## ****Draft Proposal Skripsi****

### ****Judul Penelitian****

Klasifikasi Web Phishing Berbasis URL dengan XGBoost menggunakan Metode Hybrid Feature Selection dan Optimasi Model dengan Cross-Validation

### ****1. Objek Penelitian****

Objek penelitian ini adalah **websites** yang dikategorikan sebagai **legitimate** atau **phishing** berdasarkan analisis URL. Dataset yang digunakan dalam penelitian ini akan mencakup berbagai fitur URL seperti panjang URL, penggunaan IP address, domain authority, dan beberapa indikator lainnya yang dapat membantu dalam menentukan apakah sebuah website adalah phishing atau tidak.

### ****2. Rumusan Masalah****

Penelitian ini akan mencoba menjawab beberapa pertanyaan penting:

1. Bagaimana cara mengimplementasikan metode hybrid feature selection untuk memilih fitur-fitur terbaik dalam proses klasifikasi web phishing?
2. Sejauh mana algoritma XGBoost dapat meningkatkan akurasi klasifikasi web phishing berbasis URL setelah dilakukan pemilihan fitur dengan metode hybrid feature selection?
3. Bagaimana pengaruh optimasi model menggunakan cross-validation terhadap performa model XGBoost dalam mendeteksi phishing?

### ****3. Tujuan Penelitian****

Tujuan utama dari penelitian ini adalah:

1. Mengembangkan model klasifikasi yang akurat untuk mendeteksi web phishing berbasis URL dengan menggunakan algoritma XGBoost.
2. Mengimplementasikan dan mengevaluasi metode hybrid feature selection untuk meningkatkan performa model XGBoost.
3. Melakukan optimasi model menggunakan cross-validation untuk memastikan bahwa model memiliki generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya.

### ****4. Batasan Penelitian****

Penelitian ini dibatasi oleh beberapa faktor, di antaranya:

1. **Dataset**: Penelitian ini hanya akan menggunakan dataset yang tersedia secara publik dan memiliki fitur-fitur terkait URL.
2. **Lingkup**: Penelitian ini hanya fokus pada klasifikasi phishing berdasarkan URL tanpa memperhitungkan konten halaman web atau data pengguna.
3. **Algoritma**: Algoritma yang digunakan untuk klasifikasi adalah XGBoost, dan metode hybrid feature selection untuk pemilihan fitur.
4. **~~Implementasi~~**~~: Implementasi model terbatas pada lingkungan simulasi dan belum diuji pada data real-time.~~

### ****5. Metode Pengumpulan Data****

Data yang digunakan dalam penelitian ini akan dikumpulkan dari berbagai sumber dataset publik yang berisi URL yang telah dikategorikan sebagai **phishing** atau **legitimate**. Beberapa metode pengumpulan data meliputi:

1. **Penggunaan dataset publik**: Mengakses dataset publik dari platform seperti UCI Machine Learning Repository, kaggle, ataupun PhishTank.
2. **Preprocessing**: Melakukan preprocessing data untuk mengekstrak fitur-fitur URL yang relevan seperti panjang URL, karakteristik domain, dan parameter lainnya.

### ****6. Metode Analisis Data****

Metode analisis data dalam penelitian ini akan dilakukan melalui beberapa tahapan:

1. **Exploratory Data Analysis (EDA)**: Untuk memahami karakteristik dataset, distribusi fitur, dan hubungan antara fitur.
2. **Hybrid Feature Selection**: Menggabungkan metode filter, wrapper, dan embedded untuk memilih subset fitur yang optimal.
3. **Implementasi XGBoost**: Membangun model klasifikasi menggunakan XGBoost dengan fitur yang telah dipilih.
4. **Cross-Validation**: Melakukan cross-validation untuk mengoptimalkan parameter model dan mengevaluasi performanya.
5. **Evaluasi Model**: Mengukur kinerja model dengan metrik seperti akurasi, precision, recall, dan F1-score.

