BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker merupakan salah satu penyebab utama kematian di dunia, dengan kanker payudara menjadi jenis paling umum pada wanita. Berdasarkan data dari *Global Cancer Observatory*, tercatat 2,3 juta kasus baru kanker payudara secara global, menjadikannya jenis kanker dengan jumlah kasus baru tertinggi kedua di dunia. Sementara itu, angka kematiannya mencapai 665.684 jiwa, menempatkan kanker payudara sebagai penyebab kematian akibat kanker tertinggi keempat secara global [1, 2]. Di Amerika Serikat, kanker payudara menyumbang 32% dari seluruh kasus kanker baru dan 21% dari kematian akibat kanker pada wanita [3]. Meskipun angka deteksi dini meningkat, disparitas rasial tetap menjadi tantangan. Wanita kulit putih cenderung menerima diagnosis lebih awal berkat akses layanan kesehatan yang lebih baik, sementara wanita kulit hitam memiliki risiko kematian lebih tinggi akibat keterlambatan diagnosis dan subtipe kanker yang lebih agresif [4, 5].

Skrining dini melalui mammografi masih menjadi metode utama dalam deteksi kanker payudara [6]. Teknologi ini mampu mendeteksi lesi sebelum munculnya gejala, namun memiliki keterbatasan signifikan seperti *overdiagnosis*, sensitivitas rendah pada jaringan padat, dan potensi *false positive* [7, 8]. Studi oleh Flemban mencatat bahwa *overdiagnosis* pada wanita usia 40 tahun ke atas akibat mammografi mencapai 12,6% [9]. Hal ini menandakan perlunya pendekatan pelengkap yang lebih akurat dan bersifat noninvasif.

Dalam beberapa tahun terakhir, pendekatan molekuler berbasis biomarker mulai dikembangkan. Biomarker seperti ER, PR, HER2, serta ekspresi *microRNA* (miRNA) seperti miR-21 dan miR-145 telah digunakan untuk klasifikasi subtipe kanker [10]. Namun, variasi antar individu dan laboratorium masih menjadi kendala dalam konsistensi hasil. Teknologi *next-generation sequencing* (NGS), seperti RNA-seq dan ekspresi miRNA, memiliki kemampuan untuk analisis ekspresi gen secara menyeluruh, namun data yang dihasilkan cenderung berdimensi tinggi dan tidak seimbang, sehingga sulit diolah secara konvensional [11].

Untuk mengatasi tantangan tersebut, integrasi bioinformatika dan *machine learning* mulai diterapkan. Analisis *Differentially Expressed Genes* (DEG)

digunakan untuk mengidentifikasi gen signifikan, sementara algoritma *machine learning* digunakan dalam klasifikasi stadium kanker. Kombinasi ini diperkuat oleh teknik seleksi fitur guna menyaring fitur relevan dari ribuan kandidat awal, sehingga dapat meningkatkan akurasi prediksi.

Beberapa penelitian sebelumnya telah mengimplementasikan pendekatan ini dengan berbagai variasi metode dan jenis data. Penelitian oleh Das dkk memanfaatkan pendekatan bioinformatika dan algoritma *machine learning* untuk mengklasifikasikan stadium kanker payudara berdasarkan ekspresi mikroRNA menggunakan dataset TCGA-BRCA. Dataset terdiri dari 1.224 sampel, yaitu 1.111 jaringan tumor dan 113 jaringan normal. Seleksi fitur dilakukan menggunakan metode DEG *limma* dengan ambang $|\log_2 \mathrm{FC}| > 1.0$ dan adj.p-value < 0.05. Beberapa algoritma seperti *Gaussian Naive Bayes*, *Random Forest* (RF), *Decision Tree*, *K-Nearest Neighbors*, *XGBoost*, dan *Support Vector Machine* (SVM) diuji dengan skema pembagian data 70% pelatihan dan 30% pengujian, serta menerapkan SMOTE untuk mengatasi ketidakseimbangan kelas. Dari tiga skenario klasifikasi stadium, model RF menunjukkan performa terbaik pada kategori kedua (stadium I, II–III, IV, V) dengan akurasi dan sensitivitas mencapai 97,19% [12].

Sementara itu, penelitian oleh Wu menggunakan pendekatan *machine learning* dengan dataset *Wisconsin Breast Cancer Diagnosis* (WBCD), yang terdiri atas 569 sampel pasien dan 32 indikator atau fitur. Studi ini bertujuan mengklasifikasikan kanker payudara menjadi dua kategori, yaitu *malignant* dan *benign*, dengan menggunakan dua algoritma utama, yakni *Support Vector Machine* (SVM) dan *Random Forest* (RF). Hasil evaluasi menunjukkan bahwa SVM memberikan performa lebih unggul dibandingkan RF, dengan akurasi sebesar 97% dan sensitivitas mencapai 98%, sedangkan RF mencatatkan akurasi 96% dan sensitivitas 95% [13].

