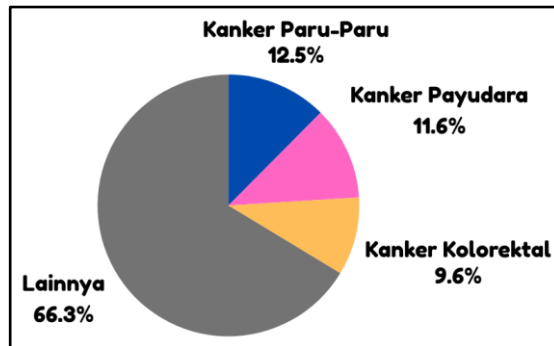


BAB I

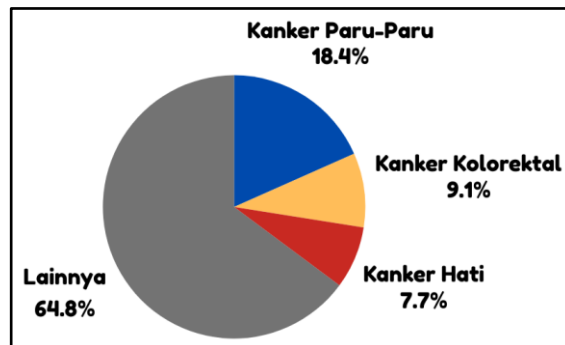
PENDAHULUAN

1.1 Latar Belakang

Kanker paru-paru merupakan salah satu masalah kesehatan yang terjadi secara signifikan pada masyarakat dengan tingkat kejadian penyakit (morbiditas) dan tingkat kematian (mortalitas) yang tinggi secara global. Menurut Woodman *et al.* (2021), kanker paru-paru merupakan jenis kanker paling umum yang menjadi penyebab kematian akibat kanker di dunia dengan peringkat kedua yang paling banyak didiagnosis pada laki-laki maupun perempuan. GLOBOCAN (*Global Cancer Observatory*) mengestimasi jumlah kasus kanker paru-paru untuk kedua jenis kelamin mencapai 2.480.675 kasus baru dengan tingkat kematiannya mencapai 1.817.469 kasus pada tahun 2022.



Gambar 1. Prevalensi Kasus Kanker Paru-paru di Dunia (2022)



Gambar 2. Prevalensi Kematian Kanker Paru-paru (2022)

Berdasarkan Gambar 1 dan Gambar 2, kanker paru-paru pada tahun 2022 menjadi kanker dengan peringkat pertama yang paling banyak didiagnosis dan penyebab utama kematian akibat kanker di seluruh dunia. Asmara *et al.* (2023) menyatakan bahwa pada tahun 2020 sebanyak 396.914 individu di Indonesia didiagnosis menderita kanker yang menewaskan sebanyak 234.511 individu. Kanker paru-paru sendiri mencapai 38.904 kasus di Indonesia pada tahun 2022. Kementerian Kesehatan Indonesia mengungkapkan bahwa penyakit tidak menular seperti kanker paru-paru menjadi salah satu dari tiga penyebab utama morbiditas dan mortalitas yang tinggi di Indonesia.

Pengobatan konvensional atau terapi menjadi standar pengobatan kanker paru-paru yang sebenarnya pengobatan ini memiliki keterbatasan yang signifikan. Menurut Zhou *et al.* (2024), terapi konvensional kanker paru-paru seperti *radiotherapy*, *chemotherapy*, dan *immunotherapy* sudah berkembang, tetapi tingkat kejadian dan kematian terus meningkat secara ekstrem di berbagai wilayah dan negara. Kegagalan terapi standar dengan pendekatan “*one-size-fits-all*” terjadi karena jarang mempertimbangkan variasi dari faktor biologis, genetik, dan lingkungan. Menurut Li *et al.* (2023), perbedaan mutasi dan karakteristik tumor kanker memberi pengaruh langsung terhadap efektivitas pengobatan, sehingga terapi standar akan menghasilkan respons yang sangat heterogen, artinya setiap pasien tidak memberikan respons sama untuk terapi seragam yang menyebabkan pengobatan menjadi tidak optimal untuk seluruh pasien. Heterogenitas biologis kanker paru-paru harus diatasi dengan adanya stratifikasi dan penilaian risiko individual sehingga efektivitas pengobatan akan meningkat dan angka mortalitas menjadi berkurang (Zhou, *et al.*, 2024).

