BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit kanker telah menjadi penyebab kematian dini paling umum, menempati posisi pertama atau kedua di sebagian besar negara di dunia. Diperkirakan jumlah pasien kanker secara global akan terus meningkat dalam 50 tahun ke depan (Soerjomataram & Bray, 2021). Berdasarkan data terakhir dari *Global Cancer Observatory* pada tahun 2022, terdapat 19.976.499 kasus baru dan 9.743.832 kematian akibat kanker di seluruh dunia. Dari jumlah data tersebut, kanker payudara menyumbang sekitar 11,5% dari total kasus yang ada, menempatkannya sebagai salah satu jenis kanker kedua yang paling umum terjadi. Di Indonesia sendiri terdapat 2.296.840 kasus dan 666.103 kematian yang disebabkan oleh kanker payudara (*Global Cancer Observatory*, 2022).

Kanker payudara adalah jenis kanker yang terjadi akibat sel-sel di dalam jaringan payudara tumbuh secara abnormal atau tidak terkendali (Suparna & Sari, 2022). Penyakit ini memiliki karakteristik yang heterogen, sehingga dapat diklasifikasikan ke dalam beberapa subtipe utama, yaitu *Luminal A, Luminal B, HER2,* dan *Triple Negative Breast Cancer* (TNBC) (Cosar et al., 2022). Setiap subtipe memiliki perbedaan signifikan dalam hal prognosis, tingkat agresivitas, serta respons terhadap terapi. Sehingga, identifikasi subtipe yang tepat sangat penting untuk menentukan pendekatan pengobatan yang optimal.

Banyak dari penelitian sebelumnya yang telah memanfaatkan *machine learning* untuk klasifikasi kanker payudara dan menunjukkan tingkat akurasi yang sangat baik. Sebagai contoh, penelitian yang dilakukan Rizkyani et al. (2021) yang menggunakan algoritma *machine learning Random Forest* dan *Adabost* untuk mendeteksi kanker payudara menghasilkan tingkat akurasi model sebesar 95% dan 70%. Pada penelitian Andryan et al. (2022) yang menggunakan algoritma *XGBoost* dan *Support Vector Machine (SVM)* untuk diagnosa kanker payudara juga menunjukkan tingkat akurasi yang sangat baik yaitu sebesar 95,12% dan 90,24%. Penelitian lain oleh Banerjee et al. (2025) yang menerapkan lima model *machine learning* untuk memprediksi kanker payudara juga menunjukkan performa yang

sangat baik, di mana KNN mencapai akurasi sebesar 94%, SVC sebesar 96%, *Random Forest* sebesar 99%, *Gradient Boosting* sebesar 96% dan *XGBoost* sebesar 95,6%. Meskipun algoritma *machine learning* terbukti efektif, sebagian besar penelitian tersebut masih terbatas pada data sitologi, yaitu data yang menggambarkan karakteristik morfologi sel (Bardales, 2022), seperti *radius, texture,* dan area, sehingga potensi informasi penting dari data genetik, seperti data *Ribonucleid acid* (RNA) sering kali terabaikan. Hal ini menunjukkan perlu adanya pendekatan baru yang memanfaatkan data genetik untuk menghasilkan klasifikasi yang lebih akurat dan mendalam.

Berbeda dengan penelitian sebelumnya, penelitian ini tidak hanya berfokus mengklasifikasikan subtipe kanker payudara menggunakan data RNA*,* tetapi juga berupaya mengidentifikasi gen-gen kunci yang dapat membantu memahami mekanisme biologis pada perkembangan sel kanker. Pemilihan data RNA sebagai fokus utama penelitian ini didasarkan pada kemampuannya untuk memberikan gambaran ekspresi genetik secara rinci, yang sangat berguna dalam memahami heterogenitas pada kanker payudara. Selain itu, integrasi data RNA dengan model *machine learning* memungkinkan analisis yang lebih akurat dan efisien untuk mengklasifikasikan subtipe kanker payudara serta mengidentifikasi gen-gen kunci yang terlibat.

