BAB 5
SIMPULAN DAN SARAN

Bab ini menyajikan rangkuman kesimpulan yang ditarik dari hasil penelitian yang telah dilaksanakan. Selain itu, dipaparkan pula serangkaian saran yang ditujukan untuk pengembangan penelitian di masa depan serta untuk implementasi praktis berdasarkan temuan-temuan yang diperoleh.

5.1 Kesimpulan

Berdasarkan analisis dan evaluasi yang telah dilaksanakan, penelitian ini menyimpulkan bahwa metode *ensemble weighted averaging* yang diusulkan mampu meningkatkan performa deteksi *deepfake* secara signifikan jika dibandingkan dengan performa model individual terbaik. Model *ensemble* ini mencapai akurasi sebesar 99,64% pada dataset pengujian, menunjukkan keunggulan 0,44% atas model individual dengan performa tertinggi, yaitu Xception (99,20%). Peningkatan ini dapat berdampak berarti terutama dalam konteks praktis deteksi *deepfake* yang sensitif terhadap *false negative*. Hal ini tercermin pada penurunan angka *false negative* sebesar 78% (dari 46 menjadi 10 kasus) dan *false positive* sebesar 46.5% (dari 114 menjadi 61 kasus).

Implementasi metodologi *ensemble weighted averaging* yang menggunakan formula perhitungan bobot berdasarkan Persamaan 2.24 telah terbukti efektif dalam mengkombinasikan kekuatan dari berbagai arsitektur yang berbeda. Proses perhitungan bobot yang proporsional terhadap validation accuracy masing-masing model memastikan bahwa kontribusi setiap model terhadap prediksi akhir sesuai dengan kemampuannya, sebagaimana dijelaskan dalam contoh perhitungan pada Bab 3.

Salah satu temuan penting dari penelitian ini adalah efektivitas arsitektur Custom CNN yang dirancang secara spesifik untuk tugas ini, yang berhasil mencapai akurasi 98,81%. Performa ini mendekati arsitektur *state-of-the-art* yang lebih kompleks seperti Xception. Hasil ini menggarisbawahi relevansi desain yang berorientasi pada domain spesifik (*domain-specific design*) dalam *deep learning*. Sebaliknya, arsitektur ResNet50 menunjukkan performa yang kurang optimal dengan akurasi 90,10%, mengindikasikan bahwa tidak semua arsitektur *pre-trained* mampu memberikan kinerja terbaik untuk tugas deteksi *deepfake* tanpa adaptasi

lebih lanjut.

Distribusi bobot *ensemble* yang relatif seimbang (berkisar antara 23,4% hingga 25,8%) mengafirmasi bahwa setiap model, termasuk ResNet50 yang berkinerja lebih rendah, memberikan kontribusi komplementer yang esensial terhadap keputusan akhir *ensemble*. Analisis *confusion matrix* menunjukkan bahwa perbedaan karakteristik kesalahan antar model individual memungkinkan *ensemble* untuk mencapai keseimbangan yang lebih baik antara metrik Presisi (Persamaan 2.19) dan Recall (Persamaan 2.20), menghasilkan F1-Score (Persamaan 2.21) tertinggi sebesar 99,65%.

Meskipun demikian, penelitian ini mengidentifikasi keterbatasan signifikan terkait generalisasi model pada evaluasi *cross-dataset*, di mana terjadi penurunan performa mencapai 50%. Hal ini mengindikasikan adanya *overfitting* terhadap karakteristik domain dataset pelatihan yang spesifik pada generator StyleGAN.

Dengan demikian, validasi hipotesis memberikan hasil yang beragam: metode *ensemble* terbukti efektif dalam meningkatkan akurasi dan menunjukkan komplementaritas model pada skenario *single-dataset*, namun kemampuan generalisasinya masih terbatas. Kontribusi utama penelitian ini terletak pada validasi efektivitas metode *ensemble* dalam skenario *high-baseline*, analisis komplementaritas antar model dengan disparitas performa, implementasi sistem dengan metodologi yang transparan dan dapat direproduksi, serta penekanan pada tantangan generalisasi dalam aplikasi dunia nyata.

5.2 Saran

Berdasarkan temuan penelitian serta keterbatasan yang teridentifikasi, berikut adalah beberapa saran untuk pengembangan riset di masa depan dan implementasi praktis:

1. **Peningkatan Generalisasi Model:** Direkomendasikan untuk melakukan pelatihan model menggunakan kombinasi beberapa dataset (misalnya, FaceForensics++, DFDC, CelebDF, dan WildDeepfake) guna meningkatkan kapabilitas generalisasi. Eksplorasi lebih lanjut terhadap teknik-teknik seperti *unsupervised domain adaptation* (UDA) dan *few-shot learning* dapat menjadi fokus untuk memungkinkan adaptasi model yang cepat terhadap domain data yang baru.

