

BAB I

PENDAHULUAN

Bab pendahuluan ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan yang digunakan dalam dokumen skripsi mengenai Prediksi Struktur Sekunder Protein Menggunakan *Bidirectional Long Short-Term Memory* (Bi-LSTM).

1.1 Latar Belakang

Protein adalah molekul biologis penting yang terlibat dalam hampir semua proses biologis di dalam sel, seperti katalisis reaksi kimia, pengangkutan molekul, dan sinyal seluler. Struktur protein memainkan peran sentral dalam menentukan fungsinya, dan pemahaman mengenai struktur protein sangat penting untuk berbagai aplikasi biomedis, termasuk pengembangan obat, desain vaksin, dan terapi penyakit (Wang dkk., 2016). Secara umum, struktur protein terdiri dari empat tingkatan, yaitu struktur primer yang merupakan urutan asam amino, struktur sekunder yang terbentuk dari pelipatan lokal menjadi *helix alfa*, lembaran *beta*, dan *loop*. Lalu terdapat struktur tersier yang merupakan konfigurasi tiga dimensi kompleks, serta struktur kuarterner yang terbentuk dari gabungan beberapa subunit polipeptida (Y. Zhao & Liu, 2021).

Struktur sekunder protein biasanya diklasifikasikan ke dalam tiga kelas, yaitu *helix*, lembaran *beta*, dan *coil* atau *loop*, yang dikenal dengan istilah skema Q3. Namun, pengklasifikasian yang lebih detail sering kali dilakukan menggunakan 8 kelas, yang dikenal dengan istilah skema Q8, yang mencakup *helix alfa* (H), *helix 3_10* (G), *helix pi* (I), *strand beta* (E), *beta-bridge* (B), *loop* (L), *turn* (T), dan *bend* (S) (Spencer dkk., 2015). Dengan klasifikasi skema 8 kelas, model dapat menangkap detail lebih halus dari lipatan lokal dalam struktur sekunder protein, sehingga meningkatkan akurasi prediksi dalam aplikasi yang lebih spesifik, seperti analisis interaksi molekul atau desain obat berbasis protein. Di samping itu, klasifikasi skema Q8 juga memerlukan kompleksitas yang lebih tinggi dibandingkan skema Q3 karena skema Q8 memerlukan pemahaman yang lebih mendalam terhadap pola-pola spesifik dalam data sekuens protein.

Dalam bioinformatika, prediksi struktur sekunder protein merupakan masalah yang penting dan kompleks. Menentukan struktur tiga dimensi protein dari urutan primernya sangat sulit, sehingga prediksi struktur sekunder menjadi perhatian utama bagi banyak peneliti (Yang dkk., 2018). Biasanya struktur protein ditentukan menggunakan beberapa teknik eskperimental, seperti kristalografi sinar-X, spektroskopi resonansi magnetik nuklir (NMR), dan metode komputasional. Tetapi, metode kristalografi sinar-X dan spektroskopi resonansi magnetik nuklir (NMR) memerlukan biaya dan waktu yang besar, sehingga mengakibatkan lambatnya kemajuan dalam penelitian protein karena kurangnya efisiensi untuk analisis skala besar dan keterbatasan akses untuk penelitian baru. Oleh karena itu, metode komputasi untuk prediksi struktur protein, khususnya dengan pendekatan *machine learning*, semakin diminati dalam beberapa tahun terakhir.

