

**PEMODELAN INTERAKSI PROTEIN-PROTEIN DARI TANAMAN
OBAT PADA PENYAKIT KANKER MENGGUNAKAN METODE
RECURRENT NEURAL NETWORK**

MHD. Hidayat¹, Kana Saputra S²

¹²Universitas Negeri Medan

E-mail: mhdhidayat@mhs.unimed.ac.id¹, kanasaputras@unimed.ac.id²

Abstrak

Salah satu penyakit yang memiliki tingkat kematian tinggi adalah kanker, dan jumlah penderitanya terus bertambah setiap tahun. Pendekatan eksperimental untuk mempelajari interaksi protein-protein membutuhkan banyak sumber daya seperti waktu, tenaga, dan uang. Penggunaan obat herbal dalam jangka waktu yang panjang menimbulkan dampak negatif yang lebih sedikit jika dibandingkan dengan penggunaan obat konvensional. Pemodelan dengan metode Recurrent Neural Network digunakan untuk mengklasifikasikan interaksi protein-protein yang berasosiasi pada penyakit kanker. Tahapan dalam melakukan pemodelan pada penelitian ini adalah pengumpulan data, pra-proses data, pembangunan model, evaluasi model, prediksi interaksi protein-protein, dan analisis. Hasil penelitian menunjukkan bahwa pemodelan Recurrent Neural Network memiliki tingkat akurasi dengan arsitektur LSTM adalah sebesar 93% serta dengan 0.05% tingkat Mean Square Error sedangkan dengan arsitektur Bidirectional LSTM adalah sebesar 97% serta dengan 0.01% tingkat Mean Square Error. Protein yang diidentifikasi sebagai protein yang penting (signifikan) adalah protein NTRK2 dan VEGFA. Tanaman obat yang cocok untuk dijadikan kandidat obat alternatif adalah *Codonopsis pilosula*., *Rehmannia glutinosa*, dan *Lycium chinense*.

Kata Kunci — Recurrent Neural Network, Protein, Kanker.

PENDAHULUAN

Penyakit kanker, khususnya kanker paru-paru, merupakan salah satu tantangan kesehatan global yang terus meningkat. Data menunjukkan bahwa setiap tahun, sekitar 1,2 hingga 1,56 juta orang meninggal akibat kanker paru-paru, dengan proyeksi peningkatan yang signifikan pada tahun 2030 [1]. Kanker, sebagai kelompok penyakit dengan pertumbuhan sel yang tidak terkendali, menyebar dan merusak jaringan normal di tubuh, mengakibatkan tingkat kematian yang tinggi. Meskipun pendekatan konvensional seperti pembedahan, terapi radiasi, dan kemoterapi telah menjadi pilihan utama, masih diperlukan solusi tambahan seperti obat herbal [2].

Studi interaksi protein-protein menjadi kunci dalam memahami dasar biologis kanker dan mengembangkan metode pengobatan yang lebih efektif. Pendekatan eksperimental untuk memahami interaksi ini membutuhkan sumber daya yang besar. Oleh karena itu, pendekatan komputasi, terutama metode deep learning, menjadi alternatif yang menjanjikan. Deep learning, sebagai bagian dari machine learning, mampu memproses informasi kompleks dan telah berhasil diterapkan dalam bioinformatika untuk memprediksi interaksi protein-protein [3].

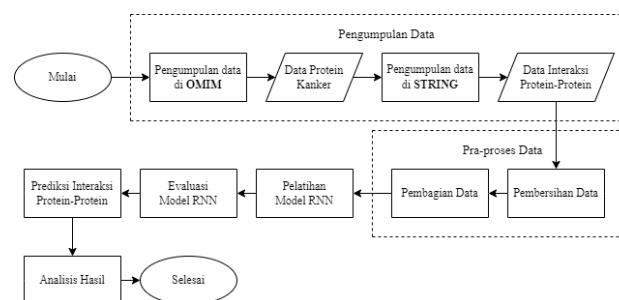
Metode deep learning, seperti Recurrent Neural Network (RNN) dengan lapisan yang dalam, menjadi fokus penelitian untuk memprediksi interaksi protein-protein. Studi sebelumnya telah menunjukkan keberhasilan penggunaan metode ini dalam mengolah data sekuensial, termasuk dalam memprediksi interaksi protein-protein [3]. Dengan memanfaatkan data dari pangkalan data OMIM dan STRING, penelitian ini bertujuan untuk mengembangkan model yang efektif menggunakan RNN, khususnya Long ShortTerm Memory (LSTM), dan mengukur kinerjanya dengan Mean Square Error.

