BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker menjadi salah satu penyebab utama kematian di dunia, dengan hampir 10 juta kematian pada tahun 2020. Di antara berbagai jenis kanker, kanker payudara merupakan penyebab kematian terbanyak pada perempuan, dengan 2,26 juta kasus secara global pada 2020 [1]. Di Indonesia, prevalensinya mencapai 16,7%, dengan Yogyakarta memiliki angka tertinggi sebesar 2,4% [2]. Sayangnya, sebagian besar kasus di Indonesia terdeteksi pada stadium lanjut, berbeda dengan negara maju yang mendeteksinya sejak dini. Keterlambatan deteksi disebabkan oleh kurangnya kesadaran masyarakat, keterbatasan fasilitas, serta hambatan ekonomi dan geografis [3, 4].

Gaya hidup juga menjadi faktor risiko utama, termasuk obesitas, merokok, konsumsi alkohol, dan jenis pekerjaan tertentu [5]. Selain itu, ketidaksetaraan akses layanan kesehatan turut memperburuk kondisi, terutama bagi perempuan dari kelompok etnis minoritas dan berpenghasilan rendah [6, 7]. Perempuan Afrika-Amerika dan Hispanik lebih sering tidak memiliki asuransi kesehatan dibandingkan perempuan Kaukasia, sehingga kurang melakukan skrining kanker payudara [6]. Perempuan dari kaum kurang mampu kurang terlayani dengan baik, dan sering kali tidak dapat mengambil cuti dari pekerjaan untuk pemeriksaan medis [7]. Dikarenakan beberapa problematik tersebut, hubungan dengan ras atau etnis tentu memiliki hubungan satu sama lain. Pada tahun 2020, angka kematian akibat kanker payudara menunjukkan variasi berdasarkan ras atau etnis. Wanita *non-Hispanic Black* memiliki angka kematian tertinggi yaitu 26,4 per 100.000, diikuti oleh *non-Hispanic White* (19,4), *non-Hispanic American Indian/Alaska Native* (13,7), *Hispanic* (13,1), dan yang terendah *non-Hispanic Asian/Pacific Islander* (11,4) [8].

Metode deteksi kanker payudara saat ini meliputi mamografi dan MRI. Mamografi efektif untuk skrining awal dan relatif terjangkau, namun kurang sensitif pada jaringan payudara padat dan sering menimbulkan ketidaknyamanan [9, 10]. Sementara MRI menawarkan akurasi lebih tinggi, namun biayanya mahal, terbatas ketersediaannya, dan berisiko menyebabkan *overdiagnosis* [11]. Kemajuan dalam teknologi molekuler seperti *Next-Generation Sequencing* (NGS) membuka peluang baru untuk mendeteksi kanker secara lebih akurat. Panel NGS memungkinkan

analisis ekspresi gen secara mendalam, termasuk mutasi gen BRCA yang berperan dalam kanker payudara herediter [12]. Namun, biaya tinggi menjadi kendala; misalnya, panel FoundationOne CDx mematok harga $5.800 untuk 324 fitur, atau sekitar $17,90 per fitur [13].

MicroRNA (miRNA) merupakan molekul *Ribo Nucleic Acid* (RNA) kecil yang berperan penting dalam regulasi ekspresi gen. MiRNA bertindak sebagai onkogen (pemicu kanker) atau penekan tumor, hal ini tergantung pada fungsi target yang diaturnya [14]. MiRNA juga telah diidentifikasi sebagai biomarker potensial. Misalnya, *miR-21* menunjukkan ekspresi tinggi pada kanker payudara tipe Luminal A dan terkait dengan proliferasi tumor serta resistensi terhadap terapi hormon [15]. Stabilitas miRNA menjadikannya kandidat kuat untuk biomarker non-invasif [16].