Berbagai pendekatan *machine learning* telah diterapkan untuk prediksi struktur sekunder protein, termasuk *transfer learning* menggunakan *embedding* berbasis ELMo yang diterapkan oleh Heinzinger dkk. (2019), *Convolutional Neural Network* (CNN) dan *Support Vector Machine* (SVM) oleh Sutanto dkk. (2021), serta gabungan *Bidirectional Temporal Convolutional Networks* (BTCN) dan *Bidirectional Long Short-Term Memory* (Bi-LSTM) oleh Yuan dkk. (2023). Pendekatan *transfer learning* menggunakan *embedding* SeqVec oleh Heinzinger dkk. (2019) menunjukkan efektivitasnya dalam menangkap properti biofisik protein melalui representasi *embedding* berbasis data besar tanpa label. Metode ini menggunakan *embedding* SeqVec sebagai fitur utama dalam jaringan saraf sederhana, namun memiliki kekurangan dalam memprediksi hubungan spasial panjang. Di sisi lain, metode CNN-SVM bekerja dengan cara menggunakan CNN 1-dimensi untuk mengekstrak fitur dan SVM untuk klasifikasi data (Sutanto dkk., 2021). Meskipun pendekatan ini dapat menghasilkan akurasi yang cukup baik, metode ini juga mengalami keterbatasan dalam menangkap hubungan spasial jangka panjang. Selain itu, pendekatan BTCN dan Bi-LSTM yang diterapkan oleh Yuan dkk. (2023) menerapkan pendekatan *ensemble deep learning* dengan menggabungkan BTCN untuk interaksi lokal dua arah dan fitur jarak jauh, dan Bi-LSTM untuk menangkap ketergantungan global antar-residu. Model ini mampu mendapatkan hasil prediksi dengan nilai yang tinggi namun kompleksitas tinggi model dapat menjadi hambatan untuk penerapan yang lebih luas, seperti tantangan dalam efisiensi pemrosesan data dalam jumlah besar dan keterbatasan dalam infrastruktur atau sumber daya komputasi pada lingkungan industri.

Untuk mengatasi kekurangan dari metode sebelumnya, penelitian ini memilih menggunakan metode *Bidirectional Long Short-Term Memory* (Bi-LSTM) murni. Bi-LSTM memungkinkan model untuk memproses data sekuensial dari kedua arah, sehingga bisa menangkap lebih banyak konteks sekuensial dari urutan asam amino. Ini memungkinkan model untuk menghasilkan prediksi yang lebih akurat dalam konteks prediksi struktur sekunder protein dibandingkan dengan metode yang sebelumnya telah digunakan tanpa harus membutuhkan komputasi yang sangat tinggi. Bi-LSTM juga efisien menangani kompleksitas urutan protein dengan lebih baik karena informasi dari kedua arah (*forward* dan *backward*) dapat digunakan untuk memahami urutan dengan lebih komprehensif (Xudkk., 2019).

Untuk mencapai model Bi-LSTM yang optimal dalam prediksi struktur sekunder protein, penelitian ini juga memfokuskan pada optimasi *hyperparameter* seperti *optimizer*, *learning rate*, *dropout rate*, dan jumlah lapisan LSTM untuk mendapatkan konfigurasi model yang paling optimal. Dengan optimasi *hyperparameter* yang tepat, diharapkan model dapat mencapai kinerja prediksi yang lebih baik, terutama dalam hal akurasi dan kemampuan generalisasi pada dataset baru, yang belum pernah digunakan dalam pelatihan dan validasi sebelumnya.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah diuraikan, maka permasalahan yang dapat dirumuskan adalah sebagai berikut:

1. Bagaimana membangun model untuk prediksi struktur sekunder protein pada skema Q8 menggunakan metode *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan mendapatkan kombinasi *hyperparameter* terbaik untuk membangun model tersebut?
2. Bagaimana performa metode Bi-LSTM dalam memprediksi struktur sekunder protein?