Obat herbal telah menjadi pilihan populer selama tiga dekade terakhir, dengan lebih dari 80% populasi global mengandalkan penggunaannya [4]. Dalam konteks pengobatan kanker, analisis interaksi obat-target menjadi krusial. Namun, studi eksperimental interaksi ini membutuhkan usaha besar. Oleh karena itu, penggunaan pendekatan machine learning untuk memprediksi interaksi obat-target menjadi solusi yang menarik [5].

Penelitian ini bertujuan untuk mengembangkan model prediksi interaksi obat-target menggunakan pendekatan machine learning, terutama RNN. Dengan menggabungkan data interaksi protein-protein dari pangkalan data OMIM dan STRING, penelitian ini berpotensi memberikan pandangan yang lebih dalam tentang protein-protein yang signifikan dalam konteks kanker. Hasilnya diharapkan dapat memberikan landasan untuk penemuan obat-obatan baru, terutama obat herbal, dengan dampak positif pada pengobatan kanker.

METODE PENELITIAN

Penelitian ini berfokus pada identifikasi (in silico) protein potensial tanaman obat sebagai komposisi utama kandidat obat penyakit kanker. Dengan menerapkan analisis in silico dapat dilihat seberapa signifikan pengaruh protein tanaman obat tersebut terhadap protein penyakit kanker. Tahapan yang dilakukan penelitian ini terdiri atas beberapa tahapan utama yaitu, pengumpulan data, pra-proses data, pembangunan model, evaluasi model, prediksi interaksi protein-protein, dan analisis hasil. Tahapan-tahapan dalam metode tersebut dapat dilihat dalam bagan berikut.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Dalam penelitian ini, data yang digunakan berasal dari sumber terbuka yang dapat diakses melalui internet. Data protein-protein terkait penyakit kanker diperoleh dari Online Mendelian Inheritance in Man (OMIM), sebuah pangkalan data yang fokus pada gen manusia dan penyakit

genetik. OMIM menyediakan informasi terkait gen, penyakit, gejala, penanganan, dan sifat genetiknya. Sementara itu, data interaksi protein-protein terkait kanker diperoleh dari pangkalan data STRING, yang merupakan sumber daya web untuk Protein-Protein Interaction (PPI) yang mencakup interaksi fisik dan fungsional. Proses pengumpulan data dilakukan menggunakan program berbahasa Python untuk memastikan efisiensi dan akurasi dalam analisis data.

Praproses Data

Praproses data dalam penelitian ini melibatkan langkah-langkah penting seperti membersihkan data dari informasi yang tidak relevan atau tidak valid, mengidentifikasi dan menghapus duplikat, menangani kesalahan penulisan atau format yang salah, membuang urutan sequence protein yang tidak sesuai kriteria, dan menyesuaikan format data agar mudah dikelola. Data yang telah dibersihkan kemudian dibagi menjadi dua bagian: satu untuk pelatihan dan evaluasi model, dan satu lagi untuk proses prediksi. Proses pembagian data dilakukan secara acak untuk memastikan representativitas contoh. Hasil dari praproses ini adalah data protein yang telah dimodifikasi dan disimpan dalam format berkas comma separated value (csv).

Pembangunan Model

Pada tahap perancangan model, digunakan arsitektur Recurrent Neural Network (RNN) dengan fokus pada penggunaan Long Short-Term Memory (LSTM). Proses pelatihan model melibatkan pemberian input ke model, perhitungan output yang diprediksi, pengukuran loss (selisih antara prediksi dan nilai sebenarnya), dan pengoptimalan parameter model dengan algoritma tertentu. Tahap evaluasi menggunakan data validasi berbeda dari data pelatihan untuk mengukur kemampuan generalisasi dan mendeteksi potensi overfitting atau underfitting. Apabila diperlukan, penyetelan ulang parameter model atau penerapan teknik seperti regularisasi dapat dilakukan untuk meningkatkan kinerja model [6].

Evaluasi Model

Model yang sudah dibangun nanti akan dievaluasi untuk menganalisis akurasi yang diperoleh, evaluasi juga dapat menggunakan Mean Squared Error (MSE). Semakin rendah nilai MSE, maka semakin tinggi kualitas model tersebut. sebagai berikut [7].

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Analisis Interaksi Protein-Protein

Analisis interaksi protein-protein dilakukan untuk mengidentifikasi protein-protein yang memiliki interaksi paling banyak, yang dianggap sebagai protein yang signifikan [8]. Protein-protein signifikan tersebut kemudian dianalisis menggunakan IJAH Analytics (<http://ijah.apps.cs.ipb.ac.id/>) untuk mengidentifikasi tanaman obat yang mengandung protein tersebut. Selanjutnya, dilakukan pencarian protein dalam tanaman obat yang telah dikumpulkan sebelumnya, dan data protein yang signifikan dicocokkan dengan kandungan protein setiap

tanaman obat. Tanaman obat yang memiliki protein signifikan diidentifikasi sebagai kandidat obat alternatif untuk penyakit kanker [9].

