

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker payudara merupakan salah satu jenis kanker yang paling banyak ditemukan pada perempuan di berbagai negara dan termasuk penyebab utama kematian akibat kanker di seluruh dunia [1]. Pada tahun 2020, kanker payudara diperkirakan menyebabkan sekitar 685.000 kematian pada perempuan di seluruh dunia, yang setara dengan sekitar 16 persen atau 1 dari setiap 6 kematian akibat kanker pada perempuan [2]. Besarnya jumlah kasus dan tingkat kematian akibat kanker payudara menunjukkan bahwa penyakit ini masih menjadi tantangan kesehatan global yang signifikan sehingga diperlukan upaya penanganan yang efektif dan berkelanjutan.

Prediksi kelangsungan hidup pasien merupakan aspek penting dalam dunia medis. Dengan adanya prediksi yang akurat, tenaga medis dapat lebih memahami kondisi pasien dan menentukan rencana pengobatan yang sesuai [3]. Namun, diagnosis dan penentuan pengobatan kanker payudara masih menjadi tantangan karena kompleksitas penyakit serta beragamnya hasil klinis yang dapat terjadi pada pasien [4]. Selain itu, metode diagnostik konvensional seperti mammography dan analisis biopsi memiliki keterbatasan dalam akurasi serta bergantung pada interpretasi tenaga medis, sehingga berpotensi menimbulkan variasi dalam hasil diagnosis [5].

Untuk menjawab permasalahan tersebut, berbagai metode *machine learning* telah dikembangkan dan diterapkan guna memprediksi tingkat kelangsungan hidup pasien kanker payudara [6]. Berbagai penelitian telah memanfaatkan metode ini untuk menganalisis data klinis maupun biologis, termasuk data *multi-omics*, dalam memprediksi tingkat kelangsungan hidup pasien melalui identifikasi pola yang kompleks [7, 8, 6].

Dalam praktik klinis, penentuan status kelangsungan hidup pasien umumnya memerlukan periode observasi hingga 5 tahun, sehingga banyak penelitian menggunakan batas tersebut sebagai acuan dalam mengukur tingkat kelangsungan hidup pasien [9]. Periode ini telah menjadi standar dalam pelaporan klinis dan dianggap cukup representatif untuk menggambarkan perkembangan penyakit. Namun, beberapa penelitian juga menggunakan horizon waktu yang lebih pendek,

seperti 3 tahun, untuk memperoleh gambaran kondisi pasien dalam jangka waktu yang lebih cepat [10, 11]. Penggunaan lebih dari satu horizon waktu juga memungkinkan analisis yang lebih komprehensif terhadap kelangsungan hidup pasien. Dengan demikian, penelitian ini menggunakan dua horizon waktu, yaitu 3 tahun dan 5 tahun, untuk mengklasifikasikan kelangsungan hidup pasien menjadi jangka pendek dan jangka panjang.

Karakteristik klinis pasien merupakan salah satu faktor yang dapat memengaruhi kelangsungan hidup. Beberapa penelitian menunjukkan adanya perbedaan tingkat kelangsungan hidup antar kelompok ras, di mana pasien dari kelompok ras Black cenderung memiliki tingkat kelangsungan hidup yang lebih rendah dibandingkan kelompok ras White, meskipun faktor lain seperti usia dan stadium tumor juga berkontribusi terhadap perbedaan tersebut [12]. Meskipun distribusi jumlah sampel pada kedua kelompok ras tidak seimbang dan didominasi oleh pasien ras White, penelitian ini tetap mengikutsertakan pasien dari kelompok ras White dan Black untuk mempertahankan keberagaman data dan memungkinkan analisis yang lebih representatif terhadap pengaruh ras pada kelangsungan hidup pasien.

Meskipun berbagai penelitian telah memanfaatkan data klinis dan data multi-omics untuk prediksi kelangsungan hidup pasien, pemanfaatan data epigenetik dalam penelitian serupa masih relatif terbatas. Salah satu jenis data epigenetik yang menjanjikan adalah DNA methylation, yang berperan penting dalam regulasi ekspresi gen dan perkembangan kanker [13]. Perubahan pola methylation dapat digunakan sebagai biomarker untuk memahami kondisi penyakit serta mendukung prediksi kelangsungan hidup pasien.

