BAB 5
SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, penelitian ini berhasil mengevaluasi kinerja dua algoritma *machine learning*, yaitu *Decision Tree* sebagai model dasar (*baseline*) dan Random Forest sebagai metode ensemble dalam mendeteksi ancaman siber terutama serangan *phishing*. Fikus utama penelitian ini adalah mengenali struktur dan pola URL *phishing* berdasarkan karakteristik URL (*lexical features*). Hasil akhir penelitian menunjukkan bahwa *Random Forest* dengan metode ensemble terbukti mampu menghasilkan nilai akurasi lebih tinggi dan kinerja generalisasi yang lebih stabil dibandingkan model *Decision Tree* dalam mengklasifikasikan URL berbahaya dan mencurigakan.

Berikut adalah kesimpulan dalam penelitian ini secara lebih rinci:

1. Dataset yang digunakan dalam penelitian ini bersumber dari Kaggle milik Shashwat Tiwari dan telah divalidasi memiliki kualitas data yang baik dengan fitur yang lengkap. Tidak ditemukan nilai hilang (*missing values*) pada atribut data, namun pada tahap pemeriksaan awal teridentifikasi adanya redudansi data berupa duplikat sebanyak 401 baris yang telah ditangani.

2. Tahapan Data Preprocessing seleksi fitur terbukti berperan penting dalam meningkatkan kualitas dataset. Pemilihan 26 fitur meliputi 23 fitur karakteristik URL dan 3 fitur eksternal reputasi domain seperti google_index dan page_rank berhasil membentuk atribut yang relevan serta kombinasi tersebut memberi kontribusi signifikan bagi model dalam membedakan pola antara URL *legitimate* dan *phishing*. Tanpa adanya 3 fitur eksternal hanya akan menurukan performa deteksi karena pada dasarnya sebuah situs URL pasti memiliki *google index* dan *page rank*, sehingga penggunaan 3 fitur eksternal ini merupakan *metadata* yang melekat pada identitas sebuah URL, bukan fitur berbasis konten.

3. Hasil evaluasi model *Decision Tree* sebagai model *baseline* menunjukkan kinerja yang cukup baik dengan akurasi sebesar 92.52%. Meskipun model ini unggul dalam sisi kecepatan dalam pelatihan dan interpretabilitas struktur pohon, namun ditemukan adanya indikasi *overfitting*. Hal ini ditandai dengan

selisih yang jauh antara data latih dan data uji. Hal ini dipastikan *decision tree* yang memiliki pohon keputusan tunggal cenderung memiliki varians tinggi dan kurang stabil dalam melakukan pemahaman terhadap data baru.

4. Algoritma *Random Forest* yang disertai *hyperparameter tuning* berhasil mengatasi kelemahan model *baseline*. Model final *Random Forest* yang terdiri dari banyak pohon keputusan mencatatkan peningkatan kinerja di seluruh bagian *confusion matrix* dengan nilai Akurasi 95.42%, *Precision* 95.34%, *Recall* 95.43%, dan *F1-score* 95.38%. Hal ini membutikan bahwa metode *ensemble learning* efektif mereduksi varians, sehingga menghasilkan prediksi yang lebih stabil dan konsisten terhadap data baru. Sehingga penetapan algoritma Random Forest lebih unggul dalam mendeteksi karakteristik URL dibandingkan algoritma Decision Tree. Dengan peningkatan akurasi dari 92.52% pada Decision Tree menjadi 95.42% pada Random Forest dengan selisih 2.9% (atau dibulatkan 3%).

5. Berdasarkan prespektif keamanan siber, *Random Forest* menunjukkan keunggulan yang krusial dengan nilai *Recall* mencapai 95.43%. Nilai yang tinggi ini membuktikan bahwa model sangat efektif dalam minimalisir tingkat *false negative* (serangan yang lolos deteksi). Kemampuan ini menjadikan *random forest* solusi yang jauh lebih baik dan aman untuk diterapkan dalam sistem deteksi URL phishing dibandingkan berbasis pohon keputusan tunggal.

5.2 Saran

Penelitian ini telah berhasil membangun model deteksi phishing dengan kinerja yang baik dengan algoritma *Random Forest* yang memiliki hasil lebih baik. Untuk pengembangan pada penelitian selanjutnya mengenai deteksi URL *phishing* berbasis karakteristik URL, berikut adalah beberapa saran yang dapat dipertimbangkan antara lain:

1. Mengingat komputasi *Random Forest* dengan 26 fitur tergolong ringan, disarankan untuk mengimplementasikan model ini menjadi *browser extension* ataupun berbasis website. Hal ini memungkinkan deteksi URL berbahaya secara langsung dan cepat saat pengguna berselancar, tanpa membebani kinerja perangkat.

2. Disarankan untuk menambahkan metode NLP dalam ekstraksi fitur guna menangkap konteks semantik string URL. Pendekatan ini efektif mendeteksi teknik *typosquatting* (seperti ”paypa1.com” atau ”g0ogle.com”) yang sering lolos jika hanya mengandalkan statistik karakter leksikal.

3. Mengingat pola serangan *phishing* berubah dengan cepat, disarankan untuk memperbarui dataset secara rutin menggunakan sumber data terkini (seperti PhishTank). Hal ini memastikan model *Random Forest* tetap adaptif dan relevan dalam mengenali tren ancaman terbaru.

4. Penelitian selanjutnya disarankan untuk melakukan evaluasi *robustness* (ketahanan) dengan menguji model terhadap skenario serangan tersebut. Penggunaan algoritma yang lebih kompleks seperti *Deep Learning* (CNN atau LSTM) juga dapat dipertimbangkan untuk meningkatkan kemampuan model dalam mengenali pola serangan *non-linear* yang rumit.

5. Model berbasis karakteristik URL memiliki keterbatasan dalam mendeteksi phishing pada domain resmi yang diretas (*compromised domains*). Penelitian selanjutnya sebaiknya menggabungkan analisis URL dengan analisis konten (kode HTML atau visual logo) untuk menutup celah deteksi tersebut.