HASIL DAN PEMBAHASAN

Ini harus mencakup temuan penelitian. Hasilnya harus disajikan dalam urutan logis dengan penjelasan tahapan yang digunakan dalam pengujian dengan memuat informasi hasil dalam teks atau bentuk tabel dan grafik.

Hasil Pelatihan dan Pengujian Model

Penelitian ini menggunakan 393 protein yang terkait dengan kanker paru, diperoleh dari pangkalan data OMIM. Selanjutnya, pencarian interaksi protein-protein dilakukan menggunakan pangkalan data STRING, yang menghasilkan 1.476 interaksi protein-protein dari total 301 protein, dengan 92 di antaranya tidak memiliki interaksi yang tercatat. Data ini kemudian dijadikan dasar untuk penelitian ini, sebagaimana terlihat dalam Tabel 1.

Tabel 1. Data Interaksi Protein-Protein

Node 1	Node 2	Node 1 Accession	Node 2 Accession	Score
CD44	CTNNB1	ENSP00000398632	ENSP00000344456	769
CD44	EGF	ENSP00000398632	ENSP00000265171	863
CD44	EGFR	ENSP00000398632	ENSP00000275493	997
CD44	ERBB2	ENSP00000398632	ENSP00000269571	999
...	...	...	...	...
STAT5A	CBL	ENSP00000293328	ENSP00000264033	525
STAT5B	CRKL	ENSP00000293328	ENSP00000346300	989
STAT5B	PIK3R1	ENSP00000293328	ENSP00000428056	513
STAT5B	STAT5A	ENSP00000293328	ENSP00000341208	943

Setelah membagi dataset, dilanjutkan dengan pembangunan model pertama menggunakan arsitektur Long Short-Term Memory (LSTM). Arsitektur LSTM ini terdiri dari beberapa lapisan, termasuk Embedding layer yang berfungsi untuk mengonversi bilangan

integer representasi sekuens protein menjadi vektor dengan dimensi 128. Dua LSTM layer digunakan untuk memodelkan urutan data, dilanjutkan dengan Dropout layer yang berperan dalam mencegah overfitting dengan mengabaikan secara acak unit dalam layer sebelumnya selama pelatihan. Selanjutnya, terdapat Dense layer dengan aktivasi Sigmoid yang menghasilkan probabilitas kelas interaksi atau tidak. Untuk simulasi pertama dengan menggunakan model RNN dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian Model RNN

Metric	Score
<i>Train MSE</i>	0.07032748311758041
<i>Test MSE</i>	0.05934073403477669
<i>Train Accuracy</i>	0.9254237413406372
<i>Test Accuracy</i>	0.9391891956329346

Dengan menggunakan Bidirectional LSTM dan penambahan Dense layer, model kedua memiliki lebih banyak lapisan dan lebih kompleks dalam memodelkan pola-pola dalam data sequence protein. Untuk simulasi pertama dengan menggunakan model RNN dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian Model Bidirectional

Metric	Score
<i>Train MSE</i>	0.03333985432982445
<i>Test MSE</i>	0.01726572960615158
<i>Train Accuracy</i>	0.9542372822761536
<i>Test Accuracy</i>	0.9763513803482056

Dari Tabel diatas, dapat disimpulkan bahwa model Bidirectional LSTM mengungguli model LSTM dari segi Mean Square Error (MSE), menunjukkan hasil yang lebih baik pada tahap pelatihan maupun pengujian. MSE yang lebih rendah pada kedua tahap tersebut mengindikasikan kemampuan model Bidirectional LSTM yang lebih baik dalam menyesuaikan diri dengan data dan menghasilkan prediksi yang lebih akurat. Selain itu, dari segi akurasi, model Bidirectional LSTM juga menunjukkan performa yang lebih tinggi dibandingkan dengan model LSTM, dengan tingkat akurasi yang tinggi pada tahap pelatihan dan pengujian. Hasil ini menandakan bahwa model Bidirectional LSTM mampu dengan baik memahami dan memodelkan pola-pola dalam data, sehingga berpotensi menghasilkan prediksi yang lebih akurat dalam konteks praktis.

Analisis Interaksi Protein-Protein

Selanjutnya melakukan analisis dari interaksi protein-protein dari data sebelumnya. Proses analisis interaksi protein-protein dilakukan dengan melakukan pengecekan setiap

protein dengan IJAH Analytics untuk melihat protein mana yang terkandung dalam tanaman obat. Berdasarkan analisis sebelumnya didapatkan beberapa tanaman obat dapat dilihat pada Tabel 4.

