

BAB 1 PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker prostat merupakan kanker kedua yang paling sering didiagnosis pada pria di seluruh dunia, dengan perkiraan sebanyak 1.414.000 kasus baru dan 375.304 kematian pada tahun 2020 [1]. Kanker ini menjadi jenis kanker yang paling sering didiagnosis di 112 negara, serta menjadi penyebab utama kematian akibat kanker di 48 negara [2]. Selain itu, terdapat berbagai faktor yang memengaruhi angka kematian akibat kanker prostat, salah satunya adalah perbedaan ras pada pasien. Saat ini, teknik machine learning telah banyak digunakan dalam penelitian untuk mengembangkan metode diagnosis kanker prostat [3, 4]. Teknik machine learning mampu menemukan pola dari dataset yang kompleks sehingga dapat digunakan untuk memprediksi hasil atau perkembangan kanker prostat secara lebih efektif [5].

Perbedaan ras memiliki peran yang signifikan dalam diagnosis kanker prostat, yang dapat memengaruhi hasil deteksi pada berbagai kelompok ras [6, 7]. Di Amerika Serikat, pria berkulit hitam memiliki kemungkinan 1,76 kali lebih tinggi untuk didiagnosis kanker prostat dan memiliki tingkat kematian 2,14 kali lebih tinggi dibandingkan pria berkulit putih [8]. Selain itu, pria berkulit hitam cenderung didiagnosis pada stadium penyakit yang lebih lanjut [9]. Perbedaan ini menunjukkan pentingnya analisis berbasis ras untuk mengurangi risiko kesalahan diagnosis serta meningkatkan akurasi deteksi. Oleh karena itu, pemahaman terhadap perbedaan ras menjadi aspek penting dalam upaya menghasilkan diagnosis kanker prostat yang lebih akurat.

Selain faktor ras, penggunaan biomarker molekuler juga menjadi perhatian dalam pengembangan metode diagnosis kanker prostat [10]. Salah satu biomarker yang banyak dimanfaatkan adalah *microRNA* (miRNA), karena perubahan pola ekspresinya berkaitan dengan perkembangan dan tingkat keparahan kanker prostat [11, 12]. Informasi yang terkandung dalam data ekspresi miRNA memungkinkan identifikasi karakteristik biologis yang sulit diamati melalui data klinis saja [13]. Oleh karena itu, data ekspresi miRNA memiliki potensi untuk dimanfaatkan sebagai fitur dalam pengembangan model klasifikasi berbasis *machine learning* guna meningkatkan akurasi prediksi stadium kanker prostat [14, 15].

Beberapa penelitian telah memanfaatkan data ekspresi miRNA untuk klasifikasi kanker prostat menggunakan berbagai algoritma *machine learning*, seperti SVM dan *logistic regression*. Penelitian yang dilakukan oleh Kalaiyarasi et al. menggunakan dataset miRNA pada kanker prostat dan melakukan seleksi fitur untuk mengidentifikasi miRNA yang paling signifikan, dengan hasil sebanyak 50 fitur miRNA terpilih. Penelitian tersebut menghasilkan akurasi sebesar 91% dan nilai AUC sebesar 94% menggunakan algoritma SVM [16]. Penelitian lain yang dilakukan oleh Fernando et al. menggunakan 4 miRNA tanpa melalui proses seleksi fitur untuk mengidentifikasi biomarker kanker prostat, dengan menghasilkan akurasi sebesar 85% menggunakan metode *logistic regression* [17].

Berbeda dengan penelitian sebelumnya, penelitian ini menggunakan data yang diperoleh dari GDC TCGA Xena Browser dan menerapkan algoritma XGBoost untuk klasifikasi stadium kanker prostat. Selain itu, jumlah fitur miRNA yang digunakan dalam penelitian ini lebih sedikit karena melalui proses seleksi fitur. Penelitian ini juga menerapkan berbagai metode seleksi fitur, seperti T-test, Mutual Information, Chi-square, dan limma, untuk mengidentifikasi fitur miRNA yang signifikan. Selain itu, digunakan berbagai teknik penyeimbangan data untuk mengatasi ketidakseimbangan kelas. Data yang digunakan berfokus pada pasien ras putih, dengan pengujian dilakukan pada pasien ras hitam untuk mengamati pengaruh perbedaan ras. Model yang dibangun juga diinterpretasikan menggunakan LIME untuk mengetahui kontribusi fitur terhadap hasil klasifikasi. Dengan kombinasi metode tersebut, penelitian ini bertujuan untuk membangun model klasifikasi stadium kanker prostat yang lebih akurat sehingga memungkinkan deteksi kanker prostat pada tahap yang lebih awal.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut.

