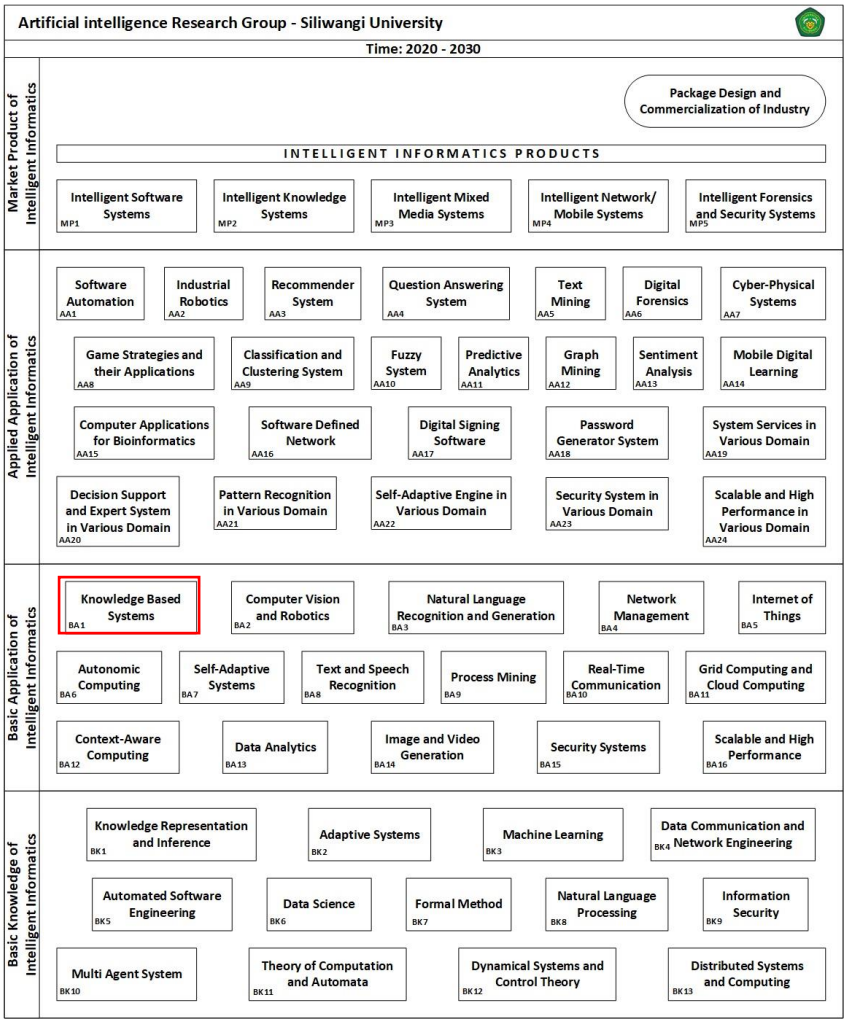
BAB III

METODOLOGI PENELITIAN

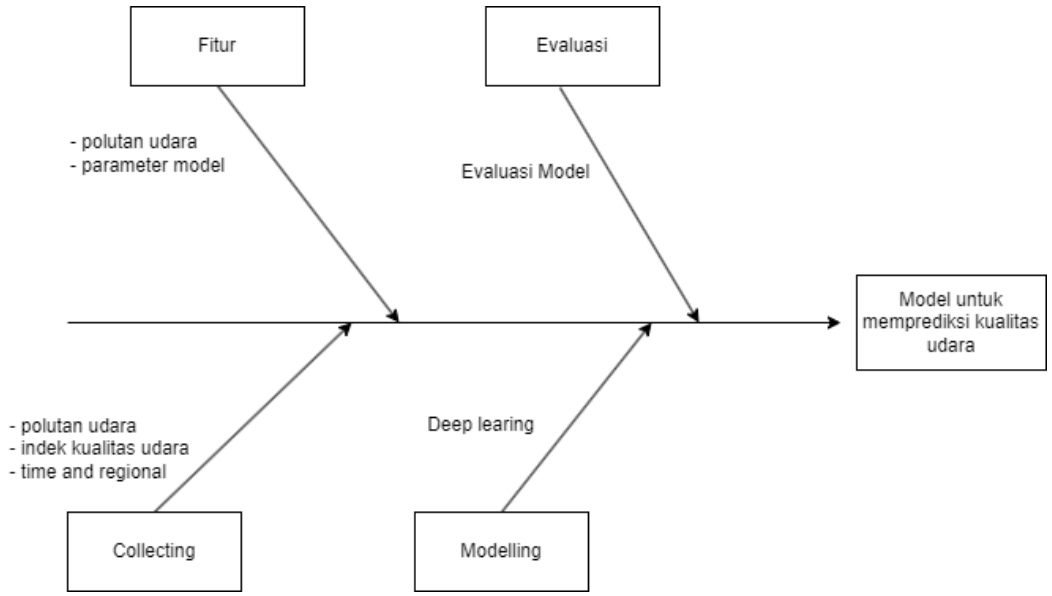
3.1 *Roadmap* Penelitian

Secara keseluruhan, rencana penelitian ini sejalan dengan *roadmap* di Universitas Siliwangi pada bagian *artificial Intelligence*. Topik penelitian yang dipilih yaitu *knowledge based system* pada ranah *basic application of intelligent informatics*. *Roadmap* penelitian dapat dilihat pada gambar 3.1.



Gambar 3. 1 *Roadmap* Penelitian (AIS Universitas Siliwangi, 2020)

Gambar 3.1, pemilihan topik *knowledge based system* pada ranah *basic applications of intelligent informatics* sehingga penelitian ini memiliki tujuan dalam membuat pemodelan yang akan menjadi dasar