Namun, data DNA methylation umumnya memiliki jumlah fitur yang sangat besar dibandingkan jumlah sampel sehingga termasuk dalam kategori data berdimensi tinggi (*high-dimensional data*) [14]. Karakteristik tersebut meningkatkan kompleksitas proses analisis dan pemodelan, yang pada akhirnya dapat memengaruhi kinerja prediksi. Pada data dengan karakteristik tersebut, algoritma *machine learning* seperti *Support Vector Machine* (SVM), *Random Forest*, dan jaringan saraf tiruan (*neural network*) dinilai lebih efektif karena mampu mempelajari hubungan nonlinier serta interaksi antarfitur pada data DNA methylation [13]. Berdasarkan hal tersebut, penelitian ini bertujuan untuk mengembangkan serta membandingkan kinerja *Support Vector Machine* (SVM), *Extreme Gradient Boosting* (XGBoost), *Random Forest*, dan *Multilayer Perceptron* (MLP) dalam memprediksi kelangsungan hidup pasien kanker payudara dengan

memanfaatkan data DNA methylation dan data klinis.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini merumuskan beberapa permasalahan yang akan dikaji sebagai berikut.

1. Bagaimana penerapan model *machine learning* untuk memprediksi kelangsungan hidup pasien kanker payudara menggunakan data DNA methylation, *clinical data*, dan *survival data*?
2. Bagaimana kinerja model dalam mengklasifikasikan kelangsungan hidup pasien berdasarkan horizon waktu 3 tahun dan 5 tahun sebagai indikator kelangsungan hidup jangka pendek dan jangka panjang?
3. Bagaimana pengaruh penggunaan fitur klinis berupa ras, usia, dan stadium kanker (*stage*) terhadap kinerja model dalam prediksi kelangsungan hidup pasien kanker payudara?
4. Bagaimana perbandingan kinerja algoritma Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest, dan Multilayer Perceptron (MLP) dalam memprediksi kelangsungan hidup pasien kanker payudara?

1.3 Batasan Permasalahan

Untuk menjaga fokus penelitian serta membatasi ruang lingkup pembahasan, penelitian ini ditetapkan beberapa batasan sebagai berikut.

1. Data yang digunakan merupakan dataset TCGA-BRCA yang diperoleh dari UCSC Xena Browser, meliputi data DNA methylation (Illumina HumanMethylation450), *clinical data*, dan *survival data*.
2. Prediksi kelangsungan hidup dilakukan berdasarkan dua horizon waktu, yaitu 3 tahun dan 5 tahun.
3. Penelitian ini membatasi penggunaan fitur klinis pada tiga aspek utama, yaitu ras (*race*), usia (*age*), serta stadium kanker (*stage*).

4. Pemodelan dilakukan menggunakan empat algoritma machine learning, yaitu Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest, dan Multilayer Perceptron (MLP).
5. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, Macro F1-Score, dan Area Under the ROC Curve (ROC-AUC).

1.4 Tujuan Penelitian

Berdasarkan permasalahan yang telah diidentifikasi, penelitian ini bertujuan untuk:

1. Menerapkan model *machine learning* untuk memprediksi kelangsungan hidup pasien kanker payudara menggunakan data DNA methylation, *clinical data*, dan *survival data*.
2. Mengevaluasi kinerja model dalam mengklasifikasikan kelangsungan hidup pasien kanker payudara berdasarkan horizon waktu 3 tahun dan 5 tahun sebagai indikator kelangsungan hidup jangka pendek dan jangka panjang.
3. Menganalisis pengaruh penggunaan fitur klinis berupa ras, usia, dan stadium kanker terhadap kinerja model prediksi kelangsungan hidup pasien kanker payudara.
4. Membandingkan kinerja algoritma Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest, dan Multilayer Perceptron (MLP) dalam memprediksi kelangsungan hidup pasien kanker payudara.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, yaitu sebagai berikut.