1. Bagaimana membangun dan mengoptimalkan model klasifikasi stadium kanker prostat menggunakan algoritma XGBoost berbasis data ekspresi miRNA?
2. Bagaimana pengaruh penerapan feature selection, teknik data balancing, dan hyperparameter tuning dalam mengoptimalkan performa model klasifikasi XGBoost?

3. Bagaimana performa model hasil optimasi pada perbedaan ras serta bagaimana interpretasi kontribusi fitur miRNA terhadap hasil prediksi menggunakan LIME?

1.3 Batasan Permasalahan

Batasan permasalahan dalam penelitian ini adalah sebagai berikut.

1. Data yang digunakan berupa data ekspresi miRNA dan data klinis kanker prostat dari The Cancer Genome Atlas (TCGA) melalui UCSC Xena Browser, dengan klasifikasi dua kelas, yaitu early-stage dan late-stage..
2. Model yang digunakan terbatas pada algoritma XGBoost dengan seleksi fitur (T-test, mutual information, chi-square, limma) dan teknik balancing (Random Oversampler, SMOTE, SMOTEENN dan BorderlineSMOTE).
3. Analisis difokuskan pada perbandingan ras kulit putih sebagai data training dan kulit hitam sebagai data testing serta interpretasi menggunakan LIME.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Membangun dan mengoptimalkan model klasifikasi stadium kanker prostat menggunakan algoritma XGBoost berbasis data ekspresi miRNA.
2. Model klasifikasi yang digunakan terbatas pada algoritma XGBoost dengan proses optimasi melalui metode seleksi fitur (T-test, Mutual Information, Chi-Square, dan Limma), teknik data balancing, serta hyperparameter tuning.
3. Analisis perbedaan ras dibatasi pada pasien berkulit putih sebagai data pelatihan dan pasien berkulit hitam sebagai data pengujian, dengan interpretasi hasil prediksi menggunakan LIME.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut.

1. Menyediakan pendekatan yang dapat dijadikan acuan dalam pengembangan model klasifikasi stadium kanker prostat berbasis data ekspresi miRNA, khususnya pada penelitian yang memanfaatkan algoritma *machine learning*.

2. Memberikan informasi mengenai pengaruh karakteristik data genomik dari kelompok ras yang berbeda terhadap kemampuan generalisasi model klasifikasi, sehingga dapat menjadi pertimbangan dalam pengembangan model yang lebih representatif.
3. Menyajikan hasil interpretasi prediksi berbasis LIME yang membantu menjelaskan kontribusi fitur miRNA terhadap keputusan model, sehingga meningkatkan transparansi dan kemudahan dalam memahami hasil klasifikasi.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

1. Bab 1 PENDAHULUAN

Bab ini menguraikan latar belakang penelitian yang melandasi pemilihan topik, identifikasi *research gap* dari penelitian-penelitian sebelumnya, serta permasalahan yang ingin diselesaikan. Selain itu, bab ini menjelaskan rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, dan sistematika penulisan skripsi.

2. Bab 2 LANDASAN TEORI

Bab ini membahas teori-teori yang menjadi dasar dan pendukung penelitian, meliputi kanker prostat, data miRNA, teknologi *Next Generation Sequencing* (NGS), sumber dan karakteristik data yang digunakan, metode seleksi fitur, algoritma XGBoost, metrik evaluasi model, serta konsep *Explainable Artificial Intelligence* (XAI). Teori-teori tersebut digunakan sebagai landasan dalam pengembangan dan evaluasi model klasifikasi pada penelitian ini.

3. Bab 3 METODOLOGI PENELITIAN

Bab ini menguraikan tahapan penelitian yang dilakukan, mulai dari pengumpulan data, pra-proses data, seleksi fitur, penyeimbangan data, pembangunan model menggunakan algoritma XGBoost, hingga proses evaluasi dan interpretasi model.

4. Bab 4 HASIL DAN DISKUSI

Bab ini menyajikan hasil implementasi penelitian, hasil pengujian model, analisis performa model berdasarkan berbagai skenario eksperimen, serta

pembahasan mengenai hasil yang diperoleh dan interpretasi model menggunakan metode LIME.

5. Bab 5 SIMPULAN DAN SARAN

Bab ini berisi simpulan yang diperoleh berdasarkan hasil penelitian yang telah dilakukan serta saran yang dapat digunakan untuk pengembangan penelitian lebih lanjut.