pengetahuan. Penelitian ini, melakukan pemodelan yang berorientasi dalam memprediksi kualitas udara untuk menjadi pengetahuan sebuah aplikasi. Berikut ini gambar 3.2, merupakan *fishbone* penelitian sebagai bagian dari *roadmap.*



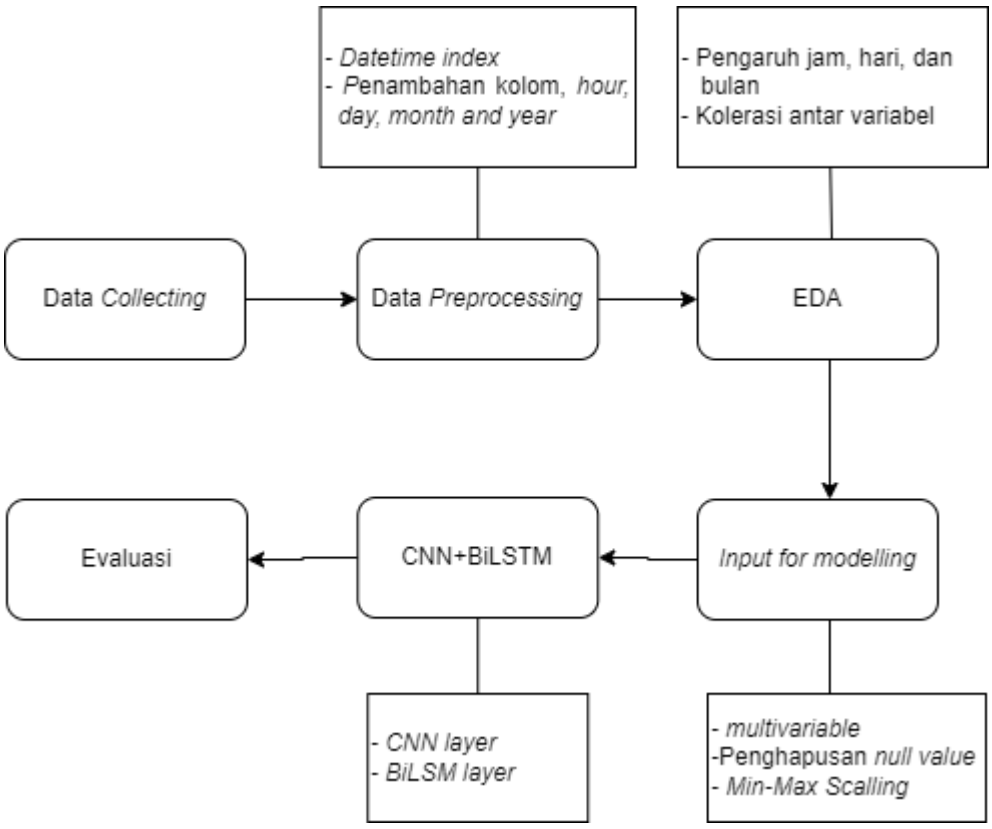
Gambar 3. 2 *Fishbone* Penelitian

Gambar 3.2, *fishbone* mempresetasikan *knowledge based system* penelitian. penelitian dirancang untuk pemodelan yang dapat memprediksi kualitas udara beserta komponennya. Pengambilan data yang digunakan memiliki informasi mengenai polusi udara beserta waktu dan wilayah. Kemudian, pemilihan fitur dari polutan udara serta pemilihan parameter untuk model dibutuhkan dalam pemodelan. Selanjutnya, hasil dari pemodelan menggunakan