Penelitian dalam bidang bioinformatika terus berkembang, khususnya dalam klasifikasi kanker payudara berbasis ekspresi gen dan miRNA. Peneliti pertama menggunakan data *miRNA-seq gene expression* dari *TCGA-BRCA* untuk klasifikasi antara *Primary Solid Tumor* (kode '01') dan *Solid Tissue Normal* (kode '11'), dengan 849 sampel dan 897 fitur awal. Seleksi fitur dilakukan menggunakan *Logistic Regression - Recursive Feature Elimination* (LR-RFE) dan *Chi-Squared Test*, sementara algoritma klasifikasi yang digunakan adalah *Support Vector Machine* (SVM). Penelitian ini juga menguji dua model *Deep Learning* (DL), yaitu *Sparse Autoencoder* (SAE) dan *Deep Belief Network* (DBN), baik dengan maupun tanpa pendekatan *Principal Component Analysis* (PCA). Kombinasi terbaik diperoleh dari pipeline SAE + LR-RFE + SVM tanpa PCA, dengan akurasi sebesar 97,46% menggunakan 30 fitur. Jika metode seleksi fitur diganti menjadi *Chi-Squared Test*, akurasi menurun menjadi 91,81% [17].

Sementara itu, peneliti kedua menggunakan data *gene expression* dari *TCGA-BRCA* sebanyak 1.224 sampel untuk melakukan klasifikasi *staging* kanker payudara. Seleksi fitur dilakukan menggunakan metode *Differentially Expressed Genes* (DEG) berbasis *Limma*, dengan kriteria $|\log_2 \text{FC}| > 1{,}0$ dan *adjusted p-value* $< 0{,}05$, sehingga diperoleh 20 gen signifikan (10 *upregulated* dan 10 *downregulated*). Ketidakseimbangan kelas diatasi dengan pendekatan *Synthetic Minority Oversampling Technique* (SMOTE). Proses klasifikasi dilakukan dalam tiga skema pembagian kelas stadium. Hasil terbaik diperoleh pada skema kedua (kelas 'I', 'II–III', 'IV', dan 'V') dengan menggunakan algoritma *Random Forest* (RF), menghasilkan akurasi 97,19%, presisi 97,20%, *recall* 97,19%, *F1-score* 97,18, dan spesifisitas 92,88% [18].

Penelitian lain oleh peneliti ketiga menggabungkan data *TCGA RNA-Seq*

dan *METABRIC Microarray* dengan total 1.793 sampel. Seleksi fitur dilakukan menggunakan pendekatan hibrida antara *minimum Redundancy Maximum Relevance* (mRMR) dan *Support Vector Machine–Recursive Feature Elimination* (SVM-RFE), menghasilkan 100 fitur terpilih. Teknik *SMOTE* digunakan untuk penyeimbangan kelas. Beberapa arsitektur *deep learning* diterapkan, termasuk *Deep Neural Network* (DNN), *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN), *modified DNN*, dan *AutoKeras*. Model DNN memberikan hasil terbaik untuk klasifikasi multi-kelas dengan akurasi 53,1% dan nilai *Cohen's Kappa* 0,303, sedangkan untuk klasifikasi biner (*early stage* vs *advanced stage*), *modified DNN* mencapai akurasi 81,0% dengan *Cohen's Kappa* sebesar 0,280 [19].

Penelitian ini berfokus pada klasifikasi stadium 2 dan 3 kanker payudara karena keduanya merepresentasikan fase transisi dari kanker lokal ke lokal lanjut yang krusial secara klinis. Stadium ini menunjukkan perbedaan signifikan dalam ukuran tumor, keterlibatan kelenjar getah bening, dan kompleksitas terapi, serta masih berada dalam rentang kuratif meskipun membutuhkan pendekatan *multimodal* yang lebih agresif. Berbeda dengan stadium 1 yang memiliki prognosis baik dan stadium 4 yang bersifat metastatik dan heterogen, pemfokusan pada stadium 2 dan 3 menghindari bias klasifikasi ekstrem dan lebih relevan secara klinis dalam pengambilan keputusan pengobatan [20]. Tujuan utama dari penelitian ini adalah untuk menemukan jumlah fitur paling minimal yang tetap mampu menghasilkan akurasi klasifikasi yang tinggi. Mengingat tingginya biaya pemeriksaan panel NGS, diperlukan metode klasifikasi yang efisien dan tetap optimal meskipun hanya menggunakan sedikit fitur.

