BAB V

KESIMPULAN DAN SARAN

Bab penutup ini merangkum seluruh hasil dan pembahasan pada Bab IV menjadi kesimpulan yang dipetakan langsung terhadap rumusan permasalahan dan tujuan penelitian (5.1), serta saran dan rekomendasi bagi penelitian selanjutnya (5.2).

5.1 Kesimpulan

Bagian ini menyajikan kesimpulan penelitian yang dipetakan langsung terhadap kelima Rumusan Permasalahan (Bab I sub-bab 1.2) dan Tujuan Penelitian (Bab I sub-bab 1.3), termasuk keterbatasan yang masih terbuka (mis. LOSO-CV, *live testing* integrasi 2GT dengan pengguna nyata).

1. **Pipeline Knowledge Distillation domain-matched (menjawab Rumusan Permasalahan #1 dan Tujuan #1):**

Penelitian ini berhasil merancang dan mengevaluasi pipeline KD domain-matched dari model *Teacher* DDAMFN++ ke arsitektur *Student* MobileNetV3 pada domain RAF-DB (Bab III sub-bab 3.3, Bab IV sub-bab 4.2). Model *Student* MobileNetV3-Large berhasil melampaui model *Teacher*-nya pada Balanced Accuracy (73,16% vs 71,30%, +1,86 poin, Tabel 4.2), memvalidasi secara empiris bahwa distilasi domain-matched dapat menghasilkan model *Student* yang menyamai atau bahkan melampaui kompetensi *Teacher*-nya sendiri pada domain evaluasi yang identik. Namun, kesimpulan ini perlu disertai satu nuansa penting (sub-bab 4.8.2): pada metrik Macro F1-Score, model *Teacher* justru sedikit lebih unggul (74,17% vs 74,06%), yang menunjukkan keunggulan *Student* lebih terasa pada *Recall* rata-rata (Balanced Accuracy) dibanding keseimbangan penuh Precision-Recall di seluruh kelas. Rumusan Permasalahan #1 dapat dijawab positif dengan nuansa: KD domain-matched terbukti efektif, namun keunggulannya bersifat metrik-spesifik, bukan keunggulan mutlak di semua ukuran evaluasi.

2. **Pengaruh kesesuaian *domain* Teacher-Student terhadap efektivitas *KD* (menjawab Rumusan Permasalahan #2 dan Tujuan #2):**

Perbandingan langsung antara Eksperimen 1 (domain-*mismatch*, *Teacher* lemah pada domain target, Balanced Accuracy 74,5%) dan Eksperimen 2 (domain-*matched*, *Teacher* kuat pada domain evaluasi, Balanced Accuracy 71,30%) memberikan bukti empiris yang jelas (Bab IV sub-bab 4.3, Tabel 4.3): pada skenario *mismatch*, KD secara konsisten merugikan kedua varian *Student* (Large −1,43 poin, Small −1,77 poin dibanding pelatihan langsung tanpa distilasi), sedangkan pada skenario *matched*, KD dapat menguntungkan *Student* berkapasitas memadai (Large +1,86 poin) namun tetap dapat merugikan *Student* berkapasitas terbatas (Small −2,63 poin). Kesimpulan atas Rumusan Permasalahan #2 adalah bahwa kesesuaian domain *Teacher-Student* merupakan faktor penentu yang perlu namun tidak selalu cukup, karena kapasitas representasi model *Student* turut menentukan keberhasilan transfer pengetahuan. Temuan inilah yang menjadi justifikasi empiris utama desain metodologis dua-tahap penelitian ini (sub-bab 4.8.3).

3. **Strategi transfer learning ke domain sasaran IMED–IMSFD (menjawab Rumusan Permasalahan #3 dan Tujuan #3):**