Penelitian lanjutan oleh Naji dkk juga menggunakan dataset WBCD, namun dengan fokus pada 11 atribut fitur utama dari total 569 sampel, yang terdiri atas 357 kasus *benign* dan 212 *malignant*. Dalam penelitian ini, lima algoritma pembelajaran mesin diuji, yaitu SVM, RF, *Logistic Regression* (LogReg), *Decision Tree*, dan *K-Nearest Neighbors* (KNN), dengan skema pembagian data 75% untuk pelatihan dan 25% untuk pengujian. Hasil menunjukkan bahwa SVM tetap menjadi algoritma dengan performa terbaik, mencatat akurasi sebesar 97,2%. Disusul oleh RF dengan akurasi 96,5%, LogReg 95,8%, *Decision Tree* 95,1%, dan KNN 93,7%. Selain itu, pada klasifikasi kelas *malignant*, SVM juga mencatatkan nilai *precision* dan *F-measure* yang tinggi, masing-masing sebesar 0,98 [14].

Namun demikian, sebagian besar penelitian tersebut masih berfokus pada klasifikasi jenis kanker, bukan pada identifikasi stadium. Penelitian yang lebih mendalam terkait klasifikasi stadium dilakukan oleh Sathipati, yang memanfaatkan data ekspresi miRNA dari platform *Illumina HiSeq 2000* yang bersumber dari TCGA. Dataset terdiri atas 503 profil dari 386 pasien kanker payudara, dengan 193 sampel stadium awal (Stadium I dan II) dan 193 stadium lanjut (Stadium III dan IV). Model klasifikasi yang dikembangkan, yaitu *SVM-BRC*, menggunakan metode seleksi fitur *Inheritable Bi-objective Combinatorial Genetic Algorithm* (IBCGA) dan divalidasi dengan *10-fold cross-validation*. Hasil terbaik diperoleh pada model *SVM-BRC-Best* dengan akurasi 83,16% dan sensitivitas 0,84. Model *SVM-BRC-Mean* menunjukkan akurasi 80,38% dan sensitivitas 0,79. Sebagai perbandingan, metode RF hanya mencapai akurasi 66,83% dan sensitivitas 0,66%. Kinerja lebih rendah tercatat pada *Multilayer Perceptron* 57,25%, *Sequential Minimal Optimization* 62,69%, serta *Naïve Bayes* 64,50% dengan sensitivitas 0,63. Model dengan performa terendah adalah *Decision Tree*, yang hanya mencatat akurasi 50,25% dan sensitivitas 0,50 [15].

Berbeda dengan sebagian besar penelitian terdahulu yang berfokus pada klasifikasi kanker *benign* dan *malignant*, penelitian ini menitikberatkan pada klasifikasi stadium kanker payudara, yang memiliki signifikansi klinis tinggi karena berkaitan langsung dengan penentuan prognosis pasien. Penelitian ini menerapkan dua skenario utama klasifikasi, yaitu Stadium I vs. III (*localized cancer* vs. *locally advanced cancer*) dan Stadium II vs. III (*early stage* vs. *late stage*). Pemilihan skenario ini didasarkan tidak hanya pada relevansi biologis antar stadium yang signifikan secara klinis [16], tetapi juga mempertimbangkan distribusi data yang memungkinkan pembentukan kelompok dengan representasi yang seimbang. Untuk meningkatkan kualitas klasifikasi, dilakukan kombinasi beberapa pendekatan seleksi fitur, yaitu *Differentially Expressed Genes* (DEG) dengan metode LIMMA, *Analysis of Variance* (ANOVA), serta *Recursive Feature Elimination* (RFE). Dua algoritma klasifikasi digunakan, yakni *Support Vector Machine* (SVM) dan *Logistic Regression* (LogReg) dengan regularisasi L2, yang sebelumnya telah terbukti efektif dalam studi serupa.

Pendekatan ini dirancang untuk mengatasi tantangan dalam analisis data biologis, yaitu ketidakseimbangan kelas, kompleksitas sinyal genetik, serta keterbatasan jumlah sampel. Selain itu, penelitian ini juga membandingkan performa dua jenis data, yaitu ekspresi gen (RNA-seq) dan ekspresi miRNA (stem-loop) dalam membedakan stadium kanker payudara, dengan fokus pada sampel dari

individu berkulit putih guna meminimalkan potensi bias genetik dalam interpretasi hasil.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka permasalahan utama dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana perbandingan efektivitas klasifikasi antara dataset Stadium I vs. III dan Stadium II vs. III dalam mengenali pola fitur stadium kanker payudara, berdasarkan metrik evaluasi akurasi, presisi, *recall*, F1-score, dan durasi komputasi?
2. Apakah data ekspresi gen RNA-seq memiliki performa lebih baik dibandingkan data ekspresi stem-loop miRNA dalam klasifikasi stadium kanker payudara, berdasarkan metrik evaluasi akurasi dan F1-score?
3. Apa saja biomarker signifikan yang dapat diidentifikasi melalui seleksi fitur dan model yang digunakan untuk mendeteksi stadium kanker payudara stadium awal dan stadium lanjut pada wanita ras kulit putih?