### ****7. Relevansi dan Kontribusi Penelitian****

Penelitian ini diharapkan memberikan kontribusi signifikan dalam bidang keamanan siber, khususnya dalam deteksi web phishing yang berbasis URL. Dengan menggabungkan metode hybrid feature selection dan optimasi model menggunakan cross-validation, penelitian ini berusaha untuk mengembangkan model yang lebih akurat dalam mendeteksi web phishing. Hasil penelitian ini juga dapat menjadi referensi bagi pengembangan sistem deteksi phishing real-time yang dapat diaplikasikan dalam industri.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author | Tittle | Method | Purpose | Conclusion |
| Korkmaz, M., Sahingoz, O. K., & Diri, B. | Detection of Phishing Websites by Using  Machine Learning-Based URL Analysis | In the machine learning based system, 8 different  algorithms were run in the experiment. These are: Logistic  Regression (LR), K-Nearest Neighborhood (KNN), Support  Vector Machine (SVM), Decision Tree (DT), Naive Bayes  (NB), XGBoost, Random Forest (RF) and Artificial Neural  Network (ANN). The models created with these algorithms  trained by using the Sklearn library in the Python  programming language. | To implement a phishing detection system using various machine learning algorithms.  To test the proposed systems with recent datasets available in the literature.  To compare the results with the latest works in the literature to evaluate the efficiency and accuracy of phishing detection. | The proposed systems improve the efficiency of phishing detection and achieve very good accuracy rates.  Future work aims to create a new, large dataset for URL-based phishing detection systems.  Plans include enhancing the system using hybrid algorithms and deep learning models as mentioned in reference. |
| Agus Fatkhurohman, Eli Pujastuti | Penerapan Algoritma Naïve Bayes Classifier Untuk  Meningkatkan Keamanan Data  Dari Website Phising | Address  Bar based Feature, Abnormal based Feature,  HTML and Javascript based Features dan  Domain based Feature |  |  |
| Vikky Aprelia Windarni, Anggit Ferdita Nugraha , Surya Tri Atmaja Ramadhani ,  Dewi Anisa Istiqomah , Fiyas Mahananing Puri , Adi Setiawan | DETEKSI WEBSITE PHISHING MENGGUNAKAN TEKNIK FILTER PADA  MODEL MACHINE LEARNING | Naïve Bayes, Decision Tree,Random Forest | Tujuan dari penelitian ini  adalah peneliti ingin menggunakan machine learning dengan memanfaatkan fitur filter yang ada didalamnya yaitu  pearson correlation dan menerapkan 3 metode Naïve Bayes, Decision Tree dan Random Forest untuk menentukan  metode yang paling efektif dalam mendeteksi web phishing | . Dari hasil penelitian  ini didapatkan bahwa penerapan metode Naïve Bayes memiliki nilai akurasi sebesar 60,4%, metode Decision Tree  memiliki nilai akurasi 94,4% dan metode Random Forest memiliki akurasi sebesar 96,3%. Sehingga dapat  disimpulkan bahwa metode yang paling efektif untuk mendeteksi web phishing adalah menggunakan Random  Forest karena memiliki tingkat akurasi sebesar 96.3%. |
| Pungkas Subarkah, Ali Nur Ikhsan | IDENTIFIKASI WEBSITE PHISHING MENGGUNAKAN ALGORITMA  CLASSIFICATION AND REGRESSION TREES (CART) | Algoritma Classification  and Regression Trees (CART). | mengidentifikasi dan mendeteksi situs web yang terindikasi sebagai phishing dengan menggunakan algoritma Classification and Regression Trees (CART). Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma tersebut dalam mengklasifikasikan situs web phishing berdasarkan pola-pola yang diekstraksi dari data. Data yang digunakan dalam penelitian ini diambil dari Phishing Website Data Set yang tersedia secara publik di UCI. | Berdasarkan hasil perhitungan yang  sudah dilakukan menggunakan Algoritma  CART pada dataset website phishing  diperoleh nilai akurasi sebesar 95.28%,  dengan rincian nilai precision sebesar  0.953%, nilai recall sebesar 0.953% dan nilai  F-Measure sebesar 0.953%. Hal ini sejalan  dari  pendapat referensi [17] bahwa  Algoritma CART cocok digunakan untuk  data numeric dan data yang berjumlah  besar. |
| Rami M. Mohammad, Fadi Thabtah, Lee McCluskey | Intelligent rule-based phishing websites classification | The first one is based on blacklists ,  in which the requested URL is compared with those in that  list.  The second approach is known as heuristic-based  method [2], where several features are collected from the website to classify it as either phishy or legitimate.  Rule-based classification algorithms for predicting phishing websites, including C4.5, RIPPER, PRISM, and CBA algorithms | aims to identify features that are effective in detecting phishing websites and evaluate how good rule-based classification techniques are at predicting phishing websites. | The results showed that we could improve the  prediction accuracy relying only on nine features, those are:  ‘Request URL, Age of Domain, HTTPS and SSL, Website  Traffic, Long URL, Sub Domain and Multi Sub Domain,  Adding prefix or Suffix Separated by (−) to Domain, URL  of Anchor and Using the IP Address’. After conducting the  experiments on the nine chosen features, the error-rate has  decreased for all the algorithms. Precisely, the CBA  algorithm has the lowest error-rate with 4.75%. |
| LizhenTang and QusayH.Mahmoud | A Survey of Machine Learning-Based Solutions for Phishing Website Detection |  |  |  |
| Eza Nanda , Istikomah ,Nurindah A.Amari,Yoga Pristyanto | PERBANDINGAN KLASIFIKASI ALGORITMA K-NN, NEURAL  NETWORK, NAÏVE BAYES, C 4.5 UNTUK MENDETEKSI WEB PHISING | meliputi Data Acquisition, Data Preprocessing, Classification Process, serta Evaluation | deteksi web phising yang berguna melindungi  dari data sensitif pengguna. | Berdasarakan pengujian model berbasiskan metode Naïve Bayes, Decision Tree  C4.5, K-NN, Neural Network menggunakan framework Weka v.3.8.2. Didapat hasil  penelitian dari 1353 dataset yang diolah dan menggunakan 10 atribut membuktikan  bahwa algoritma Decision Tree C4.5 memiliki kinerja yang lebih tinggi dibandingkan  dengan tiga algoritma lainnya yaitu Neural Network, Naïve Bayes dan k-NN. Hal ini  karena secara keseluruhan algoritma C4.5 mempunyai nilai kinerja baik akurasi, sensitifity,  presisi maupun f-measure yang paling tinggi. |
|  |  |  |  |  |