deep learning dievaluasi untuk mengetahui peforma model dalam memprediksi kualitas udara.

3.2 Metode Penelitian

Metode penelitian untuk memprediksi indeks kualitas udara beserta komponenya dengan melibatkan integrasi *convolutinal* layer pada BiLSTM. Metode ini melibatkan beberapa langkah seperti yang ditunjukkan dalam gambar 3.3, menunjukan penerapan *convolutional* pada LSTM sebagai strategi dalam meningkatkan akurasi prediksi kualitas udara.



Gambar 3. 3 Tahapan Penelitian

Gambar 3.3 mengilustrasikan beberapa tahapan utama dalam proses penelitian. Pembuatan model untuk memprediksi kualitas udara membutuhkan *dataset* berisi indeks kualitas udara beserta komponennya sehingga memerlukan proses data *collection* untuk memproses pengambilan data yang akan digunakan di dalam penelitian (Samal et al., 2021a). Setelah itu, data *preprocessing* untuk memproses data disesuaikan untuk lebih mudah dianalisis dan disesuaikan dengan kebutuhan (Arkadia et al., 2022). Kemudian, *Exploratory Data Analysis* menganalisis data hasil dari data *preprocessing* untuk lebih memahami *dataset* yang digunakan (S. Li et al., 2020). Selanjutnya, *input for modeling* merupakan tahapan memproses data disesuaikan dengan kebutuhan model (Khan et al., 2022). Setelah itu, integrasi CNN dengan BiLSTM sebagai arsitektur model dalam memprediksi kualitas udara beserta komponennya. Tahapan yang terakhir yaitu, evaluasi mengukur performa model dalam memprediksi kualitas udara (D. Li et al., 2022). Penjelasan detail dari setiap tahapannya yaitu sebagai berikut:

1. *Data Collecting*

   Proses ini mencakup pengambilan *dataset* yang diperlukan untuk penelitian. Data akan diambil dari *platform* weatherbit.io, fokus pada data historis waktu dari indeks kualitas udara beserta komponennya di setiap ibu kota provinsi di Indonesia.

2. *Data Preprocessing*

   Tahap data *preprocessing* adalah langkah penting dalam persiapan data untuk analisis lebih lanjut (Chauhan et al., 2021). Selanjutnya, untuk kebutuhan pemodelan *time series*, fitur *datetime* diubah menjadi indeks dari dataset, dan

fitur tambahan seperti *hour, day, month,* dan *year* ditambahkan untuk memudahkan *exploratory data analysis*. Setelah itu, dataset dengan mempertahankan data dari satu ibu kota untuk digunakan datanya.

3. *Exploratory Data Analysis* (EDA)

Tahap ini, fokus analisis berada pada distribusi data untuk mengetahui bagaimana jam, hari, dan bulan mempengaruhi kualitas udara (Khan et al., 2022). Selain itu, dilakukan analisis korelasi antara variabel tersebut, di mana nilai korelasi tinggi akan dipertimbangkan sebagai fitur penting dalam proses pemodelan. Berikut persamaan (7) untuk menghitung korelasi antar variabel (Tyas et al., 2023).

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{(\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2})} \quad (7)$$

Dalam persamaan (7) ini, X_i dan Y_i mewakili nilai individu dari masing-masing variabel, sedangkan $\bar{X}$ dan $\bar{Y}$ adalah rata-rata dari masing-masing variabel. Melalui langkah-langkah perhitungan yang mencakup pengurangan nilai individu dari rata-rata, perkalian selisih, dan normalisasi dengan faktor akar kuadrat, kita dapat menghasilkan koefisien korelasi Pearson (r). Rentang nilai r berada antara -1 hingga 1, dengan -1 menunjukkan korelasi negatif sempurna, 1 menunjukkan korelasi positif sempurna, dan 0 menunjukkan tidak adanya korelasi linier antara kedua variabel. Dengan demikian, rumus ini menyediakan ukuran objektif terhadap sejauh mana variabilitas antara dua variabel dapat dijelaskan dalam konteks hubungan linier.