1.3 Batasan Permasalahan

Agar penelitian lebih terfokus dan sesuai dengan tujuan yang ingin dicapai, maka ruang lingkup permasalahan dibatasi pada:

1. Data dalam penelitian ini berasal dari dataset publik *The Cancer Genome Atlas* (TCGA) melalui platform *Xena Browser*, dengan pemilihan sampel berdasarkan ras mayoritas, yaitu kulit putih.
2. Data yang digunakan terbatas pada ekspresi gen (RNA-seq) dan ekspresi miRNA (*stem-loop expression*).
3. Jumlah fitur yang digunakan dibatasi maksimal 50 fitur untuk tiap skenario klasifikasi.
4. Penelitian ini tidak mencakup pembahasan mengenai aspek klinis lanjutan, seperti efektivitas terapi maupun tingkat kelangsungan hidup pasien.

1.4 Tujuan Penelitian

Berdasarkan rumusan dan batasan masalah di atas, penelitian ini bertujuan untuk:

1. Membandingkan efektivitas klasifikasi antara dataset Stadium I vs. III dan Stadium II vs. III dalam mengenali pola fitur stadium kanker payudara, berdasarkan metrik evaluasi akurasi, presisi, *recall*, F1-score, dan durasi komputasi.
2. Mengevaluasi performa klasifikasi stadium kanker payudara menggunakan data ekspresi gen RNA-seq dibandingkan dengan data ekspresi stem-loop miRNA, berdasarkan metrik evaluasi akurasi dan F1-score.
3. Mengidentifikasi biomarker signifikan yang berkontribusi terhadap klasifikasi stadium kanker payudara stadium awal dan stadium lanjut, melalui proses seleksi fitur dan penerapan model klasifikasi pada wanita ras kulit putih.

1.5 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan wawasan ilmiah mengenai efektivitas klasifikasi dua skenario stadium kanker payudara, yaitu Stadium I vs. III dan Stadium II vs. III, berdasarkan metrik evaluasi seperti akurasi, presisi, *recall*, F1-score, dan durasi komputasi, sehingga dapat membantu pemilihan strategi klasifikasi yang lebih optimal.
2. Menyediakan bukti empiris mengenai keunggulan relatif antara data ekspresi gen RNA-seq dan data ekspresi stem-loop miRNA dalam klasifikasi stadium kanker payudara, khususnya dilihat dari akurasi dan F1-score, yang dapat menjadi dasar dalam pemilihan tipe data biologis untuk analisis serupa.
3. Menghasilkan daftar biomarker signifikan yang diidentifikasi melalui metode seleksi fitur dan algoritma klasifikasi, yang berpotensi menjadi kandidat biomarker baru untuk deteksi stadium kanker payudara stadium awal dan lanjut, khususnya pada populasi wanita ras kulit putih.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam laporan ini adalah sebagai berikut:

- **Bab 1 Pendahuluan**
  Pada bab ini memaparkan latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

- **Bab 2 Landasan Teori**
  Pada bab ini diuraikan teori dan algoritma yang mendukung penelitian, meliputi penjelasan mengenai kanker payudara, ekspresi gen (RNA-seq), ekspresi miRNA (stem-loop), *differentially expressed genes* (DEG), *feature selection*, *Analysis of Variance* (ANOVA), *Recursive Feature Elimination* (RFE), *Logistic Regression* dengan L2 *Ridge Regularization*, *Support Vector Machine* (SVM), serta *confusion matrix*.

- **Bab 3 Metodologi Penelitian**
  Pada bab ini menjelaskan langkah-langkah penelitian, spesifikasi perangkat, studi literatur, pengumpulan data, praproses data, seleksi fitur, pembangunan model, evaluasi model, skenario eksperimen, dan interpretasi biomarker.

- **Bab 4 Hasil dan Diskusi**
  Bab ini menyajikan hasil dan analisis dari seluruh tahapan penelitian, dimulai dari pemaparan karakteristik dataset, proses seleksi fitur, hingga evaluasi performa model pada berbagai skenario klasifikasi stadium kanker payudara. Pembahasan mencakup perbandingan efektivitas antar skenario stadium, jenis data ekspresi RNA-seq dan miRNA, serta pendekatan seleksi fitur yang digunakan. Di bagian akhir, dilakukan identifikasi serta interpretasi biologis terhadap biomarker signifikan yang diperoleh dari model terbaik.

- **Bab 5 Kesimpulan dan Saran**
  Pada bab ini berisi simpulan utama yang menjawab rumusan masalah, mencakup keunggulan data ekspresi gen atau miRNA, efektivitas masing-masing skenario stadium, serta potensi biomarker. Juga memuat saran untuk pengembangan metode dan penelitian di masa depan.