BAB I
PENDAHULUAN

Pada bab ini dijelaskan tentang pendahuluan pada penelitian ini. Pada bab ini dijelaskan latar belakang dari penelitian ini. diikuti dengan beberapa aspek seperti rumusan masalah, tujuan penelitan, manfaat penelitian, serta batasan masalah, untuk memperkuat landasan pada penelitan ini.

1.1. Latar Belakang

Phishing merupakan salah satu bentuk kejahatan dunia maya yang bertujuan untuk mencuri informasi sensitif, seperti *kredensial login*, data pribadi, atau informasi keuangan, dengan menyamar sebagai entitas yang tepercaya. Serangan *phishing* biasanya dilakukan melalui situs web palsu yang dirancang menyerupai situs resmi untuk mengelabui pengguna agar memberikan informasi mereka secara sukarela. Serangan ini terus berkembang dan menjadi ancaman serius bagi individu maupun organisasi, karena pelaku *phishing* terus menyempurnakan teknik mereka untuk menghindari deteksi. Oleh karena itu, diperlukan metode yang lebih *adaptif* dan efisien dalam mendeteksi situs *phishing* secara *real-time* [1].

Pendekatan tradisional dalam mendeteksi *phishing*, seperti penggunaan daftar hitam (*blacklist*), telah lama diterapkan. Namun, metode ini memiliki keterbatasan yang signifikan, terutama dalam mendeteksi situs *phishing* baru yang belum terdaftar. Sistem berbasis *blacklist* bergantung pada pembaruan manual, yang memakan waktu dan sering kali tidak efektif dalam mengatasi ancaman baru yang muncul dengan cepat [2].

Oleh karena itu, penggunaan pendekatan berbasis *machine learning* menjadi solusi yang lebih menjanjikan, karena dapat mempelajari pola dan karakteristik situs *phishing* berdasarkan data historis serta mengidentifikasi ancaman baru yang memiliki karakteristik serupa. Dalam beberapa tahun terakhir, berbagai algoritma *machine learning* telah diterapkan dalam deteksi *phishing*, termasuk *Random Forest*, *K-Nearest Neighbor*, dan *Support Vector Machine (SVM)*. Penelitian sebelumnya menunjukkan bahwa penggunaan teknik seleksi fitur dapat meningkatkan akurasi model dalam mendeteksi *phishing* [2].

Metode *Ensemble Voting*, yang menggabungkan beberapa algoritma *pembelajaran mesin* serta analisis *multi-fitur*, telah terbukti mampu meningkatkan efektivitas deteksi. Beberapa studi telah menguji berbagai pendekatan dalam meningkatkan akurasi deteksi *phishing*. Penelitian [3] mengombinasikan algoritma *Random Forest* dan *C4.5*, yang menghasilkan tingkat akurasi mendekati 97%. Penelitian [4] mengimplementasikan teknik *Logistic Regression* untuk mengekstrak fitur penting dari *URL*, meningkatkan presisi deteksi hingga 95%.

Sementara itu, penelitian [5] mengadopsi *Support Vector Machine* untuk menganalisis fitur tekstual dan visual situs *phishing*, yang menghasilkan peningkatan skor *F1* hingga 94%. Studi-studi ini menegaskan bahwa kombinasi berbagai metode dapat meningkatkan efektivitas sistem dalam mengidentifikasi situs *phishing* dengan lebih baik. Penelitian ini berfokus pada pengembangan model *ensemble machine learning* yang menggabungkan tiga algoritma utama: *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)*. Setiap algoritma memiliki keunggulan tersendiri, seperti *Logistic Regression* yang memiliki kecepatan komputasi yang tinggi, *Random Forest* yang memiliki kemampuan klasifikasi yang kuat, dan *SVM* yang memiliki akurasi tinggi dalam memisahkan data *non-linear*. Model ini dirancang untuk meningkatkan deteksi situs *phishing* melalui analisis *multi-fitur*, termasuk fitur *URL* dan *metadata* seperti sertifikat *SSL*. Dengan mengintegrasikan metode ini, penelitian tersebut diharapkan dapat meningkatkan akurasi dan efisiensi deteksi *phishing* secara signifikan.

Dalam penerapan *machine learning* untuk deteksi *phishing*, pemilihan algoritma yang tepat bukan satu-satunya faktor yang menentukan keberhasilan model. Optimalisasi *hyperparameter,* juga berpengaruh dalam meningkatkan kinerja model. *Hyperparameter tuning* memungkinkan model untuk menyesuaikan parameter seperti jumlah *estimators* dalam *Random Forest*, nilai *regularisasi* dalam *Logistic Regression*, serta jenis *kernel* dalam *SVM*. Salah satu metode yang paling umum digunakan dalam penelitian ini adalah *Grid Search Cross Validation* (*GridSearchCV*) [6], yang memungkinkan pencarian kombinasi parameter terbaik berdasarkan performa model pada data validasi. Dengan menggunakan *GridSearchCV*, model dapat menghindari *underfitting* dan *overfitting*, sehingga

meningkatkan akurasi dan generalisasi terhadap data baru.

