BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kanker payudara merupakan salah satu penyakit dengan prevalensi dan tingkat mortalitas tertinggi pada perempuan secara global. Berdasarkan data World Health Organization (WHO) tahun 2023, kanker payudara menyumbang sekitar 12% dari seluruh kasus keganasan baru dan menjadi penyebab utama kematian akibat kanker pada wanita. Diestimasi terdapat lebih dari 2,3 juta kasus baru dengan mortalitas mencapai 685.000 jiwa setiap tahunnya [1]. Penyakit ini bersifat multifaktorial, dipengaruhi oleh interaksi kompleks antara faktor genetik, hormonal, dan lingkungan yang memicu mutasi serta proliferasi sel abnormal pada jaringan payudara [2]. Deteksi dini menjadi faktor krusial dalam meningkatkan prognosis dan efektivitas terapi. Namun, modalitas diagnostik konvensional seperti mammografi dan biopsi masih memiliki keterbatasan, terutama dalam mendeteksi kanker stadium awal pada jaringan payudara yang padat (*dense breast tissue*), sehingga urgensi terhadap inovasi metode diagnostik yang presisi semakin meningkat [2].

Perkembangan pesat teknologi kecerdasan buatan (*Artificial Intelligence*) dan pembelajaran mesin (*Machine Learning*) menawarkan paradigma baru dalam akurasi diagnosis onkologi. AI memiliki kapabilitas untuk mengekstraksi pola *non-linear* yang kompleks dari data medis berskala besar yang seringkali sulit diinterpretasikan secara manual [3]. Berbagai algoritma pembelajaran mesin seperti *Support Vector Machine* (SVM), *k-Nearest Neighbor* (KNN), dan *Decision Tree* telah terbukti efektif dalam klasifikasi kanker payudara dengan akurasi tinggi, termasuk pada dataset berskala besar seperti TCGA (The Cancer Genome Atlas) [4].

Di sisi lain, pendekatan radiomik yang memanfaatkan ekstraksi fitur kuantitatif dari citra medis, seperti MRI dan mammografi, juga telah berkembang sebagai modalitas non-invasif untuk deteksi kanker. Namun, efektivitas radiomik dalam melakukan stratifikasi stadium (*staging*) kanker payudara secara presisi masih menghadapi tantangan signifikan. Pendekatan ini cenderung terbatas pada representasi fenotipe makroskopis—seperti morfologi, volumetri, dan tekstur tumor—yang tidak selalu mampu memotret heterogenitas intratumoral di tingkat seluler maupun molekuler [5]. Mengingat progresi stadium kanker sangat dipengaruhi oleh dinamika perubahan genetik dan epigenetik yang kompleks, informasi visual dari radiologi sering kali tidak cukup sensitif untuk menangkap perubahan biologis tersebut tanpa validasi molekuler yang kuat [6]. Selain itu, isu terkait reproduksibilitas akibat variabilitas parameter akuisisi citra antar-perangkat dan subjektivitas dalam segmentasi *Region of Interest* (ROI) turut membatasi konsistensi hasil diagnosis, sehingga menyulitkan generalisasi model pada data klinis yang beragam [7]. Oleh karena itu, pendekatan berbasis data molekuler menjadi alternatif krusial untuk mendapatkan gambaran patogenesis yang lebih komprehensif dan akurat dalam penentuan stadium kanker.

Dalam konteks molekuler tersebut, data ekspresi gen (*gene expression*) memegang peranan vital untuk diagnosis dan klasifikasi. Profil ekspresi gen merefleksikan aktivitas biologis yang mendasari perkembangan kanker, memungkinkan identifikasi subtipe kanker secara lebih akurat dibandingkan fitur klinis semata [8]. Kendati demikian, data genomik memiliki karakteristik intrinsik *high dimensional–low sample size* (HDLSS), di mana jumlah fitur (gen) jauh melampaui

jumlah sampel pasien. Tanpa strategi reduksi dimensi yang tepat, model klasifikasi rentan terhadap *overfitting* dan tingginya beban komputasi, yang pada akhirnya menurunkan generalisasi model [8], [9].

Sejumlah penelitian terdahulu telah mengimplementasikan metode seleksi fitur konvensional seperti *Principal Component Analysis* (PCA), ReliefF, atau LASSO untuk menangani dimensi tinggi tersebut [10]. Namun, pendekatan-pendekatan ini sering kali kurang optimal dalam mempreservasi informasi diskriminatif antar kelas, terutama pada data biologis yang mengandung *noise* tinggi dan ketidakpastian batas kelas (*class overlaps*). Metode konvensional cenderung berfokus pada variansi data atau korelasi linear, sehingga berisiko mengeliminasi gen yang memiliki signifikansi biologis namun variansinya rendah. Untuk mengatasi limitasi tersebut, pendekatan berbasis logika *fuzzy* seperti *Discriminant Fuzzy Pattern* (DFP) hadir sebagai solusi yang lebih adaptif. DFP dirancang untuk menangani ketidakpastian (*fuzziness*) pada data ekspresi gen dengan mengukur derajat keanggotaan fitur terhadap kelas tertentu, sehingga mampu menyeleksi gen yang paling relevan dan diskriminatif dalam membedakan stadium kanker secara lebih *robust* [11].

