BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Proses terbentuknya kanker atau dapat disebut karsinogenesis (*carcinogenesis*) merupakan sebuah proses kompleks yang dipengaruhi oleh banyak faktor, seperti predisposisi genetik (*genetic predispositions*) dan paparan lingkungan. Jumlah kematian akibat kanker terus meningkat setiap tahunnya, menjadikannya salah satu penyebab utama kematian di dunia. Meskipun tidak semua jenis kanker berujung pada kematian, penyakit ini tetap berdampak besar pada penurunan kualitas hidup dan menimbulkan beban biaya yang cukup besar bagi penderita. Pada tahun 2022, tercatat sebanyak 20 juta kasus kanker baru di seluruh dunia. Kanker Payudara merupakan salah satu jenis kanker yang paling umum didiagnosis dan menjadi salah satu jenis kanker yang menjadi penyebab kematian tertinggi dengan 2,3 juta jumlah kasus dan menyebabkan sekitar 685.000 angka kematian. Angka kejadian dan kematian akibat kanker payudara sangat bervariasi di berbagai wilayah dunia [1, 2].

Kanker payudara merupakan penyakit di mana sel-sel tertentu di jaringan payudara mengalami perubahan abnormal dan berkembang secara tidak terkendali hingga membentuk tumor. Pada tahap awal, kanker payudara biasanya tidak menimbulkan rasa sakit dan tidak menunjukkan gejala yang mencolok. Seiring perkembangannya, gejala yang dapat muncul meliputi benjolan atau penebalan di dalam atau sekitar payudara, perubahan ukuran atau bentuk payudara, keluarnya cairan dari puting, nyeri, atau puting tertarik ke dalam, serta iritasi kulit, perubahan tekstur seperti berlesung, kemerahan, atau bersisik. Namun, gejala-gejala ini juga bisa disebabkan oleh berbagai kondisi lain, sehingga kemunculannya tidak selalu berarti seseorang pasti menderita kanker payudara. Terdapat Tiga komponen struktural utama pada payudara, yaitu lobulus, duktus, dan jaringan lunak. Lobulus bertugas memproduksi susu, yang kemudian dialirkan ke puting melalui saluran duktus. Semua bagian tersebut dibungkus atau dihubungkan oleh jaringan ikat. Kanker payudara hampir selalu bermula di lobulus dan menyebar melalui duktus [3]. Terdapat berbagai faktor yang dapat memicu kemunculkan kanker payudara, salah satunya adalah ras dann etnis. Secara umum, angka kejadian kanker payudara tertinggi tercatat pada perempuan kulit putih non-

Hispanik. Sebaliknya, angka kematian akibat kanker ini secara signifikan lebih tinggi pada perempuan kulit hitam, yang juga memiliki tingkat kelangsungan hidup terendah dibanding kelompok etnis lainnya [2]. Terkait peningkatan angka kelangsungan hidup pada pasien dengan kanker stadium lanjut selama periode 1975-2013, tingkat kelangsungan hidup spesifik selama 5 tahun pada perempuan kulit putih non-Hispanik tercatat lebih tinggi dibandingkan kelompok etnis lain, terutama perempuan kulit hitam non-Hispanik. Perbedaan ini disebabkan oleh berbagai faktor yang saling berkaitan, termasuk predisposisi genetik, gaya hidup, serta faktor lingkungan lainnya [4].

Deteksi dini / diagnosis awal merupakan hal penting dalam pengobatan kanker payudara. Tumor stadium T1 yang berukuran kurang dari 2 cm memiliki tingkat kelangsungan hidup selama 10 tahun sekitar 85%, sedangkan tumor stadium T3 yang umumnya disebabkan oleh keterlambatan diagnosis memiliki tingkat kelangsungan hidup kurang dari 60% dalam periode yang sama [5]. Diagnosis kanker payudara melalui pemeriksaan fisik payudara, mammografi, USG payudara, MRI, serta modalitas pencitraan (*imaging modalities*) lainnya dapat membantu mengidentifikasi tumor dan kelainan jaringan. Terdapat berbagai teknik *imaging modalities* untuk *screening* kanker payudara, seperti *Mammography, Magnetic Resonance Imaging, Magnetic Resonance Spectroscopy, Breast Specific Gamma Imaging, Ultrasound* dan lain-lain. Setiap teknik memiliki keunggulanya masing-masing dan penggunaannya masing-masing teknik juga disesuaikan dengan kebutuhan. Disisi lain, teknik-teknik tersebut memiliki berbagai kelemahan, seperti dapat memberikan hasil positif palsu pada beberapa tumor jinak, dapat memberikan hasil negatif palsu pada tumor ganas, dan memberikan dosis radiasi yang tinggi [5].