Strategi *discriminative fine-tuning* / *progressive layer freezing* (Howard & Ruder, 2018; Bab II sub-bab 2.5.2, Persamaan 2.7–2.8) telah dirancang, diimplementasikan, dan dieksekusi penuh (`source_code/FER_MobileNetV3_TransferLearning_Hybrid_Kaggle.ipynb`, Bab III sub-bab 3.4, hasil lengkap Bab IV sub-bab 4.5), yaitu enam kelompok parameter dengan *learning rate* proporsional kedalaman lapisan (Gambar 2.11), dimuat dari *checkpoint* terbaik Tahap I (bukan inisialisasi ImageNet dari nol) dan diadaptasi ke dataset hibrid IMED–IMSFD (20.337 citra, leakage-safe, Tabel 3.3). Hasilnya sangat meyakinkan: Balanced Accuracy melonjak dari 33,44% (baseline *zero-shot*, checkpoint Tahap I tanpa adaptasi) menjadi 91,89% setelah *fine-tuning*, kenaikan 58,45 poin persentase, dengan kelas Netral mencapai skor sempurna (Precision=Recall=F1=1,000, Tabel 4.12) dan tanpa tanda-tanda *overfitting* pada kurva konvergensi 30 epoch (Gambar 4.12). Rumusan Permasalahan #3 terjawab positif dan tuntas secara empiris: strategi *discriminative fine-tuning* dari checkpoint Tahap I terbukti sangat efektif mengadaptasi model ke domain populasi Indonesia

tanpa kehilangan efisiensi komputasi (arsitektur dan jumlah parameter tidak berubah, sub-bab 4.5.4).

4. **Efisiensi komputasi model *Student* (menjawab Rumusan Permasalahan #4 dan Tujuan #4):**

Model *Student* MobileNetV3 terbukti jauh lebih ringkas dibanding model *Teacher* SOTA sekelas (Bab II sub-bab 2.11, Bab IV sub-bab 4.3.1): MobileNetV3-Large (4,211M parameter) dan MobileNetV3-Small (1,525M parameter) berturut-turut hanya mencapai raw accuracy 82,69% dan 80,35% pada RAF-DB, sekitar 9–10 poin di bawah SOTA berat seperti POSTER++ (92,21%) dan DDAMFN++ asli (92,04%), namun dengan ukuran model yang jauh lebih kecil. Kuantisasi INT8 terverifikasi nyata lewat dua pipeline independen (ONNX Runtime QDQ dan TFLite *full-integer*, sub-bab 4.6) namun berdampak tidak seragam: MobileNetV3-Large mempertahankan akurasi dengan wajar (−3,22 poin pipeline ONNX, kompresi 3,58×), sedangkan MobileNetV3-Small mengalami kolaps akurasi drastis pada kedua pipeline (−42,28 poin ONNX; −43,52 poin TFLite *native*, 36,83% vs 80,35% FP32, Tabel 4.16). Konsistensi lintas-*runtime* ini mengonfirmasi kolaps tersebut adalah karakteristik arsitektur, bukan artefak satu metode kuantisasi tertentu. Kesimpulan atas Rumusan Permasalahan #4 bersifat positif dengan nuansa penting yang dilaporkan terbuka (sub-bab 4.9): angka latensi yang tercatat (Tabel 4.2) diukur pada GPU lingkungan pelatihan, bukan CPU *edge* yang menjadi konteks motivasi penelitian, sehingga perbandingan latensi CPU *edge* yang valid masih memerlukan pengukuran terpisah (sub-bab 5.2). Verifikasi akurasi pasca-kuantisasi mencakup baik checkpoint Tahap I maupun Tahap II, serta berkas `.tflite` aktual yang di-*deploy* ke 2GT Smart Classroom (sub-bab 4.7).

5. **Kelayakan integrasi ke platform 2GT Smart Classroom (menjawab Rumusan Permasalahan #5 dan Tujuan #5):**

Integrasi modul FER ke arsitektur 3-*tier* platform 2GT Smart Classroom (Bab III sub-bab 3.6) terverifikasi secara struktural/fungsional: dua *endpoint* REST API nyata ditemukan langsung dari kode (`api/telemetry.php`, `api/sync_engagement.php`), dan model *Student* terkuantisasi dari kedua tahap (Tahap I Large/Small+KD dan Tahap II Large *fine-tuned*) telah ditempatkan pada

direktori *deployment* sistem, dapat dipilih Administrator secara dinamis lewat menu AI Setting (Bab III sub-bab 3.6.4, Bab IV sub-bab 4.7). Sesuai batasan yang telah dinyatakan sejak awal (Bab I sub-bab 1.5), evaluasi ini tidak mencakup pengujian langsung (*live testing*) dengan pengguna nyata (siswa/guru), sehingga Rumusan Permasalahan #5 terjawab pada tingkat kelayakan teknis (arsitektur ada dan berfungsi sesuai rancangan), bukan pada tingkat kelayakan operasional tervalidasi pengguna. Indikasi awal keadilan algoritma (sub-bab 4.10), yaitu pola kekeliruan Tahap II yang konsisten secara kualitatif dengan Tahap I dan bukan kegagalan acak, bersifat positif namun belum mencakup analisis atribut demografis (gender, usia, penggunaan hijab).