1. Memberikan pemahaman mengenai penerapan dan perbandingan berbagai metode machine learning, yaitu Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest, dan Multilayer Perceptron (MLP), dalam prediksi kelangsungan hidup pasien kanker payudara.

2. Menjadi referensi dalam pemanfaatan data DNA methylation, *clinical data*, dan *survival data* untuk analisis prediksi kelangsungan hidup pasien kanker payudara.
3. Memberikan informasi mengenai pengaruh penggunaan fitur klinis berupa ras, usia, dan stadium kanker terhadap kinerja model prediksi.
4. Memberikan kontribusi dalam pengembangan model prediksi kelangsungan hidup yang mampu menangani data DNA methylation berdimensi tinggi secara efektif dan efisien.

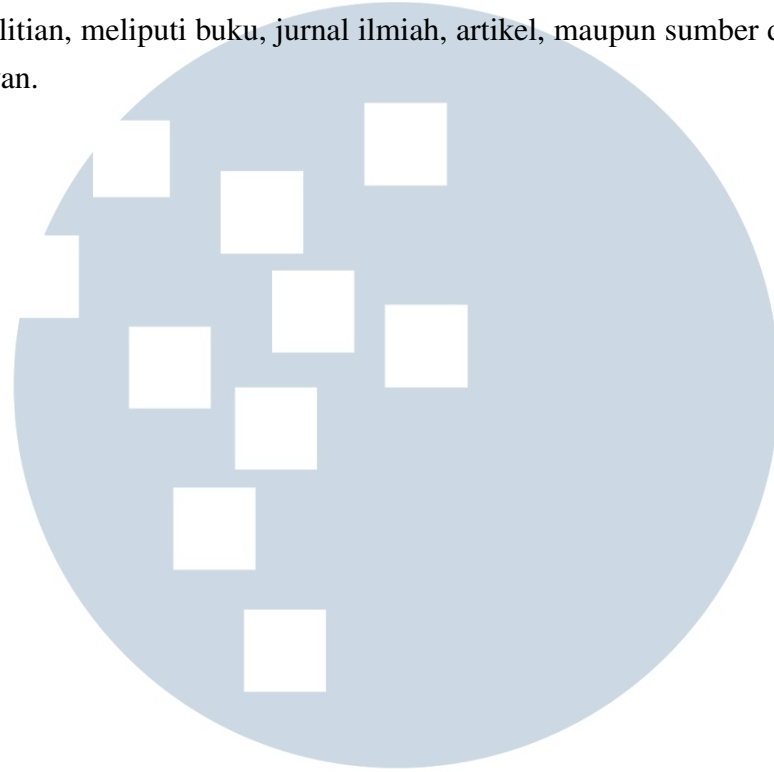
1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut.

- **Bab 1 PENDAHULUAN**
Bab ini berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
- **Bab 2 LANDASAN TEORI**
Bab ini menguraikan landasan teori yang menjadi dasar penelitian, mencakup pembahasan mengenai kanker payudara, DNA methylation, *machine learning*, serta algoritma Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest, dan Multilayer Perceptron (MLP) yang diterapkan dalam penelitian.
- **Bab 3 METODOLOGI PENELITIAN**
Bab ini menjelaskan metodologi penelitian yang digunakan, mulai dari proses pengumpulan data, tahap prapemrosesan data, pembangunan model, teknik evaluasi untuk menilai kinerja model, hingga penerapan SHAP (*SHapley Additive exPlanations*) untuk menginterpretasikan hasil prediksi model.
- **Bab 4 HASIL DAN DISKUSI**
Bab ini menyajikan hasil eksperimen yang diperoleh beserta analisis dan pembahasan untuk menginterpretasikan temuan penelitian.
- **Bab 5 KESIMPULAN DAN SARAN**
Bab ini memuat kesimpulan yang diperoleh berdasarkan hasil penelitian serta rekomendasi untuk pengembangan penelitian pada masa mendatang.

- **DAFTAR PUSTAKA**

Bagian ini berisi seluruh sumber referensi yang digunakan dalam penyusunan penelitian, meliputi buku, jurnal ilmiah, artikel, maupun sumber daring yang relevan.



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA