* untuk mencari performa model terbaik, berikut adalah bagian dari pencangan model :

3.4.1 *Training Model*

Pada fase pelatihan, langkah-langkah yang perlu dilakukan meliputi: import library yang diperlukan (tensorflow, pandas, dan keras), mempersiapkan data yang sudah diproses sebelumnya, membangun model LSTM, kompilasi model (menentukan optimizer dan fungsi loss), dan mulai melatih model sesuai nilai *epoch* dan *batch size* yang sudah dideklarasikan. Untuk lebih jelasnya, rangkaian alur pelatihan model akan dijelaskan dengan Gambar 3.2.



Gambar 3.2. Alur Pelatihan Model

3.4.2 *Hyperparameter Tuning*

Pada bagian ini, akan dilakukan pengujian untuk menentukan parameter terbaik untuk model. Parameter yang disertakan untuk dilakukan *tuning* menyangkut dengan jumlah unit dalam arsitektur model, jenis optimizer, jumlah dimensi embedding, dan *learning rate*. Dengan banyaknya kombinasi yang perlu dicoba, maka digunakanlah metode *hyperband* agar mempermudah dan mempercepat proses pencarian parameter terbaik nantinya. Model terbaik tentu saja akan digunakan untuk uji data tes, kemudian dibandingkan dengan hasil dari model sebelum *tuning* untuk melihat peningkatan performa secara konkret pada tahap evaluasi nantinya.

3.5 Evaluasi Model

Pada tahap ini, setelah berhasil mendapatkan model terbaik lewat metode *hyperparameter tuning*, data tes diuji coba secara langsung untuk mendapatkan hasil prediksi paling baik. Metode yang digunakan dalam evaluasi model adalah, MAE (*mean absolute error*), MSE (*mean squared error*), korelasi Spearman, dan *Kendall Tau*. MAE adalah rata-rata dari selisih absolut antara nilai prediksi model dan nilai aktual, yang memberikan gambaran langsung tentang seberapa jauh rata-rata kesalahan prediksi model dalam satuan data asli. MSE adalah rata-rata kuadrat dari selisih antara nilai prediksi dan nilai aktual, yang memberikan bobot lebih besar pada kesalahan prediksi yang besar dan memiliki satuan yang sama dengan variabel target. Korelasi Spearman adalah ukuran non-parametrik yang menilai kekuatan dan arah hubungan monotonik (tidak harus linear) antara dua variabel dengan menghitung korelasi antara peringkat data, bukan nilai mentahnya. Kendall Tau adalah metrik korelasi non-parametrik lain yang mengukur kekuatan hubungan ordinal antara dua variabel dengan membandingkan jumlah pasangan data yang *concordant* (memiliki urutan yang sama) dan *discordant* (memiliki urutan yang berlawanan). sg