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, penelitian ini bertujuan untuk membangun model *machine learning* yang mampu mengklasifikasikan subtipe kanker payudara serta mengidentifikasi gen-gen yang memiliki peran signifikan dalam perkembangan sel kanker, dengan memanfaatkan dan mengintegrasikan data ekspresi gen (RNA-*seq*) dari dataset *Breast Invasive Carcinoma (TCGA, PanCancer Atlas)*. Klasifikasi dilakukan menggunakan tiga algoritma *machine learning* dengan memanfaatkan data ekspresi gen (RNA-*seq*) sebagai *input* utama. Selain itu, penelitian ini juga diharapkan dapat memberi wawasan penting dalam pengembangan pengobatan kanker payudara yang lebih efektif dan personal.

1.2 Perumusan Masalah

Mengacu pada latar belakang masalah dan hasil penelitian terdahulu yang telah diuraikan sebelumnya, maka permasalahan utama yang dibahas dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana model *machine learning* dapat diterapkan untuk melakukan klasifikasi subtipe kanker payudara berbasis data RNA?
2. Algoritma manakah yang menghasilkan tingkat akurasi terbaik di antara algoritma yang digunakan untuk mengklasifikasikan subtipe kanker payudara?
3. Gen kunci apa saja yang berhasil diidentifikasi untuk setiap subtipe kanker payudara berdasarkan hasil analisis data RNA?

1.3 Batasan Masalah

Mengingat permasalahan yang dibahas dalam penelitian ini memiliki cakupan yang luas, batasan-batasan berikut ditetapkan untuk memastikan bahwa penelitian ini lebih terfokus dan terarah sesuai dengan tujuan yang ingin dicapai:

1. Data yang digunakan dalam penelitian ini merupakan data *RNA-seq* yang bersumber dari *cBioportal for Cancer Genomics* terkait kanker payudara tipe *Breast Invasive Carcinoma*.
2. Klasifikasi hanya dilakukan pada empat subtipe kanker payudara, yaitu *Luminal A, Luminal B, HER2,* dan *Basal,* sampel pasien dengan sel normal tidak digunakan dalam penelitian ini, karena sampel tersebut merepresentasikan jaringan normal, bukan jaringan tumor, sehingga kurang relevan untuk klasifikasi subtipe kanker payudara.
3. Model *machine learning* yang digunakan hanya mencakup tiga algoritma utama, yaitu *Gradient Boosting, XGBoost* dan *Random Forest.*
4. Identifikasi gen kunci dilakukan melalui pendekatan *feature importance* dan analisis diferensial ekspresi.
5. Nilai akurasi model diukur menggunakan metode *confusion matrix* dengan mempertimbangkan nilai *accuracy, F1-score*, dan *recall.*

6. Proses pengujian model dilakukan menggunakan *k-fold cross validation* dengan K=10.
7. Penelitian ini hanya mempertimbangkan nilai ekspresi dari setiap gen (RNA-*seq*) dan tidak mencakup analisis faktor lain seperti mutasi genetik dan tingkat kelangsungan hidup.
8. Analisis gen difokuskan pada 10 gen teratas yang memiliki nilai *feature importance* tertinggi yang diperoleh dari model dengan performa klasifikasi terbaik.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijelaskan sebelumnya, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Membuat dan mengidentifikasi model *machine learning* yang paling sesuai untuk klasifikasi subtipe kanker payudara berbasis data RNA.
2. Mengidentifikasi gen kunci yang relevan untuk setiap subtipe kanker payudara berdasarkan hasil analisis data RNA.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Untuk Peneliti

   Meningkatkan pengetahuan di bidang Bioinformatika dan kanker payudara, khususnya pada penerapan *machine learning* untuk analisis data biologis.
2. Untuk Pengembangan Teknologi

   Membuka peluang pengembangan alat diagnostik berbasis *machine learning* yang lebih akurat untuk deteksi dini dan klasifikasi subtipe kanker payudara berdasarkan data RNA.
3. Untuk Dunia Kesehatan
   - Membantu mengidentifikasi gen-gen penting yang dapat menjadi target potensial terapi pada setiap subtipe kanker payudara, sehingga pengobatan dapat lebih personal dan terarah.

