

BAB 1 PENDAHULUAN

1.1 Latar Belakang Masalah

Salah satu jenis kanker dengan angka kejadian kedua tertinggi di Indonesia maupun di dunia dan juga salah satu penyebab utama kematian akibat kanker pada perempuan adalah kanker payudara [1–3]. Berdasarkan data *Global Cancer Observatory* (GLOBOCAN) pada tahun 2022 yang diterbitkan oleh *World Health Organization* (WHO), kanker payudara mendominasi kasus kanker pada wanita di Indonesia dengan jumlah mencapai 66.271 kasus yang merupakan 16,2% dari seluruh kasus kanker baru yang terdeteksi. Tingginya angka kejadian ini menjadikan kanker payudara sebagai salah satu prioritas utama dalam bidang kesehatan dan penelitian klinis di Indonesia [4–6].

Diantara berbagai sub tipe kanker payudara, *Triple-Negative Breast Cancer* (TNBC) merupakan sub tipe yang paling agresif dan memiliki prognosis terburuk [7–9]. TNBC didefinisikan sebagai kanker payudara yang tidak mengekspresikan *reseptor estrogen* (ER), *reseptor progesteron* (PR), maupun protein *human epidermal growth factor receptor 2* (HER2). Kondisi ini menyebabkan terapi hormonal dan terapi bertarget yang efektif untuk sub tipe kanker payudara lainnya tidak dapat diterapkan pada pasien TNBC [10–12]. Secara global, TNBC menyumbang sekitar 10-20% dari seluruh kasus kanker payudara invasif, dengan rata-rata angka ketahanan hidup lima tahun (*5-year survival rate*) yang hanya mencapai sekitar 77% untuk stadium awal dan jauh lebih rendah pada kasus metastatik [13–15].

Karakteristik TNBC yang sangat heterogen pada tingkat molekuler membuat diagnosis dan klasifikasi sub tipe yang akurat menjadi tantangan besar dalam praktik klinis [16, 17]. Metode diagnosis konvensional seperti *imunohistokimia* (IHC) bergantung pada ekspresi protein reseptor dan memiliki keterbatasan dalam membedakan TNBC dari sub tipe kanker payudara lainnya secara menyeluruh [18, 19]. Oleh karena itu, pendekatan berbasis analisis ekspresi gen menggunakan data transkriptomik berpotensi memberikan gambaran yang lebih komprehensif dan presisi dalam membedakan TNBC, kanker payudara non-TNBC, dan jaringan payudara normal [20]. Untuk penelitian ini, ketiga kelompok tersebut selanjutnya masing-masing dinotasikan sebagai TNBC, BC, dan NM.

Perkembangan teknologi sekuensing generasi berikutnya *Next-Generation Sequencing* (NGS) dan tersedianya repositori data genomik publik berskala besar seperti *The Cancer Genome Atlas* (TCGA) yang dapat diakses melalui platform *UCSC Xena* memberikan peluang besar bagi peneliti untuk mengembangkan model klasifikasi berbasis *machine learning* yang akurat. Dataset TCGA-BRCA yang tersedia di *UCSC Xena* mencakup data ekspresi gen RNA-seq (*HiSeq V2*) dari ratusan sampel TNBC, BC, serta NM, menjadikannya sumber data yang ideal untuk penelitian klasifikasi multikelas [21].

Beberapa penelitian sebelumnya telah menggunakan pendekatan model berbasis *machine learning* yang dikombinasikan dengan berbagai metode *Feature Selection* (FS) untuk mengidentifikasi biomarker yang digunakan. Penelitian-penelitian tersebut memanfaatkan data ekspresi gen dari hasil teknologi NGS. Penelitian oleh Quist *et al.* memanfaatkan algoritma *Random Forest* (RF) pada data ekspresi gen berdimensi tinggi untuk klasifikasi kanker payudara. Penelitian tersebut menunjukkan bahwa *Random Forest* mampu menangani karakteristik data biologis yang memiliki jumlah fitur sangat besar dan hubungan antar fitur yang kompleks, sekaligus memberikan informasi mengenai pentingnya setiap fitur melalui *feature importance score* [22].

Pendekatan berbasis RF juga digunakan oleh Luo *et al.* dalam penelitian *Forest Subtype* untuk identifikasi subtype kanker menggunakan data ekspresi gen berdimensi tinggi. Pada penelitian tersebut, *parallel Random Forest* dimanfaatkan untuk melakukan seleksi fitur dan identifikasi subtype kanker payudara berdasarkan data dari *The Cancer Genome Atlas* (TCGA). Hasil penelitian menunjukkan bahwa pendekatan tersebut mampu mengidentifikasi subtype kanker secara lebih efektif dibandingkan beberapa metode pembandingan yang digunakan [23].

Selain *Random Forest*, algoritma *eXtreme Gradient Boosting* (XGBoost) juga semakin banyak digunakan dalam analisis data genomik dan transkriptomik karena kemampuannya dalam menangani data berdimensi tinggi serta memodelkan hubungan nonlinier antar gen. Deng *et al.* mengembangkan metode seleksi gen berbasis XGBoost yang dikombinasikan dengan *Multi-Objective Genetic Algorithm* (MOGA) untuk klasifikasi kanker menggunakan data ekspresi gen. Hasil penelitian menunjukkan bahwa XGBoost mampu melakukan pemeringkatan fitur yang relevan secara efektif dan menghasilkan performa klasifikasi yang lebih baik dibandingkan beberapa metode seleksi fitur lainnya [24].