1.3 Tujuan dan Manfaat

Dalam penelitian ini terdapat beberapa tujuan yang hendak dicapai. Adapun tujuan yang dimaksud antara lain:

1. Mengembangkan model dan mendapatkan kombinasi *hyperparameter* untuk prediksi struktur sekunder protein pada skema Q8 menggunakan metode *Bidirectional Long Short-Term Memory* (Bi-LSTM)

2. Menguji dan menganalisis performa metode Bi-LSTM dalam memprediksi struktur sekunder protein.

Adapun manfaat yang diharapkan dapat diperoleh dari penelitian ini adalah:

1. Memberikan panduan untuk peneliti lain dalam hal pengembangan model *deep learning* dan optimasi *hyperparameter* pada model, khususnya dalam pemrosesan data sekuensial.
2. Memperoleh pemahaman dan dapat menyediakan contoh konkret penerapan Bi-LSTM dalam prediksi struktur sekunder protein, yang dapat diadopsi dan dikembangkan lebih lanjut untuk topik yang serupa atau berbagai jenis data sekuensial lainnya

1.4 Ruang Lingkup

Ruang lingkup penelitian ini mencakup beberapa aspek utama yang menjadi batasan dalam pelaksanaan penelitian. Hal ini dilakukan untuk memfokuskan penelitian pada aspek-aspek yang relevan dengan tujuan yang ingin dicapai. Adapun ruang lingkup penelitian ini adalah sebagai berikut:

1. Dataset yang digunakan dalam penelitian ini adalah dataset CullPDB untuk proses pelatihan dan validasi, serta dataset CB513 untuk pengujian model. Kedua dataset tersebut berisi urutan asam amino dari berbagai jenis protein beserta label struktur sekundernya yang dikategorikan dalam skema Q8 dalam format npy.gz.
2. Data dalam dataset sudah mencakup informasi *Position-Specific Scoring Matrix* (PSSM), sehingga penelitian ini tidak melakukan komputasi tambahan untuk menghitung PSSM.
3. Optimasi *hyperparameter* meliputi pengujian berbagai nilai untuk *optimizer*, *learning rate*, *dropout rate*, dan jumlah lapisan LSTM yang digunakan dalam model.
4. Model terbaik yang diperoleh dari hasil akurasi validasi tertinggi langsung diuji menggunakan dataset CB513 (data uji) tanpa dilakukan pelatihan ulang.
5. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, serta analisis lebih rinci melalui *confusion matrix* dan metrik evaluasi lain seperti *precision*, *recall*, dan *F1-score*, untuk mengukur performa prediksi terhadap masing-masing kelas struktur sekunder protein dalam skema Q8.
6. Penelitian ini tidak mencakup prediksi struktur tersier atau kuarterner protein. Fokus utama adalah pada prediksi struktur sekunder protein dengan skema Q8.

1.5 Sistematika Penulisan

Sistematika penulisan digunakan untuk memberi gambaran dan alur yang runtut terkait penyusunan laporan skripsi yang berjudul *Prediksi Struktur Sekunder Protein Menggunakan Bidirectional Long Short-Term Memory (Bi-LSTM)* sehingga mudah dipahami dan dipelajari. Sistematika penulisan yang dimaksud adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup penelitian, serta sistematika penulisan dari penelitian mengenai prediksi struktur sekunder protein menggunakan *Bidirectional Long Short-Term Memory (Bi-LSTM)*.

BAB II LANDASAN TEORI

Bab ini membahas tentang tinjauan pustaka dan teori-teori yang digunakan untuk landasan penelitian mengenai prediksi struktur sekunder protein menggunakan *Bidirectional Long Short-Term Memory (Bi-LSTM)*, yang meliputi protein, struktur protein, prediksi struktur sekunder protein, PSSM, *deep learning*, dan alur serta penggunaan model yang digunakan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini membahas alur dan langkah-langkah yang digunakan sebagai acuan dalam pelaksanaan penelitian. Gambaran umum dari tahapan penelitian ini adalah pengumpulan data, prapemrosesan data, pembagian data, pembangunan dan pelatihan model, serta evaluasi model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas mengenai lingkungan dan perangkat untuk penelitian, serta menjelaskan hasil yang diperoleh dari pengujian dan analisis terkait hasil yang didapatkan untuk mendapatkan kombinasi *hyperparameter* terbaik.

BAB V PENUTUP

Bab ini membahas mengenai kesimpulan dari analisis hasil yang telah dilakukan pada bab-bab sebelumnya serta saran untuk penelitian selanjutnya.