BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil pengujian, *MobileNetV3 Large* terbukti menjadi model dasar terbaik untuk klasifikasi sampah dengan akurasi *test* tertinggi sebesar 0.8512 dan akurasi validasi 0.8568, menunjukkan kemampuan generalisasi yang baik meskipun memiliki waktu inferensi awal 0.089493 detik per citra dan ukuran model 16.03 MB. Penerapan *structured weight pruning* berbasis *magnitude* secara bertahap hingga 40% mampu meningkatkan efisiensi model secara signifikan tanpa menurunkan performa, dengan akurasi *test* tetap terjaga pada 0.86 dan waktu inferensi mengalami penurunan sebesar 93.17% per citra. Selanjutnya, *post-training quantization* dengan skema *FP16* memberikan hasil paling optimal dengan menurunkan ukuran model sebesar 63.50% tanpa penurunan akurasi yang berarti. Kombinasi *structured weight pruning* bertahap dengan *MobileNetV3 Large* sebagai *base* model dan *post-training quantization FP16* menghasilkan performa terbaik secara keseluruhan dengan akurasi test 0.86, penurunan waktu inferensi sebesar 89.45% per citra, dan penurunan ukuran model sebesar 63.50%, sehingga memberikan keseimbangan optimal antara akurasi, kecepatan inferensi, dan efisiensi penyimpanan untuk penerapan pada sistem klasifikasi sampah berbasis citra dengan keterbatasan sumber daya.

5.2 Saran

Berikut adalah saran penelitian yang sepenuhnya disusun berdasarkan hasil dan pengalaman selama penelitian optimasi *MobileNet* dengan *pruning* dan *quantization*:

1. Keterbatasan eksplorasi konfigurasi optimasi, penelitian ini masih menggunakan variasi tingkat *pruning* dan skema *quantization* yang terbatas, meskipun kombinasi *pruning* bertahap dan *quantization FP16* telah menunjukkan hasil yang optimal. Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi konfigurasi optimasi yang lebih luas,

seperti tingkat *pruning* yang lebih halus maupun skema *quantization* lain, guna memperoleh titik keseimbangan terbaik antara akurasi, ukuran model, dan waktu inferensi.

2. Sensitivitas performa terhadap konfigurasi pelatihan, selama proses penelitian, kestabilan akurasi model sangat dipengaruhi oleh pengaturan parameter pelatihan awal dan setelah optimasi. Oleh karena itu, disarankan untuk melakukan *hyperparameter tuning* secara sistematis sebelum dan sesudah penerapan *pruning* serta *quantization* agar performa model tetap stabil dan optimal.

3. Evaluasi yang masih terbatas pada lingkungan pengujian, pengukuran efisiensi model pada penelitian ini dilakukan dalam lingkungan eksperimen, sehingga belum sepenuhnya merepresentasikan kondisi penggunaan nyata. Oleh karena itu, penelitian lanjutan disarankan untuk menguji model hasil optimasi secara langsung pada perangkat dengan sumber daya terbatas, seperti perangkat *embedded*, guna memperoleh gambaran performa inferensi yang lebih realistis.

4. Ketidakseimbangan jumlah data antar kelas, selama proses evaluasi masih ditemukan kesalahan prediksi pada kelas tertentu yang memiliki jumlah data relatif lebih sedikit. Oleh karena itu, disarankan untuk menambah serta menyeimbangkan jumlah data pada setiap kelas agar kemampuan generalisasi model setelah dioptimasi dapat ditingkatkan.

5. Keterbatasan sumber daya komputasi untuk eksperimen berulang, proses optimasi dan evaluasi memerlukan waktu komputasi yang cukup besar sehingga membatasi jumlah eksperimen lanjutan. Oleh karena itu, penggunaan *environment* dengan dukungan sumber daya komputasi dan *runtime* yang lebih fleksibel sangat disarankan untuk mendukung pengujian yang lebih mendalam dan komprehensif.