2. **Pengembangan *Ensemble* Dinamis:** Penelitian selanjutnya dapat berfokus

pada pengembangan metode *ensemble* dinamis yang melampaui pendekatan *weighted averaging* statis yang digunakan dalam penelitian ini. Sistem semacam ini akan secara adaptif memilih sub-himpunan model atau menyesuaikan bobot kontribusi berdasarkan karakteristik data masukan, misalnya melalui mekanisme seperti *input-dependent model selection* atau *confidence-based weighting*. Hal ini dapat mengoptimalkan formula pada Persamaan 2.12 dengan bobot yang bersifat dinamis.

3. **Optimisasi untuk Implementasi Praktis:** Untuk memfasilitasi penerapan di lingkungan produksi, disarankan untuk mengimplementasikan teknik *knowledge distillation*. Teknik ini bertujuan mentransfer pengetahuan dari model *ensemble* yang kompleks ke sebuah model tunggal yang lebih efisien. Selain itu, teknik kompresi model seperti *pruning* dan *quantization* perlu dieksplorasi untuk mengurangi beban komputasi tanpa degradasi akurasi yang signifikan.

4. **Evaluasi Robustisitas Komprehensif:** Perlu dilakukan pengujian yang lebih luas untuk mengevaluasi robustisitas sistem terhadap berbagai tantangan, termasuk serangan adversarial (*adversarial attacks*), variasi kualitas dan tingkat kompresi citra, serta skenario penerapan di dunia nyata. Evaluasi ini harus mencakup analisis performa pada berbagai metrik yang telah didefinisikan dalam Persamaan 2.18 hingga **??** untuk memastikan keandalan sistem dalam kondisi operasional yang beragam.

5. **Analisis Lanjutan Model Individual:** Disarankan untuk melakukan investigasi mendalam terhadap performa model individual. Analisis ini dapat mencakup visualisasi *feature map* untuk memahami penyebab kinerja ResNet50 yang kurang optimal, serta melakukan eksperimen dengan berbagai strategi *fine-tuning*. Di sisi lain, studi ablasi (*ablation study*) terhadap komponen arsitektur Custom CNN dapat memberikan wawasan mengenai faktor-faktor yang berkontribusi pada efektivitasnya yang tinggi.

6. **Pengembangan Metodologi Ensemble Lanjutan:** Berdasarkan keberhasilan implementasi *weighted averaging* yang dijelaskan dalam kerangka metodologi penelitian, penelitian selanjutnya dapat mengeksplorasi metode ensemble yang lebih sophisticated seperti *stacking*, *boosting*, atau *mixture of experts*. Setiap metode dapat dievaluasi menggunakan kerangka kerja yang sama untuk membandingkan efektivitasnya.

7. **Implementasi Pembelajaran Berkelanjutan (*Continual Learning*):** Mengingat pesatnya evolusi teknik pembuatan *deepfake*, perlu dikembangkan sebuah kerangka kerja pembelajaran berkelanjutan. Sistem ini harus mampu beradaptasi dan mempelajari pola-pola *deepfake* baru secara inkremental tanpa mengalami *catastrophic forgetting* terhadap pengetahuan yang telah dimiliki, sambil mempertahankan keseimbangan metrik evaluasi yang optimal.

8. **Pengembangan Aplikasi Dunia Nyata:** Sistem deteksi yang diusulkan dapat diintegrasikan ke dalam platform nyata, seperti sistem moderasi konten media sosial, platform verifikasi berita, atau alat analisis forensik digital. Proses integrasi ini harus mempertimbangkan aspek pengalaman pengguna (*user experience*) dan menyertakan mekanisme pemantauan performa secara kontinu dalam lingkungan produksi. Implementasi harus mencakup pipeline lengkap sesuai dengan metodologi yang telah dikembangkan dalam penelitian ini.

9. **Pertimbangan Aspek Etis dan Keamanan:** Implementasi sistem harus diiringi dengan pertimbangan etis yang matang. Strategi mitigasi bias perlu diterapkan untuk mengurangi potensi diskriminasi demografis. Selain itu, teknik perlindungan privasi seperti *federated learning* dan pengembangan kerangka kerja transparansi untuk melacak jejak audit (*audit trail*) dari setiap keputusan deteksi menjadi krusial.

10. **Standardisasi Metrik dan Benchmarking:** Berdasarkan pengalaman dalam penelitian ini, disarankan untuk mengembangkan standar evaluasi yang konsisten dalam domain deteksi deepfake. Hal ini mencakup standardisasi penggunaan metrik evaluasi yang telah didefinisikan dalam Bab 2, protokol pengujian cross-dataset, dan benchmark yang dapat diandalkan untuk perbandingan yang adil antar metode yang berbeda.