Sebagai ringkasan kontribusi Tahap I dan Tahap II, sesuai kerangka dua-tahap penelitian ini (Bab I sub-bab 1.1), kedua tahap kini telah terbukti secara empiris. Tahap I memvalidasi skema KD domain-matched dan pengaruh kuantitatif kesesuaian domain terhadap efektivitas distilasi (Rumusan Permasalahan #1–#2, Balanced Accuracy 73,16% pada RAF-DB). Tahap II, adaptasi domain ke populasi Indonesia yang menjadi motivasi utama relevansi praktis penelitian ini (Bab I sub-bab 1.1), berhasil dieksekusi dengan hasil yang melampaui ekspektasi (Balanced Accuracy 91,89% pada IMED–IMSFD, +58,45 poin dari baseline *zero-shot*, Rumusan Permasalahan #3), sekaligus menjadi bukti kuantitatif paling kuat dalam laporan ini bahwa skema dua-tahap (KD domain-matched diikuti Transfer Learning) adalah pendekatan yang tepat untuk membangun model FER ringkas yang benar-benar relevan bagi konteks edukasi digital Indonesia. PTQ (Rumusan Permasalahan #4) dan pembaruan model *deployment* di 2GT Smart Classroom (Rumusan Permasalahan #5) keduanya telah dieksekusi dengan hasil dilaporkan pada Bab IV sub-bab 4.6–4.7.

5.2 Saran dan Rekomendasi Penelitian Selanjutnya

Berdasarkan batasan dan ancaman validitas yang telah dilaporkan secara terbuka sepanjang laporan ini (Bab I sub-bab 1.5, Bab III sub-bab 3.2.3 dan sub-bab 3.5.1, Bab IV sub-bab 4.9), berikut rekomendasi konkret bagi kelanjutan penelitian ini maupun penelitian sejenis di masa mendatang:

1. **Validasi independensi subjek penuh (LOSO-CV):** Sebagaimana dicatat sejak Bab I sub-bab 1.5 dan Bab III sub-bab 3.2.3, penelitian ini secara sadar tidak menerapkan *Leave-One-Subject-Out Cross-Validation* mengingat cakupan penelitian yang sudah luas (dua tahap metodologi, kuantisasi, integrasi sistem). Penelitian selanjutnya disarankan menerapkan LOSO-CV, khususnya pada dataset hibrid IMED–IMSFD yang identitas subjeknya diketahui, untuk mengukur generalisasi model secara lebih ketat melampaui protokol *leakage-safe* berbasis klaster visual yang digunakan saat ini.
2. **Pengujian signifikansi statistik formal dan eksperimen *multi-seed*:** Hasil Tahap I (Tabel 4.2) berasal dari satu kali pelatihan per konfigurasi (sub-bab 4.9 poin 2). Penelitian selanjutnya disarankan mengulang pelatihan dengan beberapa *seed* acak berbeda untuk melaporkan rata-rata ± deviasi standar, serta menerapkan uji statistik formal (mis. uji McNemar berpasangan) untuk memvalidasi signifikansi selisih performa *Teacher-Student*.
3. **Menangani kolaps akurasi PTQ INT8 pada MobileNetV3-*Small*:** Pipeline PTQ independen, baik ONNX Runtime QDQ maupun TFLite *full-integer* (Bab III sub-bab 3.5.4–3.5.5, Bab IV sub-bab 4.6), keduanya mengonfirmasi MobileNetV3-Small kolaps drastis pasca-kuantisasi INT8 (−42,28 poin ONNX; −43,52 poin TFLite *native*, Tabel 4.16), sementara Large relatif tahan (−3,22 poin ONNX; kompresi serupa pada TFLite). Konsistensi lintas-*runtime* ini menunjukkan kolaps tersebut adalah karakteristik arsitektur, bukan artefak satu metode kuantisasi. Direkomendasikan: (a) menyelidiki penyebab sensitivitas Small terhadap kuantisasi (mis. per-lapisan mana yang paling terdampak, melalui analisis rentang aktivasi), dan (b) mencoba *Quantization-Aware Training* (QAT) atau eksklusi lapisan sensitif dari kuantisasi sebagai mitigasi.
4. **Pengukuran latensi pada perangkat CPU *edge* sesungguhnya:** Angka latensi yang dilaporkan (Tabel 4.2) diukur pada GPU lingkungan pelatihan (sub-bab 4.9 poin 4), bukan perangkat *edge* CPU yang menjadi motivasi utama penelitian (Bab I sub-bab 1.1). Pengukuran *end-to-end* pada perangkat kelas Chromebook/laptop sekolah standar Indonesia, menggunakan model

INT8 hasil *deployment* aktual, diperlukan untuk memvalidasi klaim kelayakan *real-time*.

