BAB 5
SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil dan evaluasi model dalam penelitian ini, dapat disimpulkan beberapa hal berikut:

1. Penerapan *Adversarial Training* terbukti meningkatkan ketahanan (*robustness*) model EfficientNet-B0 terhadap serangan *Projected Gradient Descent* (PGD). Model baseline menunjukkan akurasi sangat tinggi pada data bersih sebesar 98,7%, tetapi mengalami penurunan ekstrem hingga 0,0% pada seluruh iterasi PGD, menandakan ketidakmampuannya mempertahankan prediksi terhadap perturbasi adversarial. Model AT mempertahankan akurasi relatif stabil pada data adversarial, yaitu 49,7% untuk PGD-5, PGD-10, dan PGD-20, meskipun akurasi pada data bersih lebih rendah dibandingkan baseline. Model Mixed AT menunjukkan kinerja seimbang dengan mempertahankan akurasi tinggi pada data bersih sebesar 95,2%, sekaligus mencapai akurasi 50,4%, 49,5%, dan 32,2% pada PGD-5, PGD-10, dan PGD-20. Peningkatan ketahanan ini dapat dijelaskan karena integrasi contoh adversarial dalam pelatihan memaksa model AT dan Mixed AT belajar representasi fitur yang lebih stabil dan tahan terhadap perturbasi, sehingga prediksi tetap konsisten pada data adversarial. Temuan ini secara empiris membuktikan bahwa *adversarial training* efektif meningkatkan *robustness* EfficientNet-B0.

2. Perbandingan performa model EfficientNet-B0 dengan dan tanpa *adversarial training* menunjukkan perbedaan yang signifikan dalam mendeteksi citra *deephoax* pada data bersih dan data yang telah mengalami serangan PGD. Model tanpa AT mencapai akurasi 98,7% pada data bersih, namun akurasinya turun drastis menjadi 0,0% pada semua konfigurasi serangan PGD, menunjukkan ketahanan yang sangat rendah terhadap perturbasi adversarial. Sebaliknya, model AT mempertahankan akurasi yang relatif stabil pada data adversarial sebesar 49,7% pada PGD-5 dan PGD-10, dan PGD-20, meskipun akurasi pada data bersih menurun menjadi 49,7%. Model Mixed AT menawarkan keseimbangan yang lebih baik dengan mempertahankan akurasi tinggi pada data bersih sebesar 95,2% dan akurasi

kompetitif pada serangan PGD dengan iterasi rendah hingga menengah, sebelum menurun pada PGD-20 menjadi 32,2%. Hasil ini menegaskan adanya *trade-off* antara akurasi pada data bersih dan ketahanan terhadap serangan PGD, di mana peningkatan *robustness* dicapai dengan konsekuensi penurunan performa pada kondisi normal.

5.2 Saran

Berdasarkan temuan penelitian ini, beberapa rekomendasi untuk pengembangan selanjutnya adalah:

1. Mengeksplorasi ketahanan model terhadap berbagai jenis serangan adversarial, termasuk serangan *white-box* seperti *FGSM*, *C&W*, dan *AutoAttack*, serta serangan *black-box*, untuk memahami performa model terhadap ancaman yang berbeda.
2. Mengembangkan dan menerapkan teknik pertahanan lain terhadap serangan adversarial untuk mengeksplorasi peningkatan robustnes model.
3. Mengeksplorasi pengembangan deteksi *deephoax* pada domain lain, seperti video dan audio, untuk menguji ketahanan model terhadap serangan adversarial di berbagai jenis media.