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada pengembangan model deteksi kanker payudara untuk membedakan stadium awal dan lanjut dengan memanfaatkan data ekspresi gen. Guna mengatasi tantangan karakteristik data HDLSS, penelitian ini mengusulkan integrasi metode seleksi fitur DFP dan klasifikasi SVM. DFP diterapkan untuk mereduksi dimensi sekaligus mengisolasi fitur gen yang paling diskriminatif terhadap tahapan kanker, sementara SVM dipilih karena keandalannya dalam menangani ruang fitur berdimensi tinggi dengan margin pemisah yang optimal. Pendekatan hibrida ini diharapkan mampu meningkatkan akurasi, sensitivitas, dan spesifisitas diagnosis, serta memberikan kontribusi teoretis maupun praktis dalam pengembangan sistem pendukung keputusan medis berbasis bioinformatika.

1.2 Rumusan Masalah

Berdasarkan pemaparan latar belakang masalah di atas, teridentifikasi bahwa penggunaan data ekspresi gen untuk diagnosis kanker payudara memiliki tantangan tersendiri, terutama terkait karakteristik data berdimensi tinggi dengan jumlah sampel yang terbatas. Oleh karena itu, permasalahan utama yang dikaji dalam penelitian ini dirumuskan dalam pertanyaan penelitian sebagai berikut:

1. Bagaimana metode *feature selection* pada data ekspresi gen menggunakan DFP dapat diterapkan untuk memperoleh fitur yang paling relevan dalam diagnosis kanker payudara?
2. Bagaimana penerapan algoritma SVM pada data ekspresi gen dapat meningkatkan kemampuan sistem dalam mendeteksi serta mengklasifikasikan kanker payudara?

1.3 Batasan Permasalahan

Mengingat luasnya domain permasalahan dalam diagnosis kanker dan analisis bioinformatika, penelitian ini perlu dibatasi ruang lingkupnya agar pembahasan tetap terarah dan fokus pada tujuan yang ingin dicapai. Pembatasan ini juga dilakukan untuk menjaga kedalaman

analisis serta efektivitas model yang dikembangkan. Adapun batasan-batasan masalah yang ditetapkan dalam penelitian ini adalah sebagai berikut:

1. Data Penelitian

   Penelitian ini menggunakan dataset publik *The Cancer Genome Atlas – Breast Invasive Carcinoma* (TCGA-BRCA) yang diakses melalui portal UCSC Xena Browser, yang terdiri dari kombinasi data ekspresi gen dan data fenotipe klinis. Data ekspresi gen berformat RNA-seq (IlluminaHiSeq) yang telah dinormalisasi ke dalam satuan *log2(x+1)* dari nilai *Transcripts Per Million* (TPM) digunakan sebagai fitur input utama. Sementara itu, data fenotipe klinis yang mencakup informasi stadium patologis (*pathologic stage*) dan karakteristik pasien digunakan sebagai label referensi (*ground truth*) dalam penentuan kelas. Penelitian ini tidak menggunakan data citra medis (seperti MRI, CT-Scan, atau Mammografi) maupun atribut klinis lain yang tidak relevan dengan proses klasifikasi stadium.

2. Metode Seleksi Fitur

   Proses reduksi dimensi dan seleksi fitur difokuskan pada penggunaan metode DFP. Metode ini digunakan untuk menangani karakteristik data *high-dimensional low-sample size* (HDLSS) serta menyeleksi gen yang paling relevan dan diskriminatif dalam membedakan stadium kanker. Metode seleksi fitur lain seperti PCA, LASSO, atau ReliefF hanya dibahas sebagai studi literatur pembanding dan tidak diimplementasikan secara mendalam.

3. Algoritma Klasifikasi

   Model klasifikasi yang dikembangkan berfokus pada algoritma SVM. Algoritma ini dipilih karena kemampuannya dalam menangani data berdimensi tinggi. Penggunaan algoritma *machine learning* lain (seperti *Random Forest* atau *Neural Network*) tidak menjadi fokus utama penelitian ini, kecuali jika diperlukan untuk perbandingan performa dasar (*baseline*).

4. Lingkup Diagnosis (Staging)

   Penelitian ini berfokus pada klasifikasi tahapan (*staging*) kanker payudara berdasarkan sistem AJCC (*American Joint Committee on Cancer*). Fokus klasifikasi dibatasi pada pembedaan antara Stadium Awal (*Early Stage*) yang mencakup *Stage I dan II*, dengan Stadium Lanjut (*Late Stage*) yang mencakup *Stage III*. Penelitian ini tidak mencakup sampel dengan label *Stage 0* (in situ), *Stage IV* (metastatik), atau sampel dengan data stadium yang tidak lengkap (*missing value*).

5. Evaluasi Performa

   Evaluasi performa model dilakukan menggunakan metrik umum dalam diagnosis medis berbasis AI, yaitu akurasi, sensitivitas (recall), spesifisitas, dan *Area Under Curve* (AUC). Analisis klinis lanjutan seperti uji prospektif, validasi eksternal, atau implementasi dalam sistem klinis nyata tidak termasuk dalam lingkup penelitian ini.