Seiring dengan kemajuan teknologi, pendekatan diagnostik kanker payudara tidak lagi terbatas pada teknik pencitraan saja, melainkan telah berkembang ke arah analisis molekuler yang lebih maju. Perkembangan teknologi *Next-Generation Sequencing* (NGS) telah membuka peluang baru dalam pemahaman molekuler kanker payudara. Next-Generation Sequencing (NGS) merupakan teknologi revolusioner dalam sekuensing DNA dan RNA yang memungkinkan analisis ratusan hingga ribuan gen atau bahkan seluruh genom dalam waktu singkat. Teknologi ini menggabungkan keunggulan kimia sekuensing unik, matriks sekuensing yang beragam, dan bioinformatika untuk mencapai paralelisasi masif, sehingga memberikan informasi yang komprehensif untuk diagnosis, prognosis, dan terapi yang dipersonalisasi [6]. NGS telah menjadi tulang punggung dalam penelitian kanker, termasuk kanker payudara, karena kemampuannya

mengidentifikasi mutasi somatik, variasi genetik, dan perubahan epigenetik yang berperan dalam perkembangan kanker. Dengan pendekatan seperti whole-genome sequencing, whole-exome sequencing, atau targeted panel sequencing, NGS memungkinkan deteksi mutasi pada gen seperti BRCA1, BRCA2, dan TP53 yang terkait dengan risiko kanker payudara herediter maupun sporadis [6].

Penggunaan data RNA-seq dalam konteks *machine learning* memberikan peluang yang sangat menjanjikan untuk pengembangan model klasifikasi kanker payudara yang lebih akurat dan presisi. Data ekspresi gen yang dihasilkan dari RNA-seq mengandung ribuan fitur (gen) yang dapat dianalisis menggunakan algoritma *machine learning* untuk mengidentifikasi pola kompleks yang mungkin tidak terdeteksi melalui analisis konvensional. RNA Sequencing (RNA-Seq) adalah aplikasi NGS yang berfokus pada analisis transkriptom untuk mengukur ekspresi gen secara kuantitatif dan mendeteksi transkrip alternatif, fusi gen, serta varian splicing. Pada kanker payudara, RNA-Seq telah digunakan untuk mengkategorikan subtipe molekuler seperti luminal A, luminal B, HER2-positif, dan triple-negative breast cancer (TNBC), yang memiliki implikasi klinis dalam pemilihan terapi [6]. Teknologi ini juga memungkinkan identifikasi biomarker baru dan pemahaman mekanisme resistensi obat melalui analisis dinamika transkriptom sel tumor [7]. Selain itu, integrasi data RNA-Seq dengan machine learning telah memperluas potensi klasifikasi kanker yang lebih akurat dan prediksi respons terapi, membuka jalan untuk pendekatan precision medicine yang lebih canggih.

Penelitian terdahulu oleh Yu et al. melakukan klasifikasi subtibe kanker payudara berbasis data RNA-seq dari TCGA (*The Cancer Genome Atlas Program*). Dataset yang digunakan mencakup ekspresi gen dari lima subtipe BRCA (*Basal-like (192 samples), Her2 (82 samples), LumA (564 samples), LumB (207 samples), dan Normal-like (40 samples)*). Untuk *feature selection*, menggunakan *Differentially Expressed Genes* (DEG) memilih 1.000 gen dengan bobot tertinggi yang memiliki nilai *absolute logFC* $\geq 0,5$ dan *adjusted P value* $\leq 0,01$ [8]. Metode klasifikasi yang diuji meliputi Naive Bayes (nb), Random Forest (rf), dan SVM dengan kernel radial (svmRadial). Akurasi tertinggi yang didapatkan adalah 98,61% dalam memprediksi subtipe *Basal-like* dengan klasifikasi Random Forest.

Penelitian terdahulu oleh Zhang et al. mengklasifikasikan kanker payudara (*cancer vs non-cancer*) dengan melakukan hybrid feature selection dan melakukan klasifikasi menggunakan metode Support Vector Machine (SVM). Dataset yang digunakan adalah data RNA-seq dari TCGA dengan 1178 sampel (1080 kanker dan 98 sehat). Seleksi fitur awal menggunakan *edgeR package* dengan penyaringan

nilai *q-value* < 0.001. Metode seleksi fitur yang digunakan adalah *Support Vector Machine-Recursive Feature Elimination with Parameter Optimization* (SVM-RFE-PO), yang menggabungkan algoritma SVM-RFE dengan tiga teknik optimasi parameter: Grid Search (GS), Particle Swarm Optimization (PSO), dan Genetic Algorithm (GA) [9]. Hasil terbaik yang didapatkan adalah dengan SVM-RFE-PSO mencapai akurasi 91.68%.