5. **Pengujian integrasi langsung (*live testing*) dengan pengguna nyata:** Evaluasi integrasi 2GT Smart Classroom pada laporan ini bersifat struktural/fungsional, bukan studi pengguna (Bab I sub-bab 1.5). Penelitian selanjutnya disarankan melibatkan siswa dan guru nyata dalam sesi pembelajaran nyata untuk mengevaluasi akurasi *real-world* dan latensi yang dirasakan pengguna. Penelitian lanjutan juga disarankan mengimplementasikan formula multimodal *Student Engagement Index* (Persamaan 2.11) secara nyata pada sistem, mencakup komponen M_{att} dan V_{act} yang terpisah dari skor FER, mengingat verifikasi kode (Bab III sub-bab 3.6.3) menunjukkan implementasi saat ini hanya memakai skor FER tunggal berbobot tetap tanpa mengikuti formula tiga-komponen tersebut.
6. **Investigasi lebih lanjut atas domain gap silang-domain***:* Temuan Bab IV sub-bab 4.4.1 menunjukkan penurunan akurasi tajam (39,5–65,7 poin, protokol N=30/kelas identik kedua sisi) ketika keempat model penelitian ini (*Teacher*, kedua *Student* Tahap I, dan *Student* Tahap II) diuji silang ke domain yang tidak dilatihnya (RAF-DB↔IMED–IMSFD). Temuan ini mencakup fakta bahwa *Student* Tahap I Large+KD justru sedikit mengungguli *Teacher* dalam generalisasi ke IMED–IMSFD, serta bahwa *Student* Tahap II (adaptasi domain paling agresif, kesenjangan 65,71 poin saat diuji balik ke RAF-DB) justru menggeneralisasi paling buruk dari keempatnya meski memiliki performa dalam-domain tertinggi. Penelitian lanjutan disarankan menyelidiki teknik regularisasi tambahan (mis. *domain adaptation* eksplisit, augmentasi lintas-domain) untuk mengurangi kesenjangan ini, di luar strategi *transfer learning* satu-domain-sasaran yang diterapkan pada Tahap II penelitian ini.
7. **Perluasan cakupan dataset populasi Indonesia:** Dataset hibrid IMED–IMSFD (Bab III sub-bab 3.2.2) dibentuk dari kombinasi dua dataset dengan jumlah subjek terbatas; penelitian selanjutnya disarankan memperluas variasi subjek, kondisi pencahayaan, dan sudut pengambilan citra populasi Indonesia untuk memperkuat validitas eksternal model hasil Tahap II.

8. **Penyelarasan *pipeline face alignment* dengan RetinaFace 5-titik:** Bab IV sub-bab 4.3 (Tabel 4.5) mengonfirmasi kesenjangan reproduksi *Teacher* (82,40% vs 92,04% resmi) berasal terutama dari perbedaan pipeline *alignment*: *checkpoint* resmi DDAMFN++ dilatih di atas RAF-DB yang di-*realign* lewat RetinaFace (MobileNet0.25) dengan *5-point similarity transform* ke templat pose kanonik (`Retinaface_alignment_ver2.0/face_align_v2.0.py`), sedangkan penelitian ini memakai crop `*_aligned.jpg` bawaan rilis resmi RAF-DB yang lebih sederhana. Karena *ceiling* performa *Student* Tahap I dan, secara berantai, *Student* Tahap II turut dibatasi oleh kekuatan *Teacher* yang direproduksi (sub-bab 4.3), penelitian selanjutnya disarankan menyelaraskan pipeline *alignment* seluruh dataset (RAF-DB maupun IMED–IMSFD) dengan pendekatan RetinaFace 5-titik yang sama sebelum pelatihan ulang, berpotensi mempersempit kesenjangan ini dan menaikkan *ceiling* akurasi kedua tahap sekaligus.