Pengobatan presisi dengan pendekatan dasar molekuler dan genetik tumor kanker lebih efektif dibandingkan pengobatan standar yang bersifat “*one-size-fits-all*”. Namini *et al.* (2025) mengungkapkan bahwa penerapan farmakogenomik merupakan kunci penting dalam terapi kanker, artinya pengobatan dengan konsep personalisasi yang mengidentifikasi mutasi genetik kanker paru-paru secara spesifik dapat meningkatkan efektivitas terapi dan mencegah efek samping. Hal ini menegaskan bahwa pengobatan presisi menjadi fokus penting untuk pengembangan terapi kanker paru-paru.

Pendekatan pengobatan presisi dengan model prediktif harus dikembangkan untuk mempelajari efektivitas obat dengan mempertimbangkan profil molekuler sel kanker paru-paru. Penelitian Yu *et al.* (2023) mengungkapkan bahwa algoritma *machine learning* diterapkan untuk memetakan hubungan antara jalur biologis tertentu dengan resistensi obat. Karakteristik biologis ditemukan pada data omik yang memiliki volume dan kompleksitas tinggi, sehingga *machine learning* menjadi kunci untuk analisis data ini. Pendekatan *machine learning* berbasis integrasi data omik seperti proteomik telah menjadi strategi yang menjanjikan untuk membangun model prediksi respons obat. Sebagian besar obat berinteraksi dengan protein sebagai target utama, sehingga data proteomik dapat merepresentasikan keadaan sel secara akurat. Hal ini ditegaskan oleh penelitian komparatif bahwa model berbasis data proteomik (protein) menunjukkan kekuatan prediktif yang lebih tinggi dibandingkan dengan model berbasis data transkriptomik (RNA) (Zheng, *et al.*, 2024).

Sebagian sistem yang kompleks, termasuk struktur kimia obat dalam format *Simplified Molecular Input Line Entry System* (SMILES) memerlukan tahap

ekstraksi fitur yang mengubah representasi molekul menjadi vektor numerik agar dapat dimodelkan menggunakan *machine learning*. Pendekatan konvensional seperti Morgan Fingerprint telah lama digunakan karena lebih sederhana dan efisien dalam merepresentasikan struktur molekul, serta mudah diimplementasikan pada berbagai model prediktif (Gil and Rowley, 2024). Metode ini memiliki keterbatasan mendasar karena representasinya yang bersifat statis. Representasi berbasis fingerprint hanya bergantung pada sub-struktur kimia yang telah ditentukan sebelumnya, sehingga informasi topologis molekuler yang bersifat kontekstual berisiko terabaikan (Jiang, *et al.*, 2021). Keterbatasan tersebut diatasi dengan pendekatan *Graph Neural Networks* (GNN) dikembangkan sebagai alternatif representasi molekul yang lebih adaptif. GNN memiliki kemampuan untuk mempelajari fitur langsung dari struktur graf molekul dan menangkap hubungan kontekstual yang relevan dalam interaksi antara obat dan protein (Jiang, *et al.*, 2021).

Pendekatan model *machine learning* diperlukan untuk menemukan model klasifikasi respons obat yang tepat. Model klasifikasi yang digunakan dalam penelitian ini yaitu Random Forest, XGBoost, dan *Multi-layer Perceptron* (MLP). Random Forest merupakan metode berbasis *bagging* yang sudah dikenal menjadi *baseline* kuat dan sering digunakan untuk memodelkan data berdimensi tinggi seperti proteomik (Orreliid, *et al.*, 2025). Model ini memberikan performa yang stabil dalam menangani data dengan fitur yang besar. Model XGBoost menawarkan metode yang kompetitif pada tabular data menggunakan pendekatan *boosting* yang dilengkapi fitur regulasi kuat untuk menghindari *overfitting* (Hakkal and Lahcen, 2024). Kedua model berbasis pohon keputusan yaitu Random Forest dan XGBoost

dibandingkan dengan model *Multi-layer Perceptron* (MLP) dengan arsitektur yang sangat berbeda yaitu jaringan syaraf tiruan. Model *Multi-layer perceptron* (MLP) merupakan bentuk dasar dari *feedforward neural network* dengan komposisi tiga jenis lapisan, yaitu *input layer*, *hidden layer*, dan *output layer* (Fikadu, *et al.*, 2024). Model ini menangkap pola non-linier yang tidak dipelajari oleh kedua model di atas. Ketiga model klasifikasi ini memiliki beberapa keunggulan, yaitu lebih adaptif untuk data yang berdimensi tinggi, proses pemodelan cepat, dan menghasilkan probabilitas secara natural untuk prediksi respons obat.

