

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker prostat adalah jenis kanker yang paling sering didiagnosis pada pria dan menjadi penyebab utama fatalitas terkait kanker, khususnya di Amerika Serikat [1]. Penyakit ini umumnya terdapat pada kelompok usia lanjut sebab faktor usia, riwayat keluarga, serta gaya hidup seseorang [2]. Secara biologis, kanker prostat memiliki sifat heterogen, memiliki banyak subtype dan pola ekspresi gen yang berbeda, sehingga mempengaruhi tingkat keparahan, respons terhadap terapi, dan prognosis pasien [3, 4]. Variasi tipe kanker prostat menimbulkan tantangan dalam proses diagnosis dan klasifikasi yang akurat, terutama ketika sebuah metode diterapkan pada populasi yang beragam.

Dalam praktik klinis, klasifikasi kanker prostat biasanya didasarkan pada *Gleason score* dan status reseptor androgen yang mengindikasikan tingkat agresivitas tumor. Selain itu, mutasi genetik tertentu seperti pada gen *TMPRSS2-ERG*, *PTEN* dan, *SPOP* juga berperan dalam perkembangan penyakit. Meskipun kemajuan teknologi seperti *next generation sequencing* (NGS) telah memungkinkan identifikasi biomarker baru, banyak metode klasifikasi kanker prostat yang belum mampu menangkap kompleksitas variasi biologis antar stadium kanker secara optimal [5] [6] [7]. Salah satu pendekatan yang berkembang adalah analisis metilasi *deoxyribonucleic acid* (DNA), yaitu mekanisme epigenetik yang mempengaruhi ekspresi gen tanpa mengubah urutan DNA. Data metilasi membuka potensi besar untuk mendeteksi kanker prostat secara dini dan memahami pertumbuhan penyakit pada tingkat lebih dalam [8]. Namun, kompleksitas data metilasi DNA yang cenderung tinggi (*high dimensional*) menuntut penggunaan metode komputasi yang mampu mengekstrak informasi yang relevan secara klinis [9]. Analisis *differentially methylated positions* (DMP) menjadi salah satu solusi untuk mengidentifikasi CpG *site* yang signifikan, tetapi penggunaannya masih lebih jarang dikombinasikan dengan metode *machine learning* seperti *multilayer perceptron* (MLP) untuk klasifikasi stadium kanker [10] [11].

Penerapan model MLP untuk klasifikasi data metilasi DNA dapat menguntungkan karena kemampuannya dalam memproses data yang kompleks dan berdimensi tinggi. MLP juga mampu mengidentifikasi pola kompleks yang sulit

dideteksi menggunakan metode konvensional. Namun, performa MLP dipengaruhi oleh beberapa faktor, seperti jumlah data *training*, keseimbangan kelas, dan tingginya dimensi data. Faktor-faktor tersebut dapat memengaruhi kemampuan model dalam melakukan generalisasi dan meningkatkan risiko terjadinya *overfitting* maupun *underfitting* [12]. Selain itu, perbedaan karakteristik biologis antar kelompok etnis dan ras juga dapat memengaruhi pola metilasi DNA sehingga berpotensi menurunkan kemampuan generalisasi model ketika diterapkan pada populasi yang berbeda.

Pada sisi lain, penelitian menunjukkan bahwa pola metilasi DNA bervariasi secara signifikan antar kelompok ras atau etnis. Perbedaan ini dipengaruhi oleh faktor genetik, lingkungan, serta gaya hidup individu, sehingga menghasilkan variasi epigenetik yang tidak universal [13]. Hal ini menunjukkan bahwa model prediksi berbasis data biologis tidak dapat secara optimal digunakan secara general untuk semua populasi tanpa mempertimbangkan faktor ras. Klasifikasi kanker menggunakan metode *machine learning*, khususnya MLP telah banyak digunakan dalam klasifikasi kanker dan terbukti efektif dalam menangani data berdimensi tinggi, sebagian besar penelitian sebelumnya masih memiliki keterbatasan dimana banyak studi berfokus pada data citra medis atau memiliki tujuan untuk klasifikasi umum kanker jinak (*benign*) dan ganas (*malignant*) tanpa secara spesifik mengklasifikasikan stadium kanker atau mempertimbangkan perbedaan ras [14]. Selain itu penggunaan data metilasi DNA dalam klasifikasi stadium kanker prostat masih relatif terbatas, dan upaya untuk meminimalisir jumlah fitur demi meningkatkan interpretabilitas model juga belum banyak dilakukan [15].

