BAB 5
SIMPULAN SARAN

5.1 Simpulan

Berdasarkan perancangan arsitektur dan hasil pengujian yang telah dijelaskan pada bab-bab sebelumnya, penelitian ini memberikan gambaran menyeluruh mengenai performa sistem AIRA sebagai platform *backend* untuk pemrosesan *ML inference* di bidang kesehatan. Evaluasi dilakukan melalui *load testing*, *stress testing*, serta pengujian komparatif antar model untuk memperoleh pemahaman empiris mengenai kapasitas, stabilitas, dan karakteristik komputasi sistem.

1. Dari sisi perancangan sistem, arsitektur *backend* yang dikembangkan berhasil memisahkan *service* antara *Cancer Gateway* dan AI *Backend*. *Cancer Gateway* terdiri dari tujuh *service* utama, sedangkan AI *Backend* menyediakan dua *service* inti dengan total tujuh model prediksi kanker. Pendekatan *layered architecture* dengan pola API *gateway* dan pemanfaatan Docker pada lingkungan *single-server* menghasilkan struktur yang modular, mudah dipelihara, dan fleksibel untuk menangani berbagai konfigurasi model. Arsitektur ini juga menjadi dasar penting bagi proses pengujian performa yang dilakukan pada penelitian ini.

2. Dari sisi performa pada skenario *load test* (skenario 1 dan skenario 2 bagian *load*), sistem AIRA mampu menangani hingga sekitar 100 *concurrent users* tanpa penurunan performa yang signifikan. Model kanker prostat menunjukkan waktu respons rata-rata yang rendah pada kisaran 50–70 ms, dengan *throughput* tinggi dan *error rate* mendekati nol. Sebaliknya, model kanker payudara memiliki waktu respons rata-rata yang jauh lebih tinggi, yaitu sekitar 480–3600 ms. Hasil ini konsisten dengan skenario pengujian yang menunjukkan bahwa kompleksitas model berperan langsung terhadap kapasitas pemrosesan dan stabilitas latency pada sistem.

3. Pada skenario *stress test* (skenario 2 bagian *stress*), batas kapasitas sistem mulai terlihat pada rentang 100 hingga 200 pengguna. Waktu respons meningkat secara drastis pada beban tinggi, terutama untuk model kanker payudara yang mencapai lebih dari 36 detik pada 1000 pengguna. Selain itu,

error mulai muncul pada beban di atas 500 pengguna, meskipun tetap berada pada tingkat yang rendah. Sementara itu, model prostat tetap stabil hingga skenario tertinggi tanpa menghasilkan *error*. Pola ini menunjukkan bahwa *bottleneck* utama berada pada proses *inference* di AI *Backend*, bukan pada jaringan atau mekanisme API *gateway*. Hasil dari skenario ini menegaskan bahwa optimasi perlu difokuskan pada komponen *inference*, terutama untuk model dengan kompleksitas tinggi.

5.2 Saran

Berdasarkan hasil penelitian dan pengalaman selama pelaksanaan kegiatan, beberapa saran yang dapat diberikan untuk pengembangan sistem AIRA adalah sebagai berikut.

1. Sistem AIRA dapat ditingkatkan performanya dengan melakukan optimasi pada proses *inference*, terutama untuk model dengan kompleksitas tinggi seperti kanker payudara. Teknik seperti *model optimization*, *quantization*, atau pemanfaatan *batching* dapat dipertimbangkan untuk menurunkan waktu respons pada beban tinggi.

2. Untuk mendukung skenario penggunaan yang lebih besar, sistem dapat dikembangkan menuju arsitektur *multi-server* atau *distributed deployment* sehingga proses *scaling* dapat dilakukan lebih fleksibel dan tidak terbatas pada satu lingkungan *single-server* seperti pada penelitian ini.

3. Penambahan *monitoring tools* yang lebih komprehensif, seperti *logging* terpusat dan *performance metrics*, akan membantu analisis *bottleneck* dan pemantauan sistem ketika berjalan dalam jangka panjang atau pada lingkungan produksi.

4. Dokumentasi teknis, hasil pengujian, serta praktik terbaik yang diperoleh selama penelitian disarankan untuk dirapikan dan disimpan sebagai referensi bagi pengembangan selanjutnya, sehingga mempermudah proses iterasi baik oleh peneliti maupun tim pengembang berikutnya.