Penelitian ini bertujuan untuk melakukan evaluasi terhadap berbagai arsitektur model *machine learning* untuk memprediksi respons obat pada kanker paru-paru berbasis data proteomik. Fokus utama penelitian adalah membandingkan dua metode representasi fitur obat, yaitu Morgan Fingerprint sebagai pendekatan klasik dan *Graph Neural Networks* (GNN) sebagai pendekatan modern. Dua representasi fitur obat ini dikombinasikan dengan tiga algoritma klasifikasi, yaitu Random Forest, XGBoost, dan *Multi-layer Perceptron* (MLP). Seluruh kombinasi representasi molekul dan model prediktif dievaluasi menggunakan *Area Under Precision Recall Curve* (AUPRC) dan *Confusion Matrix* dengan harapan kombinasi model ini mampu menghasilkan model prediktif yang akurat dan *robust* untuk mendukung pengembangan pengobatan presisi pada kanker paru-paru. Model prediktif terbaik yang dihasilkan dalam penelitian ini dapat diimplementasikan pada bidang kesehatan untuk menemukan obat yang efektif dengan langsung menyeleksi obat yang sensitif tanpa harus *trial-error* obat terlebih dahulu pada pasien.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, dapat dirumuskan permasalahan sebagai berikut.

1. Bagaimana perbandingan kualitas representasi molekul obat antara metode *Graph Neural Networks* (GNN) dengan Morgan Fingerprint dalam menangkap informasi fitur berdasarkan nilai tingkat *sparsity*?
2. Bagaimana performa tiga model prediktif *machine learning* yaitu Random Forest, XGBoost, dan *Multi-layer Perceptron* (MLP) yang dikombinasikan dengan *Graph Neural Networks* (GNN) dan Morgan Fingerprint berdasarkan evaluasi nilai AUPRC?
3. Bagaimana arsitektur model *machine learning* yang paling optimal dalam memprediksi respons obat kanker paru-paru berdasarkan nilai evaluasi pada *Confusion Matrix*?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini bertujuan agar analisisnya lebih fokus dan tidak meluas dari pembahasan yang dimaksud. Oleh karena itu, penulis membatasi penelitian ini dalam ruang lingkup sebagai berikut.

1. Data dalam penelitian ini yaitu data sekunder *in-vitro* (eksperimen di luar tubuh) yang berasal dari *cell line* kanker pada basis data publik yaitu DepMap dan GDSC. Model yang dihasilkan bersifat *in-silico* (komputasi) sehingga tidak divalidasi menggunakan data pasien klinis (*in-vivo*).
2. Ruang lingkup penelitian dibatasi hanya untuk kasus kanker paru-paru, sesuai dengan proses seleksi dan sinkronisasi data berdasarkan metadata *cell line* dan tabel pemetaan uniprot.

3. Data yang digunakan dibagi menjadi data *training* dan *testing* dengan perbandingan 80:20 yang dibagi secara acak untuk pelatihan model *machine learning*.
4. Variabel prediktor yang digunakan dalam penelitian ini yaitu representasi struktur molekul obat dan data omik yang terbatas pada proteomik, tidak melibatkan jenis data omik lain seperti genomik (DNA) atau transkriptomik (RNA).
5. Pendekatan pembelajaran prediksi yang digunakan adalah klasifikasi biner kelas respons obat (sensitif dan resisten).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, tujuan dari penelitian ini sebagai berikut.

1. Membandingkan kualitas dari representasi molekul obat antara metode *Graph Neural Networks* (GNN) dengan Morgan Fingerprint dalam menangkap informasi fitur berdasarkan nilai tingkat *sparsity*.
2. Mengevaluasi dan membandingkan kinerja tiga model prediktif *machine learning* yang diuji yaitu Random Forest, XGBoost, dan *Multi-layer Perceptron* (MLP) yang dikombinasikan dengan Morgan Fingerprint dan *Graph Neural Networks* (GNN) berdasarkan evaluasi nilai metrik AUPRC.
3. Mengevaluasi satu arsitektur model *machine learning* yang paling optimal untuk memprediksi respons obat kanker paru-paru berdasarkan nilai evaluasi pada *Confusion Matrix*.