BAB I
PENDAHULUAN

A. Latar Belakang

Perkembangan teknologi informasi yang pesat telah membawa transformasi besar dalam berbagai aspek kehidupan, mulai dari komunikasi, transaksi keuangan, hingga layanan pemerintahan berbasis daring. Sayangnya, kemajuan ini juga diiringi oleh peningkatan keamanan siber, salah satunya adalah serangan *phishing*. *Phishing* merupakan bentuk kejahatan siber yang berusaha mencuri data sensitif pengguna melalui situs *web* tiruan yang menyerupai situs resmi. Menurut data *Anti-Phishing Working Group* (*APWG*) , pada Q4 2024 terdapat 989.123 serangan *phishing*, naik dari 877.536 pada Q2 2024 dan 932.923 di Q3 2024, mencerminkan pertumbuhan serangan yang signifikan dibandingkan kuartal-kuartal sebelumnya [1].

Serangan *phishing* berbasis *URL* tetap menjadi ancaman siber yang signifikan karena teknik ini mudah diimplementasikan dan sering kali menyerupai struktur situs *web* yang sah, sehingga sulit dikenali oleh pengguna. *URL* palsu memanfaatkan elemen-elemen seperti panjang *domain*, jumlah *subdomain*, serta informasi *host* untuk mengecoh pengguna. Pendekatan tradisional seperti *blacklist* memiliki keterbatasan karena hanya mendeteksi *URL phishing* yang sudah diketahui sebelumnya. Sebaliknya, pendekatan berbasis *machine learning* menawarkan solusi yang lebih adaptif dengan kemampuan menganalisis pola dan fitur *URL* yang kompleks guna mengidentifikasi potensi serangan *phishing* yang baru muncul [2].

Salah satu pendekatan yang telah digunakan dalam mendeteksi *phishing* adalah *deep learning*. Penelitian sebelumnya mengembangkan model berbasis *Convolutional Neural Network* (*CNN*) untuk mendeteksi *URL phishing* dan menunjukkan bahwa model tersebut mampu menghasilkan performa deteksi yang baik. Namun, penggunaan model *deep learning* umumnya memerlukan sumber daya komputasi yang lebih besar serta waktu pelatihan yang relatif lebih lama dibandingkan metode *machine learning* berbasis pohon keputusan [3].

Berbeda dengan pendekatan tersebut, penelitian ini berfokus pada perbandingan dua algoritma *ensemble learning*, yaitu *LightGBM* dan *Random*

Forest, yang dikenal memiliki performa klasifikasi yang baik serta efisiensi komputasi yang tinggi. *LightGBM* merupakan algoritma berbasis *gradient boosting* yang dirancang untuk mempercepat proses pelatihan dan prediksi pada *dataset* berskala besar tanpa mengurangi performa klasifikasi [4]. Sementara itu, *Random Forest* merupakan algoritma berbasis *bagging* yang memiliki kemampuan generalisasi yang baik, tahan terhadap *overfitting*, serta banyak digunakan dalam berbagai penelitian klasifikasi karena menghasilkan performa yang stabil [5].

Berdasarkan penelitian-penelitian terdahulu, sebagian besar penelitian lebih berfokus pada pengembangan model deteksi *phishing* menggunakan satu algoritma tertentu atau membandingkan algoritma yang berbeda tanpa mempertimbangkan aspek efisiensi komputasi secara lebih mendalam. Selain itu, penelitian yang menggunakan pendekatan *deep learning* umumnya memerlukan sumber daya komputasi yang lebih besar. Oleh karena itu, penelitian ini melakukan analisis perbandingan antara algoritma *LightGBM* dan *Random Forest* pada kasus deteksi *phishing* berbasis URL dengan mengevaluasi performa menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian ini diharapkan dapat menjadi referensi dalam menentukan algoritma yang tidak hanya memiliki performa klasifikasi yang baik, tetapi juga efisien untuk diimplementasikan pada sistem deteksi *phishing* berbasis *web*.

B. Perumusan Masalah

Dari penjelasan latar belakang di atas, rumusan masalah dalam penelitian ini adalah "Bagaimana perbandingan performa algoritma *LightGBM* dan *Random Forest* dalam mendeteksi *phishing* berbasis *URL* berdasarkan *accuracy*, *precision*, *recall*, dan *F1-score*??"

C. Batasan Penelitian

Penelitian ini akan membatasi, cakupannya untuk menghindari pembahasan yang terlalu luas. Berikut adalah batasan yang diberlakukan dalam penelitian ini :

1. Data yang digunakan berasal dari sumber publik *Kaggle* tahun 2021 yang berupa *URL phishing* dan *legitimate*.
2. Algoritma yang digunakan adalah algoritma *LightGBM* dan *Random Forest*.

3. Penelitian ini hanya berfokus pada deteksi *phishing* berbasis struktur *URL* dan tidak mencakup masalah *typography URL*, analisis isi halaman *website*, status *hosting*, *HTTP* response, reputasi *domain*, pengecekan aktif atau tidaknya *website* secara real-time, maupun kemiripan tampilan *website* tiruan terhadap layanan asli.
4. Evaluasi performa model hanya dilakukan berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, untuk membandingkan kinerja kedua algoritma tersebut.

D. Tujuan Penelitian

Penelitian ini bertujuan Untuk menganalisis perbandingan performa algoritma *LightGBM* dan *Random Forest* dalam mendeteksi *phishing* berbasis *URL* berdasarkan *accuracy*, *precision*, *recall*, dan *F1-score*, sehingga dapat diketahui algoritma yang memiliki kinerja lebih optimal dalam mengklasifikasikan *URL phishing* dan *legitimate*.

E. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat baik secara praktis maupun konseptual, dengan beberapa hasil penelitian yang dapat diterapkan. Manfaat dari penelitian ini meliputi:

1. Bagi peneliti dan akademisi, penelitian ini diharapkan menjadi referensi mengenai penerapan algoritma *machine learning*, khususnya *LightGBM* dan *Random Forest*, dalam mendeteksi *phishing* berbasis URL serta sebagai acuan bagi penelitian selanjutnya di bidang keamanan siber.
2. Bagi pembaca dan masyarakat, penelitian ini diharapkan dapat meningkatkan kesadaran terhadap ancaman *phishing* serta membantu masyarakat melakukan pemeriksaan awal terhadap *URL* yang mencurigakan melalui aplikasi *web* deteksi *phishing*, sehingga dapat mengurangi risiko pencurian data pribadi maupun informasi sensitif.