Tabel 4. Hasil Pengujian Model Bidirectional

Nama Protein	Nama Tanaman Obat
NTRK2	Codonopsis pilosula (Dangsam atau Dangshen)
VEGFA	Rehmannia glutinosa dan Lycium chinense (Gojiberi)

Berdasarkan Tabel 4, dapat disimpulkan bahwa terdapat korelasi antara nama protein dan tanaman obat yang relevan. Protein NTRK2 dikaitkan dengan *Codonopsis pilosula* (Dangsam atau Dangshen), sementara protein VEGFA dikaitkan dengan *Rehmannia glutinosa* dan *Lycium chinense* (Gojiberi). Informasi ini mencerminkan potensi interaksi antara protein-protein tersebut dengan senyawa-senyawa yang terkandung dalam tanaman obat tertentu. Relevansi ini menjadi sangat penting dalam bidang farmakologi, membantu peneliti memahami potensi pengaruh atau manfaat dari tanaman obat terhadap aktivitas protein spesifik. Sebagai contoh, pemahaman lebih lanjut mengenai hubungan ini dapat mendukung pengembangan obat-obatan atau terapi berbasis tanaman, memanfaatkan sifat-sifat bioaktif yang dimiliki tanaman obat untuk mendukung kesehatan dan pengobatan.

KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa pemodelan interaksi protein-protein menggunakan data dari pangkalan data STRING menghasilkan 1.476 interaksi protein-protein dalam format yang sesuai. Tahapan penelitian, mulai dari pengumpulan data hingga analisis, telah dilaksanakan. Evaluasi model Recurrent Neural Network dengan arsitektur LSTM mencapai tingkat akurasi sebesar 93%, dengan Mean Square Error (MSE) sebesar 0.05%. Sementara itu, dengan arsitektur Bidirectional LSTM, tingkat akurasi meningkat menjadi 97%, dengan MSE sebesar 0.01%. Analisis interaksi protein-protein menunjukkan bahwa protein NTRK2 dan VEGFA diidentifikasi sebagai protein yang signifikan dan terdapat dalam tanaman obat. *Codonopsis pilosula* (Dangsam atau Dangshen), *Rehmannia glutinosa*, dan *Lycium chinense* (Gojiberi) diidentifikasi sebagai tanaman obat yang potensial sebagai kandidat obat alternatif untuk penyakit kanker.

DAFTAR PUSTAKA

- Jawas, N., & Sentana, I. W. B. (2018). Deteksi Lung Nodule Otomatis dari Citra Medis Computed Tomography (CT) Scan dengan Operasi Morfology dan SVM. *Jurnal Sistem Dan Informatika*, 12(2), 140–148.
- Golbeck, J., Frago, G., Hartel, F., Hendler, J., Oberthaler, J., & Parsia, B. (2018). The National Cancer Institute's Thesaurus and Ontology. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3199007>

- Sun, T., Zhou, B., Lai, L., & Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, 18(1), 1–8. <https://doi.org/10.1186/s12859-017-1700-2>
- Ekor, M. (2014). The growing use of herbal medicines: issues relating to adverse reactions and challenges in monitoring safety. *Frontiers in Pharmacology*, 4, 177. <https://doi.org/10.3389/fphar.2013.00177>
- Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics*, 17(4), 696–712. <https://doi.org/10.1093/bib/bbv066>
- Min, S., Park, S., Kim, S., Choi, H. S., Lee, B., & Yoon, S. (2021). Pre-Training of Deep Bidirectional Protein Sequence Representations with Structural Information. *IEEE Access*, 9, 123912–123926. <https://doi.org/10.1109/ACCESS.2021.3110269>
- Zhao, L., Wang, J., Hu, Y., & Cheng, L. (2020). Conjoint Feature Representation of GO and Protein Sequence for PPI Prediction Based on an Inception RNN Attention Network. *Molecular Therapy - Nucleic Acids*, 22, 198–208. <https://doi.org/10.1016/j.omtn.2020.08.025>
- Diansyah, M. R., Kusuma, W. A., & Annisa, A. (2021). Identification of significant protein in protein-protein interaction of Alzheimer disease using top-k representative skyline query. *Jurnal Teknologi Dan Sistem Komputer*, 9(3), 126–132. <https://doi.org/10.14710/jtsiskom.2021.13985>
- Tsukiyama, S., Hasan, M. M., Fujii, S., & Kurata, H. (2021). LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. *Briefings in Bioinformatics*, 22(6). <https://doi.org/10.1093/bib/bbab228>