BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan metode *Knowledge Distillation (KD)*, *Post-Training Quantization (PTQ)*, serta kombinasi keduanya (*hybrid*) menghasilkan perubahan kinerja model yang terukur secara persentase apabila dibandingkan dengan model *baseline MobileNetV3Small*. Evaluasi dilakukan terhadap tiga aspek utama, yaitu waktu inferensi, ukuran model, dan akurasi, dengan nilai persentase dihitung relatif terhadap performa model baseline.

Dari sisi waktu inferensi, metode optimasi menunjukkan dampak yang berbeda-beda. Model berbasis *Post-Training Quantization* dan *hybrid* menghasilkan nilai persentase penurunan waktu inferensi yang bernilai negatif, yaitu berada pada rentang –43.11% hingga –51.52%, yang menunjukkan adanya percepatan waktu inferensi dibandingkan model baseline. Penurunan waktu inferensi terbesar dicapai oleh model PTQ2 sebesar –51.52%, diikuti oleh Hybrid1 sebesar –50.15% dan PTQ1 sebesar –49.86%. Sebaliknya, seluruh model berbasis *Knowledge Distillation* murni menunjukkan nilai persentase positif yang cukup besar, yaitu berada pada rentang 271.00% hingga 294.77%, yang mengindikasikan bahwa waktu inferensi model KD lebih lambat dibandingkan model *baseline MobileNetV3Small*.

Dari sisi ukuran model, seluruh metode optimasi berhasil menghasilkan pengurangan ukuran model dibandingkan baseline. Model *berbasis Knowledge Distillation* menghasilkan penurunan ukuran model yang konsisten sebesar –35.95%. Metode *Post-Training Quantization* memberikan dampak pengurangan ukuran model paling signifikan, dengan penurunan mencapai –73.65% pada model PTQ1 dan –47.81% pada PTQ2. Model *hybrid* juga menunjukkan pengurangan ukuran model yang cukup besar, dengan persentase penurunan berada pada rentang –44.39% hingga –71.25%, sehingga menghasilkan model yang lebih ringan dibandingkan baseline.

Dari sisi akurasi, seluruh model hasil optimasi menunjukkan nilai persentase kenaikan akurasi yang bernilai negatif, yaitu berada pada rentang –17.07% hingga –25.26%. Hal ini menunjukkan bahwa secara relatif, akurasi model hasil optimasi lebih rendah dibandingkan model baseline. Penurunan akurasi paling kecil terjadi pada model berbasis *Knowledge Distillation*, khususnya KD1 dan KD4, dengan nilai masing-masing –17.07% dan –17.12%, serta pada model Hybrid1 dan Hybrid2 yang berada pada kisaran –17.11% hingga –17.17%. Sementara itu, model berbasis *Post-Training Quantization* mengalami penurunan akurasi terbesar, yaitu sebesar –25.26%.

Secara keseluruhan, penelitian ini menyimpulkan bahwa penerapan teknik optimasi model memberikan dampak yang berbeda terhadap efisiensi komputasi dan performa klasifikasi. Metode *Post-Training Quantization* dan pendekatan *hybrid* terbukti paling efektif dalam meningkatkan efisiensi, dengan penurunan waktu inferensi sebesar 43.11% hingga 51.52% dan pengurangan ukuran model hingga 73.65%, meskipun disertai penurunan akurasi relatif sebesar 17.11% hingga 25.26%. Di sisi lain, *Knowledge Distillation* mampu mempertahankan akurasi yang lebih baik dengan penurunan sebesar 17.07% hingga 18.70%, namun mengakibatkan peningkatan waktu inferensi yang signifikan, yaitu 271.00% hingga 294.77% dibandingkan model baseline. Temuan ini menegaskan adanya trade-off yang jelas antara efisiensi komputasi dan performa klasifikasi, serta secara langsung menjawab rumusan masalah penelitian mengenai sejauh mana peningkatan efisiensi waktu inferensi dapat dicapai melalui penerapan *quantization* dan *knowledge distillation* pada model *MobileNetV3Small*.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa hal yang masih dapat dikembangkan dan ditingkatkan pada penelitian selanjutnya, antara lain sebagai berikut:

1. Evaluasi performa model pada perangkat nyata, seperti *edge device* atau perangkat embedded, disarankan untuk dilakukan agar hasil pengukuran efisiensi waktu inferensi dan konsumsi sumber daya dapat lebih merepresentasikan kondisi implementasi sebenarnya.

2. Penelitian ini belum mempertimbangkan aspek konsumsi daya (*power consumption*) selama proses inferensi. Oleh karena itu, penelitian lanjutan dapat memasukkan analisis konsumsi energi sebagai metrik tambahan, terutama untuk mendukung penerapan sistem pada perangkat dengan keterbatasan daya.