BAB 5
SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil analisis, eksperimen, dan evaluasi yang telah dilakukan pada penelitian ini, maka dapat disimpulkan beberapa hal penting sebagai berikut:

1. Berdasarkan hasil dari evaluasi pengujian menunjukan bahwa metode seleksi fitur berbasis statistik memberikan hasil yang lebih baik dibandingkan dengan pendekatan metode DEG. Kemudian, jumlah pemilihan fitur awal berpengaruh pada performa model. Model yang menggunakan seleksi fitur awal *Logistic Regression* 2000 memiliki performa yang lebih baik dibandingkan model yang menggunakan seleksi fitur awal *Logistic Regression* 4000. Kemudian, pada kombinasi seleksi fitur menggunakan ANOVA menghasilkan model yang lebih baik ketika menggunakan jumlah 4000 pemilihan fitur awal dibandingkan dengan 2000 pemilihan fitur awal. Hal ini dapat terjadi karena *Logistic Regression* dalam sifat dasarnya mempertimbangkan korelasi antar fitur dan efek multivariat, sedangkan ANOVA hanya memperhitungkan variansi antar kelas secara individual. Maka, memperbesar jumlah fitur awal pada ANOVA cenderung lebih menguntungkan, selama RFE mampu menyaring fitur-fitur terbaik dari kumpulan yang lebih besar tersebut.

2. Berdasarkan hasil penelitian pendekatan dengan Stacking Classifier dapat mengklasifikasikan stadium II dan stadium III kanker payudara pada data berbasis ekspresi gen RNA-seq. Hal ini terbukti dari model terbaik yang telah dibangun menghasilkan nilai-nilai metrik evaluasi yang tinggi dengan menggunakan 48 fitur, model tersebut berhasil mencapai *accuracy* 93,47%, *precision* 93,52%, *recall* 93,47%, dan *f1-score* 93,49%. Model terbaik dicapai menggunakan kombinasi seleksi fitur *Logistic Regression* dengan pemotongan fitur awal sebanyak 2000, lalu proses lebih lanjut untuk seleksi fitur kembali menggunakan *Recursive Feature Elimination* (RFE) dan metode klasifikasi *Stacking Ensembel Learning* dengan *base learners* yang terdiri dari, *Xgboost, Random Forest, SVM, dan Logistic Regression*. Kemudian, *meta classifier* yang digunakan adalah Logistic Regression dengan *hyperparameter* C bernilai 1, penalty l2, dan *solver* liblinear.

3. Performa terbaik secara keseluruhan diperoleh kombinasi seleksi fitur *Logistic Regression* 2000 + RFE dengan model *Stacking Classifier*. Model tersebut menggunakan 48 fitur. Sebanyak 48 gen tersebut, 7 diantaranya, yaitu CRYGN, LINC01606, PXN, ACTL7A, HNF4A, LINC02268, dan SPPL2C telah ditemukan dalam penelitian/litaratur terdahulu sebagai gen-gen yang berhubungan dengan kanker payudara atau progresi kanker. Lalu, analisis lebih lanjut dilakukan dengan melakukan perhitungan nilai AUC pada masing-masing gen di 48 gen terpilih. Hasil analisis tersebut menunjukan bahwa tidak ada gen yang secara individual memiliki nilai AUC $>$ 0,7. Hal ini menunjukan bahwa secara individual, gen-gen tersebut memiliki kemampuan diskriminatif yang rendah terhadap stadium II dan stadium III kanker payudara. Namun, jika diperhatikan lebih lanjut korelasi dan kontribusi kolektif dari gen-gen tersebut dapat memberikan hasil performa model yang lebih tinggi dibandingkan per individu gen tersebut.

5.2 Saran

Berdasarkan hasil evaluasi dan temuan dalam penelitian ini, terdapat sejumlah rekomendasi yang dapat dijadikan acuan untuk pengembangan studi di masa mendatang. Salah satu poin utama adalah bahwa pilihan metode seleksi fitur memiliki dampak signifikan terhadap kinerja model, baik dalam hal akurasi prediksi maupun efisiensi proses komputasi. Oleh karena itu, sangat dianjurkan untuk mengeksplorasi lebih jauh penggunaan metode seleksi fitur yang lebih adaptif dan konsisten, seperti pendekatan berbasis ensemble yang lebih canggih dan kompleks atau teknik berbasis deep learning, yang berpotensi menghasilkan fitur-fitur dengan kekuatan prediktif yang lebih tinggi dalam membedakan stadium kanker payudara. Mengingat sifat data ekspresi genetik yang kompleks dan jumlah fitur yang sangat besar, diperlukan strategi seleksi fitur yang mampu mengidentifikasi perbedaan ekspresi antar gen, sehingga analisis dapat mencerminkan kondisi biologis secara lebih menyeluruh dan akurat.