Penelitian akan menggunakan data ekspresi gen dan miRNA dari TCGA-BRCA, dan mengimplementasikan metode bioinformatika DEG dengan *Limma*, tanpa *balancing* data seperti menggunakan *SMOTE*, seleksi fitur LR-RFE, serta model klasifikasi SVM, *Random Forest*, *Logistic Regression* (LR), dan *Soft Voting Classifier*. Dengan menggunakan metode seleksi fitur seperti DEG dan LR-RFE, serta model klasifikasi *Soft Voting* yang menggabungkan kekuatan beberapa algoritma, penelitian ini bertujuan untuk membangun sistem prediksi stadium kanker payudara yang lebih efisien. Fokus pada pasien berkulit putih dilakukan untuk meningkatkan homogenitas data, sehingga model yang dihasilkan dapat memiliki presisi dan validitas yang lebih tinggi. Sebagian besar studi sebelumnya tidak secara eksplisit mempertimbangkan faktor homogenitas ras, dari teori perbedaan etnis dapat memengaruhi pola ekspresi gen secara signifikan. Pendekatan ini mampu menghasilkan model klasifikasi yang akurat dan efisien

dari segi biaya, sehingga dapat menjadi solusi yang lebih terjangkau bagi layanan kesehatan. Selain itu, pendekatan ini juga mendukung pengembangan sistem diagnosis kanker payudara yang lebih personal dan presisi, dengan mempertimbangkan karakteristik biologis individu serta keterbatasan sumber daya di berbagai fasilitas medis.

1.2 Rumusan Masalah

Berdasarkan latar belakang, berikut merupakan rumusan masalah pada penelitian ini:

1. Apakah kombinasi metode *Feature Selection* berbasis statistik, dan *Differentially Expressed Genes* (DEG) mampu meningkatkan akurasi klasifikasi *early dan advanced stage* kanker payudara?
2. Bagaimana performa algoritma klasifikasi seperti *Support Vector Machine* (SVM), *Random Forest* (RF), *Logistic Regression* (LR), dan *Soft Voting Classifier* dalam membedakan stadium awal dan lanjut kanker payudara, ditinjau dari akurasi, harmonisasi metrik evaluasi (*precision*, *recall*, dan *F1-score* makro), serta efisiensi waktu eksekusi?
3. Berapa jumlah fitur optimal yang dapat digunakan untuk menjaga akurasi klasifikasi tinggi, namun tetap efisien secara biaya untuk implementasi deteksi kanker payudara?
4. Apa saja kombinasi *biomarker* signifikan yang diperoleh dari data ekspresi gen dan *miRNA* yang mampu membedakan kanker payudara stadium awal (*early stage*) dan stadium lanjut (*advanced stage*) secara akurat dan konsisten?

1.3 Batasan Permasalahan

Terdapat beberapa batasan masalah yang ditetapkan pada penelitian ini, dengan fokus menyelesaikan masalah utama:

1. Penelitian ini hanya menggunakan data dari *The Cancer Genome Atlas* (TCGA) untuk kasus *Breast Invasive Carcinoma* (BRCA), dengan jenis data berupa ekspresi gen (*Gene Expression RNA-seq*) dan ekspresi miRNA

(*miRNA Expression Quantification*). Informasi tahapan kanker (staging) diperoleh dari data fenotipe yang tersedia melalui platform UCSC Xena.

2. Klasifikasi kanker payudara dibatasi menjadi dua kelompok, yaitu *Early Stage* Stage 2 dan *Advanced Stage* Stage 3 berdasarkan label *ajcc_pathologic_stage*. Selain itu, penelitian ini hanya menggunakan sampel dengan label ras *white*, agar hasil yang diperoleh lebih homogen dan mengurangi potensi bias yang muncul akibat perbedaan etnisitas dalam profil genetik.

3. Penelitian ini tidak menerapkan teknik penyeimbangan data (*data balancing*) seperti SMOTE atau metode sintetik lainnya. Hal ini dilakukan untuk menjaga kemurnian distribusi data asli dan menghindari potensi bias yang dapat muncul akibat manipulasi data secara artifisial.

