ABSTRAK

Kanker paru-paru merupakan salah satu penyakit dengan tingkat kematian tertinggi di dunia, dan pemahaman terhadap mekanisme biologisnya di tingkat molekuler menjadi kunci dalam pengembangan strategi diagnosis dan terapi yang lebih efektif. Pada tingkat molekuler, perkembangan kanker paru-paru dipicu oleh gangguan pada jaringan interaksi antar protein (*Protein-Protein Interaction*/PPI) di dalam sel. Akan tetapi, jaringan PPI yang tersedia saat ini masih belum lengkap karena metode pengujian laboratorium bersifat mahal, memakan waktu, dan rentan menghasilkan kesalahan deteksi. Kondisi tersebut menuntut pendekatan komputasional yang lebih efisien untuk memprediksi interaksi protein yang belum terdeteksi. Selain itu, belum banyak penelitian yang membandingkan berbagai metode berbasis topologi secara sistematis pada jaringan PPI penyakit tertentu seperti kanker paru-paru, sehingga sulit menentukan pendekatan yang paling sesuai untuk konteks tersebut.

Penelitian ini menerapkan pendekatan kuantitatif dengan desain komputasional secara *in silico* yang mengikuti kerangka kerja CRISP-DM. Data interaksi protein diperoleh dengan memasukkan daftar protein terkait kanker paru-paru ke dalam platform basis data STRING pada tiga variasi *confidence score*, yaitu 0,40, 0,70, dan 0,90, untuk menganalisis pengaruh tingkat kepercayaan data terhadap performa prediksi. Data mentah diproses melalui tahapan pembersihan, penyaringan, dan pembentukan graf tak berarah dan tak berbobot, dilanjutkan dengan pemisahan data latih dan uji serta pembuatan sampel negatif. Sistem prediksi dikembangkan dengan mengimplementasikan tujuh metode *link prediction* berbasis indeks kemiripan topologis, yaitu *Common Neighbors*, *Jaccard Coefficient*, *Adamic-Adar*, *Resource Allocation*, *Preferential Attachment*, *Local Path*, dan L3 *Index*, serta satu model *machine learning* berbasis *Random Forest* yang menggabungkan ketujuh fitur topologi tersebut. Performa seluruh metode dievaluasi menggunakan empat metrik yang saling melengkapi, yaitu AUC-ROC, AUPR, *Accuracy*, dan *Top-k Precision*.

Hasil pengujian menunjukkan seluruh metode mencapai AUC-ROC di atas 0,80 pada ketiga skenario, membuktikan bahwa pendekatan berbasis topologi relevan untuk memprediksi interaksi protein pada jaringan PPI kanker paru-paru. *Resource Allocation* menjadi metode tunggal terbaik pada skenario kepercayaan sedang dan tinggi, sedangkan *Random Forest* unggul pada skenario kepercayaan tertinggi yang bersifat paling *sparse*, sehingga penggabungan fitur terbukti lebih menguntungkan pada jaringan yang jarang. Analisis kontribusi fitur secara konsisten menunjukkan dominasi L3 *Index* yang semakin meningkat pada data berkualitas lebih tinggi, dan peningkatan *confidence score* secara konsisten menaikkan performa seluruh metode. Sebagai kontribusi, penelitian ini menghasilkan analisis komparatif yang terukur sekaligus daftar kandidat interaksi protein baru yang dapat dijadikan acuan awal bagi validasi eksperimental lanjutan.

Kata Kunci: *link prediction*, *Protein-Protein Interaction*, kanker paru-paru, fitur topologi jaringan, *Random Forest*

ABSTRACT

Lung cancer is one of the diseases with the highest mortality rates worldwide, and understanding its biological mechanisms at the molecular level is key to developing more effective diagnostic and therapeutic strategies. At the molecular level, lung cancer progression is triggered by disruptions in the protein-protein interaction (PPI) network within cells. However, currently available PPI networks remain incomplete because laboratory-based testing methods are costly, time-consuming, and prone to detection errors. This condition demands a more efficient computational approach to predict undetected protein interactions. Furthermore, few studies have systematically compared various topology-based methods on the PPI network of a specific disease such as lung cancer, making it difficult to determine the most suitable approach for that context.

This study applies a quantitative approach with an in silico computational design following the CRISP-DM framework. Protein interaction data were acquired by inputting a list of lung cancer related proteins into the STRING database platform at three confidence score variations, namely 0.40, 0.70, and 0.90, to analyze the effect of data confidence levels on prediction performance. The raw data were processed through cleaning, filtering, and construction of an undirected and unweighted graph, followed by training and testing data splitting and negative sample generation. The prediction system was developed by implementing seven topological similarity-based link prediction methods, namely Common Neighbors, Jaccard Coefficient, Adamic-Adar, Resource Allocation, Preferential Attachment, Local Path, and L3 Index, along with a Random Forest machine learning model that combines all seven topological features. The performance of all methods was evaluated using four complementary metrics, namely AUC-ROC, AUPR, Accuracy, and Top-k Precision.

The test results show that all methods achieved an AUC-ROC above 0.80 across the three scenarios, proving that the topology-based approach is relevant for predicting protein interactions in the lung cancer PPI network. Resource Allocation was the best single method in the medium and high confidence scenarios, while Random Forest was superior in the sparsest highest confidence scenario, demonstrating that feature combination is more advantageous in sparse networks. Feature contribution analysis consistently showed the dominance of the L3 Index, which increased on higher-quality data, and increasing the confidence score consistently improved the performance of all methods. As its contribution, this study produced a measurable comparative analysis as well as a list of novel candidate protein interactions that can serve as an initial reference for further experimental validation.

Keywords: *link prediction, protein-protein interaction, lung cancer, network topological features, Random Forest*