4. *Input for Modelling*

Setelah melalui proses *preprocessing*, langkah berikutnya adalah mengubah membuat rangkaian data *multivariate* menjadi sampel yang dapat digunakan dalam pemodelan terhadap data deret waktu. Dalam proses ini, langkah-langkah termasuk menghapus data yang kosong dan fitur yang tidak diperlukan. Setelah itu, dilakukan penskalaan data menggunakan *Min-Max scaling* untuk mendapatkan rentang nilai antara 1 dan 0 (D. Li et al., 2022). Selanjutnya, data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian (Arkadia et al., 2022). Berikut persamaan (8) untuk menghitung *Min-Max scaling* (Tang et al., 2023).

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

Permasaan (8) *Min-Max scaling* digunakan untuk menormalkan nilai-nilai dalam suatu dataset ke dalam rentang tertentu, umumnya [0, 1]. Setiap nilai pada suatu fitur (X) dalam dataset dikurangkan dengan nilai minimum $(X_{\min})$ dari fitur tersebut, lalu hasilnya dibagi dengan selisih antara nilai maksimum X_{max} dan nilai minimum. Ini menghasilkan nilai yang terukur dan terstandarisasi dalam rentang yang diinginkan. Penerapan rumus ini pada setiap fitur dalam dataset membantu mencegah dominasi fitur tertentu dan memastikan konsistensi skala, sehingga meningkatkan kinerja model machine learning yang sensitif terhadap skala input.

5. Integrasi BiLSTM dengan CNN

Integrasi antara *Bidirectional Long Short-Term Memory* (BiLSTM) dan *Convolutional Neural Network* (CNN) pada model *neural network*

dikembangkan secara spesifik untuk memproyeksikan kualitas udara dalam data *time series*. CNN berfungsi untuk mengekstrak fitur spasial dari data waktu, seperti pola distribusi polutan udara, sedangkan BiLSTM memiliki peran utama dalam memahami dan memodelkan hubungan sekuensial yang terdapat dalam perubahan kualitas udara sepanjang waktu (D. Li et al., 2022). Dengan memadukan keunggulan CNN dalam menangkap pola spasial dan kemampuan BiLSTM dalam mengelola data sekuensial. CNN menggunakan 64 filter untuk menangkap pola yang berbeda dari data input (Samal et al., 2021b), menggunakan 3 *kernel size* untuk ekstraksi 3 data secara berturut turut (Choi et al., 2021) dan *pool size* 2 untuk mengubah ukuran dimensi (Samal et al., 2021a). Layer BiLSTM menggunakan 200 unit untuk setiap layer LSTM sehingga untuk menjadi BiLSTM membutuhkan 400 *unit* (Samee et al., 2022). Dalam proses *training* model menggunakan optimasi Adam (Seng et al., 2021) dengan 50 *epochs* (Maryam Zare et al., 2022).

6. Evaluasi

Hasil dari pemodelan akan diukur performanya menggunakan *Mean Absolute Percentage* Error (MAPE). MAPE adalah sebuah metrik evaluasi kinerja model yang umumnya digunakan dalam bidang statistika atau *deep learning* (H. Liu et al., 2021). Berikut persamaan (9) proses dari perhitungan MAPE (Tang et al., 2023).

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

Pada persamaan (9), n menyatakan total observasi atau data yang dievaluasi. Simbol y_i mengacu pada nilai aktual dari observasi ke-i, sedangkan

$\hat{y}_i$ adalah nilai prediksi yang dihasilkan oleh model untuk observasi ke-i. Setiap observasi, selisih persentase antara nilai aktual dan prediksi dibagi oleh nilai aktual, kemudian diambil nilai absolutnya. Selanjutnya, semua nilai absolut persentase ini dijumlahkan, hasilnya dikalikan dengan 100, dan dibagi dengan total jumlah observasi n. MAPE mengukur seberapa besar rata-rata persentase kesalahan prediksi model dari nilai aktual, dan nilai MAPE yang lebih rendah menunjukkan kinerja model yang lebih baik. Validitas MAPE dapat dinilai berdasarkan konteks dan sifat data yang dievaluasi.