1.2 Rumusan Masalah

Bagian ini merumuskan permasalahan utama yang menjadi fokus penelitian berdasarkan celah yang telah diidentifikasi pada studi sebelumnya. Permasalahan tersebut dinyatakan dalam bentuk pertanyaan ilmiah sebagai berikut:

- Bagaimana membangun model klasifikasi stadium kanker prostat berbasis data metilasi DNA menggunakan MLP yang mampu membedakan stadium awal dan stadium lanjut secara akurat?
- Bagaimana mengidentifikasi dan meminimalkan jumlah fitur metilasi DNA yang digunakan tanpa menurunkan kinerja model, sehingga meningkatkan interpretabilitas dan mengurangi risiko *overfitting*?

- Bagaimana kinerja model yang dilatih pada kohort ras White ketika diaplikasikan pada kohort ras Asian dan Black atau African American?
- Apakah diperlukan pendekatan model spesifik ras untuk meningkatkan akurasi dan keadilan dalam klasifikasi stadium kanker prostat berbasis data metilasi DNA?

1.3 Batasan Permasalahan

Bagian ini menjabarkan batasan penelitian untuk memperjelas ruang lingkup dan mempertaham fokus sesuai dengan rumusan masalah yang telah ditetapkan. Adapun batasan permasalahan dalam penelitian ini adalah sebagai berikut:

- Data yang digunakan berasal dari UCSC Xenabrowser dan terbatas pada nilai metilasi DNA dengan jumlah 556 pasien, serta memiliki ketidakseimbangan distribusi ras dan *missing value*.
- Variabel penelitian dibatasi pada CpG site hasil seleksi menggunakan metode DMP dan *Recursive Feature Elimination* (RFE) sebagai variabel bebas, serta klasifikasi stadium kanker prostat menjadi *early stage* (1–2) dan *late stage* (3–4) sebagai variabel terikat.
- Model yang digunakan terbatas pada MLP dan difokuskan hanya untuk tugas klasifikasi tanpa perbandingan model lain.
- Model dilatih menggunakan kohort ras White dan diuji pada ras lain tanpa pembangunan model terpisah akibat keterbatasan data.
- Penelitian tidak mempertimbangkan variabel klinis maupun faktor non-genetik, serta tidak menggunakan jenis data selain metilasi DNA.

1.4 Tujuan Penelitian

Bagian ini menjelaskan tujuan yang ingin dicapai dalam penelitian berdasarkan rumusan masalah yang telah ditetapkan. Adapun tujuan penelitian ini adalah sebagai berikut:

- Mengembangkan model klasifikasi stadium kanker prostat berbasis data metilasi DNA menggunakan *artificial neural network*, khususnya MLP, untuk membedakan stadium awal (*early*) dan stadium lanjut (*late*).

- Mengidentifikasi dan memperoleh koleksi jumlah minimum CpG site yang tetap informatif melalui kombinasi metode DMP dan RFE.
- Mengevaluasi efektivitas metode seleksi fitur dalam meningkatkan performa model sekaligus menjaga interpretabilitas.
- Meneliti pengaruh perbedaan ras terhadap performa model serta kebutuhan pengembangan model spesifik ras dalam klasifikasi stadium kanker prostat berbasis metilasi DNA.

1.5 Manfaat Penelitian

Bagian ini menguraikan manfaat yang diharapkan dari hasil penelitian yang dilakukan. Adapun manfaat penelitian ini adalah sebagai berikut:

- Memberikan kontribusi dalam pengembangan model klasifikasi stadium kanker prostat berbasis data metilasi DNA sebagai alternatif selain data pencitraan.
- Menyediakan pendekatan seleksi fitur yang lebih efisien melalui kombinasi metode DMP dan RFE untuk menghasilkan fitur yang minimal dan informatif.
- Meningkatkan pemahaman mengenai pengaruh perbedaan ras terhadap performa model *machine learning* dalam analisis data biologis.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
Bab ini berisi pendahuluan yang mencakup latar belakang masalah, rumusan masalah, batasan penelitian, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.
- Bab 2 LANDASAN TEORI
Bab ini membahas landasan teori yang mendukung penelitian, meliputi

konsep bioinformatika, metilasi DNA, machine learning, metode seleksi fitur, arsitektur MLP, serta metrik evaluasi dan teknik interpretasi model.

- **Bab 3 METODOLOGI PENELITIAN**

Bab ini menjelaskan metodologi penelitian yang digunakan, termasuk desain studi, pengumpulan dan praproses data, pelabelan data, seleksi fitur menggunakan DMP dan RFE, pengembangan model MLP, serta metode evaluasi performa model.

- **Bab 4 HASIL DAN DISKUSI**

Bab ini menyajikan hasil penelitian dan pembahasan, meliputi hasil pengolahan data, performa model yang dihasilkan, analisis hasil, serta perbandingan dengan penelitian terdahulu.

- **Bab 5 KESIMPULAN DAN SARAN**

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan serta saran untuk pengembangan penelitian selanjutnya.