1.4 Tujuan Penelitian

Mengacu pada perumusan masalah yang telah dipaparkan, penelitian ini memiliki sasaran utama untuk membangun model komputasi yang efektif dalam membedakan stadium kanker

payudara berbasis data ekspresi gen. Secara spesifik, tujuan yang ingin dicapai melalui penelitian ini adalah:

1. Mengembangkan metode *feature selection* pada data ekspresi gen menggunakan pendekatan DFP, guna memperoleh fitur yang paling representatif dan relevan dalam proses diagnosis kanker payudara.
2. Menerapkan algoritma SVM untuk klasifikasi data ekspresi gen, sehingga dapat meningkatkan kemampuan sistem dalam mendeteksi serta mengklasifikasikan kanker payudara secara akurat.

1.5 Urgensi Penelitian

Urgensi penelitian ini didasarkan pada tingginya angka insidensi dan mortalitas kanker payudara serta keterbatasan metode diagnostik konvensional dalam membedakan stadium kanker secara presisi. Perbedaan stadium memiliki implikasi klinis yang krusial terhadap pemilihan terapi dan prognosis pasien, namun pendekatan berbasis klinis dan radiologis masih kurang sensitif dalam menangkap perubahan biologis, terutama pada stadium awal. Data ekspresi gen menawarkan potensi diagnosis yang lebih akurat, tetapi karakteristiknya yang berdimensi tinggi dengan jumlah sampel terbatas serta mengandung ketidakpastian biologis menimbulkan tantangan tersendiri dalam proses klasifikasi. Oleh karena itu, diperlukan pendekatan seleksi fitur yang adaptif dan *robust* untuk mengekstraksi informasi diskriminatif yang relevan secara biologis. Penelitian ini menjadi penting karena mengusulkan integrasi DFP dan SVM sebagai solusi komputasional untuk meningkatkan akurasi klasifikasi stadium kanker payudara serta mendukung pengembangan sistem pendukung keputusan medis berbasis bioinformatika yang lebih objektif dan presisi.

1.6 Luaran Penelitian

Sebagai hasil dari pelaksanaan penelitian ini, diperoleh sejumlah luaran yang merepresentasikan capaian ilmiah dan teknis dari pendekatan yang diusulkan. Luaran penelitian ini dirancang untuk mendokumentasikan proses penelitian secara sistematis serta menyajikan hasil evaluasi model secara objektif, sehingga dapat dimanfaatkan sebagai referensi bagi pengembangan penelitian lanjutan di bidang bioinformatika dan kecerdasan buatan dalam diagnosis kanker.

1. Laporan Ilmiah: Naskah ilmiah yang mendokumentasikan metodologi penelitian, proses seleksi fitur menggunakan DFP, pembangunan model klasifikasi berbasis SVM, serta hasil analisis dan kesimpulan penelitian, yang dapat dijadikan referensi untuk penelitian selanjutnya.
2. Hasil Evaluasi Model: Serangkaian luaran evaluasi yang mencakup metrik kinerja seperti accuracy, precision, recall, F1-score, confusion matrix, dan ROC pada setiap skenario percobaan, baik dengan maupun tanpa penerapan seleksi fitur, guna menunjukkan perbandingan performa model secara komprehensif dan terukur.

1.7 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan, baik dari segi pengembangan keilmuan maupun penerapan praktis di lapangan. Kontribusi tersebut mencakup pengayaan literatur dalam bidang informatika medis dan bioinformatika, serta penyediaan alternatif solusi teknologi untuk mendukung diagnosis klinis. Secara lebih rinci, manfaat penelitian ini dijabarkan ke dalam dua kategori utama sebagai berikut:

1. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang *Artificial Intelligence* (AI) dan *Bioinformatics*, khususnya dalam penerapan metode *feature selection* berbasis DFP dan integrasi multimodal antara data citra medis dan data genomik. Hasil penelitian ini juga dapat menjadi referensi ilmiah bagi penelitian selanjutnya yang berfokus pada diagnosis kanker atau penyakit kompleks lainnya menggunakan pendekatan kecerdasan buatan.
2. Penelitian ini berpotensi mendorong inovasi dalam pengembangan sistem diagnosis berbasis AI di bidang onkologi. Integrasi antara deep learning dan machine learning dengan teknik feature selection canggih diharapkan dapat menghasilkan model yang lebih adaptif, interpretatif, dan memiliki potensi untuk diterapkan pada sistem pendukung keputusan medis berbasis digital di masa depan.
3. Hasil penelitian ini diharapkan dapat membantu tenaga medis dan peneliti dalam meningkatkan akurasi diagnosis kanker payudara melalui penerapan algoritma SVM pada data ekspresi gen. Dengan memanfaatkan informasi profil molekuler tersebut secara optimal, sistem yang dikembangkan dapat berperan sebagai alat bantu diagnosis yang objektif dan efisien, sehingga mempercepat proses deteksi dini dan mendukung pengambilan keputusan klinis yang lebih presisi.