Penelitian-penelitian sebelumnya menjadi dasar untuk penelitian ini. Sebelum membangun model klasifikasi, identifikasi gen-gen yang secara statistik

signifikan dalam membedakan ketiga kelas tersebut merupakan langkah krusial. Analisis *Differentially Expressed Genes* (DEG) menggunakan *Linear Models for Microarray and RNA-seq* (limma) telah terbukti efektif sebagai strategi seleksi fitur berbasis biologi untuk data ekspresi gen berukuran tinggi. Disisi lain, metode regularisasi statistik seperti *Least Absolute Shrinkage and Selection Operator* (LASSO) juga banyak digunakan sebagai pendekatan matematis untuk seleksi fitur pada data berdimensi tinggi seperti data ekspresi gen [25]. Perbandingan kedua metode seleksi fitur ini dalam konteks klasifikasi multikelas TNBC, BC, dan NM belum banyak dilakukan dalam literatur yang ada.

Salah satu tantangan utama dalam pemodelan *machine learning* pada data TCGA *Breast Invasive Carcinoma* (BRCA) adalah ketidakseimbangan kelas (*class imbalance*), dimana jumlah sampel TNBC dan NM jauh lebih sedikit dibandingkan sampel BC. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas dan menghasilkan performa yang buruk pada kelas minoritas. Teknik *oversampling* seperti *Synthetic Minority Oversampling Technique* (SMOTE) telah terbukti efektif dalam mengatasi permasalahan ini pada berbagai studi klasifikasi data transkriptomik [26, 27]. Selain performa prediksi, aspek interpretabilitas model menjadi semakin penting dalam aplikasi biomedis. Metode *Explainable Artificial Intelligence* (XAI) seperti *SHapley Additive exPlanations* (SHAP) memungkinkan identifikasi kontribusi masing-masing gen terhadap keputusan klasifikasi model, sehingga hasil analisis dapat dikaitkan kembali dengan pengetahuan biologis yang ada dan berpotensi mendukung penemuan biomarker baru untuk TNBC [28]. Penelitian ini mengusulkan sebuah model klasifikasi tiga kelas (TNBC, BC, dan NM) berbasis data ekspresi gen TCGA-BRCA dari UCSC Xena, dengan membandingkan tiga strategi seleksi fitur yaitu DEG berbasis limma, DEG berbasis limma dengan LASSO, dan juga LASSO dengan *recursive feature elimination* (RFE). Model klasifikasi yang digunakan dalam penelitian adalah RF dan XGBoost yang telah terbukti efektif oleh penelitian sebelumnya, serta *stacking* RF dan XGBoost untuk melihat efektivitas apabila kedua model digabung. Penelitian ini diharapkan dapat membantu dalam mengidentifikasi biomarker potensial yang tidak hanya mendukung klasifikasi kanker payudara, tetapi juga membuka peluang untuk pengembangan metode diagnostik yang lebih akurat dan terpersonalisasi.

1.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana performa model RF dan XGBoost dalam mengklasifikasikan tiga kelas (TNBC, BC, dan jaringan normal/NM) menggunakan data ekspresi gen TCGA-BRCA dari UCSC Xena?
2. Apakah terdapat perbedaan performa klasifikasi yang signifikan antara penggunaan DEG berbasis limma, DEG dengan LASSO, dan LASSO dengan RFE sebagai metode seleksi fitur?
3. Gen-gen apa saja yang paling berkontribusi dalam membedakan TNBC, BC, dan jaringan normal berdasarkan analisis *biomarker* dan interpretabilitas SHAP?

1.3 Batasan Permasalahan

Penelitian ini memiliki batasan sebagai berikut.

1. Data yang digunakan adalah data ekspresi gen RNA-seq HiSeq V2 dari dataset TCGA-BRCA yang diakses melalui platform UCSC Xena.
2. Klasifikasi dilakukan untuk tiga kelas yaitu TNBC, kanker payudara non-TNBC (BC), dan jaringan mamari normal (NM).
3. Penelitian ini tidak mencakup validasi eksternal menggunakan dataset dari luar TCGA-BRCA melalui platform UCSC Xena.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Membangun dan mengevaluasi model RF, XGBoost dan stacking RF dengan XGBoost untuk klasifikasi tiga kelas (TNBC, BC, dan NM) menggunakan data ekspresi gen TCGA-BRCA dari UCSC Xena.
2. Membandingkan efektivitas metode seleksi fitur berbasis DEG (limma), DEG dengan LASSO, dan LASSO dengan RFE dalam klasifikasi *multiclass* kanker payudara berdasarkan jumlah gen hasil seleksi serta kinerja model.

3. Mengidentifikasi gen-gen paling informatif yang berkontribusi pada klasifikasi menggunakan *pathway tools* dan analisis SHAP sebagai pendekatan *Explainable AI*.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut.

1. Memberikan kontribusi terhadap pengembangan sistem diagnosis kanker payudara.
2. Menambah wawasan ilmiah mengenai pengaruh ekspresi gen terhadap diagnosis subtype kanker payudara.
3. Membantu dalam identifikasi biomarker potensial berupa gen yang relevan pada kanker payudara.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut.

1. Bab 1 PENDAHULUAN
Berisi latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat, dan sistematika penulisan.
2. Bab 2 LANDASAN TEORI
Mengulas penelitian-penelitian terkait dan studi literatur yang mendasari penelitian ini.
3. Bab 3 METODOLOGI PENELITIAN
Menjelaskan secara rinci mengenai data ekspresi gen, metode preprocessing dan model yang digunakan.
4. Bab 4 HASIL DAN DISKUSI
Menyajikan dan membahas hasil eksperimen, termasuk perbandingan performa model antara jalur DEG dan LASSO serta analisis interpretabilitas SHAP.
5. Bab 5 KESIMPULAN DAN SARAN
Membahas hasil eksperimen, kontribusi penelitian, keterbatasan, serta saran untuk penelitian selanjutnya.