Evaluasi kinerja model akan dilakukan menggunakan metrik seperti *akurasi*, *presisi*, *recall*, dan *F1-score*. *Dataset* yang digunakan mencakup situs *phishing* dan *non-phishing* yang diperoleh dari *Kaggle*. Pendekatan *ensemble* ini, dengan kombinasi prediksi dari ketiga algoritma, bertujuan untuk mengurangi *bias* serta meningkatkan kemampuan model dalam mendeteksi variasi taktik *phishing* yang terus berkembang. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan keamanan siber, membantu pengguna menghindari situs *phishing*, serta melindungi infrastruktur digital organisasi dari serangan yang semakin kompleks..

1.2. Rumusan Masalah

Dalam melakukan penelitian ini, rumusan masalah yang dikemukakan adalah seperti berikut:

1. Bagaimana implementasi algoritma *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)* dalam model *ensemble* untuk mendeteksi situs *phishing*?
2. Bagaimana efektivitas penerapan model *ensemble* dalam mendeteksi situs *phishing* dibandingkan dengan metode *Random Forest*, *Logistic Regression*, dan *Support Vector Machine (SVM)* secara independen?
3. Bagaimana pengaruh *hyperparameter tuning GridSearchCV* pada parameter algoritma *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)*?

1.3. Tujuan Penelitian

Adapun penelitian ini memiliki tujuan sebagai berikut:

1. Untuk mengimplementasikan model *ensemble* dengan algoritma *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)* dalam mendeteksi situs *phishing*.
2. Untuk mengevaluasi performa model *ensemble* yang mengintegrasikan algoritma *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)* dalam deteksi situs

phishing dibandingkan dengan penggunaan masing-masing algoritma secara independen.

3. Untuk menganalisis pengaruh *hyperparameter tuning* menggunakan *GridSearchCV* terhadap kinerja algoritma *Logistic Regression*, *Random Forest*, dan *Support Vector Machine (SVM)* dalam deteksi situs *phishing*.

1.4. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

Penelitian tentang deteksi situs *phishing* menggunakan teknologi *Machine Learning* memberikan manfaat yang signifikan baik bagi akademisi maupun praktisi di bidang keamanan siber. Untuk kalangan akademisi, hasil penelitian ini memberikan kontribusi penting ke dalam literatur akademik, menyediakan referensi berharga untuk pengembangan sistem deteksi *phishing* yang lebih lanjut dan membuka peluang untuk eksplorasi teknologi deteksi yang baru. Penelitian ini juga menawarkan wawasan mendalam tentang pengaruh fitur-fitur tertentu terhadap efektivitas deteksi, mendorong inovasi dan eksperimen lebih lanjut dalam peningkatan metode deteksi.

Bagi praktisi, khususnya yang bekerja di perusahaan teknologi dan keamanan siber, penelitian ini menawarkan solusi praktis yang dapat diimplementasikan dalam operasional sehari-hari untuk meningkatkan keamanan data. Sistem yang dikembangkan mampu melakukan deteksi *phishing* secara *real-time*, memungkinkan perusahaan untuk mendeteksi dan merespons ancaman *phishing* dengan cepat dan akurat. Implementasi sistem ini tidak hanya meningkatkan keamanan data perusahaan tetapi juga mengurangi risiko kerugian finansial dan kerusakan reputasi akibat serangan siber.

Secara Keseluruhan, penelitian ini berpotensi memberikan kontribusi yang signifikan dalam mengatasi tantangan keamanan siber yang semakin berkembang, dengan menawarkan wawasan baru dan solusi konkret untuk menghadapi dan mencegah serangan *phishing* di masa depan.

1.5. Batasan Masalah

Agar penelitian ini lebih terfokus dan tidak meluas dari pembahasan

yang dimaksudkan, maka penelitian ini membataskan beberapa macam batasan masalah, antara lain:

1. Penelitian ini hanya fokus pada deteksi situs *phishing* berbasis web dan tidak mencakup serangan *phishing* melalui media lain seperti *email* atau pesan instan.
2. *Dataset* yang digunakan dalam penelitian ini terbatas pada data fitur *URL* yang telah dikumpulkan dari sumber terbuka dan bersifat statis selama proses eksperimen berlangsung.
3. Evaluasi model hanya dilakukan berdasarkan metrik performa klasifikasi, yaitu *akurasi*, *presisi*, *recall*, *F1-score*, dan *ROC-AUC*, tanpa mempertimbangkan faktor eksternal seperti efisiensi komputasi atau waktu proses yang dibutuhkan oleh model.
4. Penelitian ini tidak melakukan pengujian implementasi model secara langsung pada lingkungan nyata atau aplikasi *online*, sehingga hasil yang diperoleh terbatas pada simulasi dan pengujian secara *offline*.

Halaman ini sengaja dikosongkan