4. Penelitian ini hanya menggunakan metode *Limma* untuk *Differentially Expressed Genes* (DEG), mengingat dataset yang digunakan telah melalui proses normalisasi logaritmik $\log_2(\text{FPKM-UQ} + 1)$. Oleh karena itu, metode berbasis count seperti *DESeq2* atau *edgeR* tidak diterapkan dalam penelitian ini guna menjaga konsistensi dan kesesuaian asumsi metode statistik yang digunakan.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini berdasarkan rumusan masalah adalah:

1. Mengevaluasi efektivitas kombinasi metode *Feature Selection* berbasis statistik dan filtrasi *Differentially Expressed Genes* (DEG) dalam meningkatkan akurasi klasifikasi stadium kanker payudara.

2. Melakukan analisis performa sejumlah algoritma klasifikasi, seperti *Support Vector Machine* (SVM), *Random Forest* (RF), *Logistic Regression* (LR), dan *Soft Voting Classifier*, dalam membedakan kanker payudara stadium awal dan lanjut, berdasarkan akurasi, harmonisasi metrik evaluasi (*precision*, *recall*, dan *F1-score* makro), serta efisiensi waktu eksekusi masing-masing algoritma.

3. Menentukan jumlah fitur optimal yang mampu mempertahankan akurasi klasifikasi yang tinggi, dan efisiensi biaya untuk keperluan implementasi klinis dalam deteksi stadium kanker payudara.

4. Mengidentifikasi kombinasi *biomarker* signifikan dari data ekspresi gen dan *miRNA* yang memiliki kemampuan untuk membedakan kanker payudara stadium awal dan stadium lanjut secara akurat dan konsisten.

1.5 Manfaat Penelitian

Manfaat yang diberikan penelitian ini dibagi menjadi tiga kategori yaitu, manfaat teoritis, manfaat praktis, dan manfaat pengembangan metode:

1. **Manfaat Teoritis** Penelitian ini diharapkan dapat menambah wawasan dalam bidang bioinformatika, khususnya terkait pemilihan fitur (*Feature Selection*) menggunakan pendekatan *Differentially Expressed Genes* (DEG), serta memberikan kontribusi dalam pengembangan metode klasifikasi kanker payudara berdasarkan tahap perkembangannya.

2. **Manfaat Praktis** Hasil penelitian ini diharapkan dapat digunakan sebagai dasar dalam pengembangan sistem pendukung keputusan medis yang lebih akurat dan efisien untuk membantu dokter atau peneliti dalam mendeteksi kanker payudara dan menentukan tahap perkembangannya (*Early Stage* dan *Advanced Stage*) dengan lebih cepat dan tepat.

3. **Manfaat Pengembangan Metode** Penelitian ini diharapkan dapat membuka peluang pengembangan metode diagnostik yang lebih presisi dan personal dengan pemanfaatan kombinasi biomarker yang teridentifikasi, sehingga memungkinkan deteksi dini kanker payudara dan mendukung pengembangan terapi yang lebih tepat sasaran.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini berisi latar belakang penelitian, rumusan masalah, tujuan penelitian, batasan masalah, serta manfaat penelitian.

- Bab 2 LANDASAN TEORI
  Bab ini berisi landasan teori dan penelitian terdahulu yang mendukung penelitian ini. Penjelasan mencakup konsep dasar bioinformatika, metode *Feature Selection* serta algoritma klasifikasi. Selain itu, bab ini juga

membahas lebih dalam mengenai *Differentially Expressed Genes* (DEG), termasuk bagaimana DEG digunakan dalam mengidentifikasi biomarker potensial yang berperan dalam membedakan *Early Stage* dan *Advanced Stage* pada kanker payudara.

- Bab 3 METODOLOGI PENELITIAN
  Bab ini menjelaskan tahapan penelitian, mulai dari pengolahan data, metode pemilihan fitur, hingga model klasifikasi yang digunakan. Selain itu, dijelaskan pula rancangan eksperimen dan evaluasi model.
- Bab 4 HASIL DAN DISKUSI
  Bab ini berisi pemaparan hasil eksperimen, analisis performa model, serta pembahasan terkait efektivitas metode yang diusulkan dalam membedakan *Early Stage* dan *Advanced Stage* kanker payudara.
- Bab 5 KESIMPULAN DAN SARAN
  Bab ini berisi kesimpulan dari hasil penelitian serta saran untuk pengembangan lebih lanjut agar metode yang dikembangkan bisa diimplementasikan lebih luas.