- Memberikan data pendukung yang dapat digunakan untuk merancang strategi pengobatan yang lebih efisien dan terjangkau bagi pasien dengan subtipe kanker payudara tertentu.

1.6 Metodologi Penelitian

Penelitian ini dilakukan melalui beberapa tahapan, yang melibatkan pendekatan studi literatur, simulasi dan analisis data. Berikut merupakan tahapan-tahapan yang dilakukan dalam penelitian ini:

1. Pengumpulan data RNA-*seq* dan data klinis pasien terkait kanker payudara dari berbagai sumber.
2. *Preprocessing* data RNA-*seq* mencakup penghapusan ID pasien dengan subtipe yang tidak diketahui, normalisasi data menggunakan metode *DESeq2,* serta penghapusan data gen dengan nilai ekspresi yang rendah.
3. Pembagian dataset menjadi data *training* dan *testing* dalam proporsi tertentu.
4. Pengembangan model *machine learning* menggunakan algoritma *Gradient Boosting, XGBoost,* dan *Random Forest* dengan pendekatan *OnevsRest Classifier*.
5. Identifikasi gen kunci yang paling berkontribusi pada setiap subtipe kanker payudara menggunakan teknik *feature importance* dan *Differential Expression Analysis* (DEA).
6. Evaluasi model menggunakan metrik akurasi, F1-*Score*, *precision* dan *recall. Confusion matrix* digunakan untuk menganalisis distribusi prediksi model. Selain itu, penelitian ini juga menerapkan metode *k-fold cross-validation* untuk mengevaluasi stabilitas dan generalisasi model.
7. Gen-gen yang teridentifikasi pada setiap subtipe, dianalisis lebih lanjut untuk memahami bagaimana mereka berperan terhadap pertumbuhan dan perkembangan kanker payudara melalui studi literatur dan mengakses *data base* terkait *pathway* kanker payudara.

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir ini dikelompokkan ke dalam lima bab utama yang saling berkaitan satu sama lain, dengan rincian sebagai berikut:

BAB I : PENDAHULUAN

Bab ini menyajikan pemaparan awal terkait topik yang diangkat dalam penelitian, yang mencakup latar belakang masalah, rumusan masalah, batasan penelitian, tujuan dan manfaat yang ingin dicapai, serta uraian mengenai sistematika penulisan tugas akhir secara menyeluruh.

BAB II : TINJAUAN PUSTAKA

Bab ini memuat berbagai teori, konsep dan hasil kajian dari penelitian terdahulu yang relevan dengan topik penelitian. Seluruh teori yang disajikan digunakan sebagai landasan yang kuat untuk menjelaskan permasalahan yang diteliti serta mendukung proses analisis pada penelitian ini.

BAB III : METODOLOGI PENELITIAN

Pada bab ini dijelaskan secara sistematis langkah-langkah yang dilakukan dalam proses penelitian. Setiap tahapan yang dilakukan pada proses penelitian diuraikan secara detail dan runtut untuk memberikan gambaran yang jelas mengenai alur pelaksanaan penelitian secara keseluruhan.

BAB IV : HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari proses penelitian yang telah dilakukan, disertai dengan analisis dan interpretasi terhadap data yang diperoleh. Hasil penelitian akan disajikan dalam bentuk gambar, tabel, dan grafik untuk mempermudah pemahaman terhadap hasil penelitian.

BAB V : PENUTUP

Bab ini menyajikan kesimpulan dari hasil penelitian yang telah dijelaskan pada bab-bab sebelumnya. Selain itu, bab ini juga menyampaikan saran-saran yang dapat menjadi masukan bagi peneliti selanjutnya yang tertarik untuk mengeksplorasi topik serupa.