Penelitian terdahulu oleh Kumar et al. melakukan prediksi dan klasifikasi kanker payudara dengan pendekatan *Stacking Ensembel Learning Model*. Dataset yang digunakan berasal dari Repositori data Breast Cancer Wisconsin dari UC Irvine (UCI) dengan total 569 sampel, sebanyak 357 sampel (62,7%) termasuk dalam kategori jinak, sedangkan 212 sampel (37,3%) tergolong ganas. Klasifikasi dilakukan dengan membandingkan beberapa algoritma seperti *k-Nearest Neighbor* (k-NN), *Random Forest*, *Logistic Regression*, *Support Vector Machine*, *Decision Tree*, dan *Stacking Ensemble Learning* [10]. Hasil akurasi tertinggi didapat dengan menggunakan *Stacking Ensembel Learning* yang mencapai 99,45%.

Penelitian-penelitian terdahulu tersebut menjadi dasar dilakukannya penelitian ini. Pada penelitian ini akan melakukan klasifikasi stadium kanker payudara (stage II dan stage III). Pendekatan klasifikasi stadium II dan III dilakukan sebagai pembeda / *research gap* pada penelitian terdahulu, sekaligus dikarenakan kedua stadium ini menunjukan perbedaan karakteristik pada kanker payudara. Dataset yang digunakan pada penelitian ini adalah dataset ekspresi gen RNA-seq (STAR - Counts) dan data fenotipe dari *The Cancer Genome Atlas Breast Cancer* (TCGA-BRCA) spesifik pada ras kulit putih *non-hispanic* untuk meminimalkan bias karena ekspresi gen dapat bervariasi antar kelompok ras/etnis [11]. Penelitian ini menerapkan kombinasi seleksi feature berbasis statistik dan bioinformatika (*Differentially Expressed Genes* (DEG)). Seleksi fitur DEG dilakukan dengan menggunakan metode Limma sebagai dan seleksi fitur berbasis statistik menggunakan *Logistic Regression*, *Analysis of Variance* (ANOVA), dan *Recursive Feature Elimination* (RFE). Seleksi fitur ini digunakan untuk meningkatkan performa model klasifikasi dan metode-metode fitur seleksi tersebut telah terbukti memberikan hasil yang baik dalam performa model pada penelitian terdahulu. Kemudian, menggunakan *Stacking Ensemble Learning* sebagai metode klasifikasinya dengan *base learner* yang terdiri dari *Random Forest*, *XGBoost*, *Support Vector Machine*, dan *Logistic Regression*. Lalu, dengan Logistic Regression sebagai *meta classifier*. *Stacking Ensemble* dapat meningkatkan akurasi, stabilitas, dan generalisasi model dengan cara menggabungkan kelebihan

dari beberapa algoritma klasifikasi (*base learner*) tersebut. Pendekatan tersebut juga dilakukan karena telah terbukti dalam mengatasi data yang kurang seimbang dan keterbatasan jumlah sampel.

Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam bidang bioinformatika, khususnya dalam upaya identifikasi biomarker potensial yang memiliki peran penting dalam membedakan stadium kanker payudara. Dengan memanfaatkan pendekatan machine learning yang dikembangkan, penelitian ini tidak hanya bertujuan untuk meningkatkan akurasi dalam proses klasifikasi stadium kanker payudara, akan tetapi juga membuka peluang baru dalam pengembangan metode diagnostik yang lebih canggih, akurat, dan bersifat terpersonalisasi. Harapannya, hasil dari penelitian ini dapat mendorong penerapan teknologi kecerdasan buatan dalam bidang kesehatan secara lebih luas di masa depan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana menentukan metode *feature selection* untuk menangani data RNA-seq dengan dimensionalitas tinggi dalam konteks klasifikasi kanker payudara?
2. Bagaimana performa algoritma Stacking Ensembel Learning dalam mengklasifikasikan tingkat stadium *(stage level)* kanker payudara berdasarkan data ekspresi gen?
3. Apa saja biomarker yang dapat diidentifikasi melalui proses seleksi fitur dan pembangunan model yang digunakan untuk klasifikasi kanker payudara?

1.3 Batasan Permasalahan

Agar mengarahkan penelitian agar lebih terfokus dan terukur sesuai dengan rumusan masalah, maka penelitian ini memiliki beberapa batasan sebagai berikut:

1. Dataset yang digunakan pada penelitian ini adalah dataset ekspresi gen RNA-seq (STAR - Counts) dari *The Cancer Genome Atlas Breast Cancer* (TCGA-BRCA).

2. Penelitian ini hanya menggunakan sampel dari pasien berlabel ras White non-Hispanic guna menjaga homogenitas data serta meminimalkan potensi bias yang mungkin timbul akibat perbedaan genetik antar kelompok etnis.
3. Klasifikasi stadium kanker payudara difokuskan pada labeling antara stadium II dan stadium III, sesuai dengan kriteria yang tersedia dalam dataset.
4. Metode seleksi fitur yang diterapkan dalam penelitian ini dibatasi pada kombinasi metode statistik (seperti ANOVA dan RFE) dan *Differentially Expressed Genes* (DEG).
5. Model klasifikasi hanya dibangun menggunakan algoritma machine learning tertentu, seperti *Random Forest*, *XGBoost*, dan *Stacking classifier*.

1.4 Tujuan Penelitian

Berikut merupakan tujuan penelitian yang ingin dicapai dari rumusan masalah yang telah dirancang sebelumnya:

1. Mengidentifikasi metode feature selection dalam menentukan data ekspresi RNA-seq yang relevan terhadap kanker payudara sehingga dapat meningkatkan akurasi klasifikasi kanker payudara.
2. Mengevaluasi performa algoritma Stacking Ensemble dalam mengklasifikasikan tingkat stadium kanker payudara berdasarkan ekspresi genetik, sehingga dapat diperoleh model prediktif yang optimal.
3. Mengidentifikasi biomarker potensial yang telah melalui proses fitur seleksi dan pembangunan model klasifikasi kanker payudara.

1.5 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan kontribusi baik secara teoritis maupun praktis dalam ranah bioinformatika dan deteksi kanker payudara. Beberapa manfaat yang dapat diperoleh dari penelitian ini antara lain sebagai berikut:

1.5.1 Manfaat Teoritis

1. Memberikan kontribusi ilmiah terhadap pengembangan pengetahuan di bidang bioinformatika, studi kanker payudara, melalui penerapan metode

seleksi fitur statistik dan pendekatan *Differentially Expressed Genes* (DEG).

2. Memberikan landasan ilmiah yang dapat dijadikan acuan dalam mengevaluasi efektivitas metode seleksi fitur dan algoritma klasifikasi untuk membangun model diagnosis kanker yang berbasis pada data RNA-seq.

1.5.2 Manfaat Praktis

1. Mengembangkan klasifikasi dengan pendekatan berbasis machine learning yang bertujuan untuk menghasilkan model klasifikasi yang memiliki akurasi dan efisiensi dalam proses diagnosis stadium II dan III pada kanker payudara.
2. Mengidentifikasi biomarker potensial dalam data ekspresi gen RNA seq pada kanker payudara.

1.6 Sistematika Penulisan

Berikut ini merupakan deskripsi singkat dalam sistematika penulisan laporan:

1. Bab 1 PENDAHULUAN
   Berisi latar belakang masalah, penelitian terdahulu yang mendukung penelitian ini, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian yang ingin dicapai.
2. Bab 2 LANDASAN TEORI
   Berisi landasan teori dan konsep-konsep yang mendukung penelitian seperti kanker payudara, ekspresi gen RNA-seq, *differentially expressed genes* (DEG), *Analysis of Variance*, *Recursive Feature Elimination* (RFE), *Random Forest*, *XGBoost*, *Support Vector Machine* (SVM), *Logistic Regression*, dan Stacking Assemble Learning.
3. Bab 3 METODOLOGI PENELITIAN
   Berisi penjelasan mengenai alur kerja penelitian, mulai dari pengumpulan data, pengolahan data, proses dan metode pemilihan fitur, model klasifikasi yang digunakan, serta berbagai skenario eksprimen dan evaluasi yang dilakukan.
4. Bab 4 HASIL DAN DISKUSI
   Berisi pemaparan hasil eksperimen, analisis performa model, serta

pembahasan terkait efektivitas metode yang diusulkan dalam diagnosis stadium II dan III pada kanker payudara.

5. Bab 5 KESIMPULAN DAN SARAN
   Berisi kesimpulan dari hasil penelitian yang telah dilakukan, serta saran untuk pengembangan lebih lanjut.