TEKNOKOM : Jurnal Teknologi dan Rekayasa Sistem Komputer
Vol. 6, No. 2, September 2023, pp. 123 - 126
ISSN: 2686-3219, DOI: 10.31943/teknokom.v6i2.153

IMPLEMENTASI METODE ADASYN DALAM DETEKSI URL BERBAHAYA MENGGUNAKAN MACHINE LEARNING: DEMI MENINGKATKAN KEAMANAN SIBER DI ERA DIGITAL

Gilang Dwi Setyawan[1], Andrie Yuswanto[2], Ahmad Maulid Ridwan[1], Budi Wibowo[2], Maman Firmansyah[1]
[1]Department of Computer Science, Universitas Nusa Mandiri, Indonesia
[2]Department of Informatics Engineering,Institut Teknologi Budi Utomo, Indonesia

A R T I C L E I N F O

History of the article:
Received July 13, 2023
Revised July 28, 2023
Accepted August 1, 2023
Published August 4, 2023

Keywords:
Adasyn method
URL detection
Cyber Security
Machine Learning

ABSTRACT

Cybercriminals exploit malicious URLs as a distribution channel to spread harmful software across the internet. They take advantage of vulnerabilities in browsers to install malicious software with the aim of gaining remote access to the victims' computers. Typically, this malicious software aims to gain access to networks, steal sensitive information, and silently monitor targeted computer systems. In this research, a data mining approach known as Classification Based on Association (CBA) is employed to detect malicious URLs using both the URL itself and the features of the presented web pages. The CBA algorithm utilizes a training dataset of URLs as historical data to discover association rules that can be used to create an accurate classifier. By detecting dangerous URLs and malicious software, this contribution can assist organizations and individual users in enhancing the security of their computer systems and networks, thereby protecting sensitive data and reducing the risk of security incidents. The experimental results demonstrate that CBA achieves performance on par with tested classification algorithms, achieving an accuracy of 99% and low rates of false positives and false negatives. Future research could expand its focus to detect malicious URLs and software on mobile devices and embedded systems, as they have become significant targets for cybercriminals.

Kata Kunci :

Metode Adasyn
Deteksi URL
Keamanan Siber
Machine Learning

ABSTRAK

Penjahat siber memanfaatkan URL berbahaya sebagai jalur distribusi untuk menyebarkan perangkat lunak berbahaya melalui internet. Mereka mengeksploitasi kerentanan dalam browser untuk menginstal perangkat lunak berbahaya dengan tujuan mendapatkan akses ke komputer korban dari jarak jauh. Umumnya, perangkat lunak berbahaya bertujuan untuk mendapatkan akses ke jaringan, mencuri informasi sensitif, dan secara diam-diam memantau sistem komputer yang ditargetkan. Dalam penelitian ini, digunakan pendekatan data mining yang dikenal sebagai klasifikasi berdasarkan asosiasi (CBA) untuk mendeteksi URL berbahaya dengan menggunakan URL dan fitur konten halaman web yang disajikan. Algoritme CBA menggunakan kumpulan data pelatihan URL sebagai data historis untuk menemukan aturan asosiasi yang dapat digunakan untuk membuat pengklasifikasi yang akurat. Hasil percobaan menunjukkan bahwa CBA memberikan kinerja yang setara dengan algoritme klasifikasi yang sudah teruji, dengan mencapai akurasi 99% dan tingkat positif dan negatif palsu yang rendah.

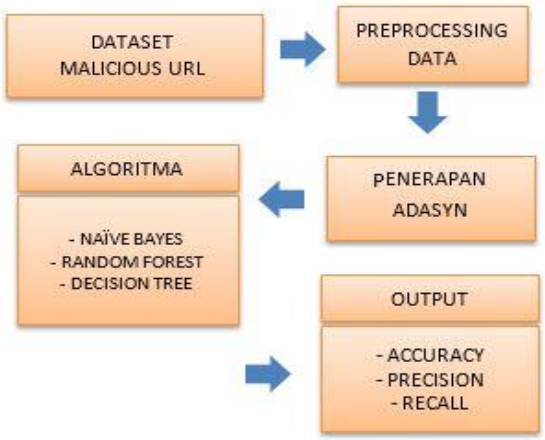
Correspondece:
Gilang Dwi Setyawan,
Department of Computer Science,
Universitas Nusa Mandiri, Indonesia
Email : ggilang57@gmail.com

PENDAHULUAN

Internet menjadi hal yang penting dan signifikan terhadap kehidupan kita sehari-hari. Banyak layanan yang dapat dilakukan internet yang bergantung pada fungsionalitas dan keamanannya, misalnya bisnis, pembelajaran, perbankan, jejaring sosial, kesehatan dan banyak lainnya yang merupakan aplikasi berbasis web. [1] Web menjadi semakin penting, penjahat dunia maya secara ilegal dapat megeksploitasi kerentanan dan memiliki peluang untuk melakukan banyak serangan terhadap aplikasi web. [2] Saat ini, keamanan Web menjadi faktor kunci dalam pengembangan web. Untuk meningkatkan keamanan situs Web, maka digunakan firewall pada suatu situs untuk keamanan, dan mencoba untuk mendeteksi adanya celah keamanan yang dapat dimanfaatkan oleh penyerang atau orang yang tidak bertanggung jawab. [3] Modus kejahatan di dunia cyber saat ini sangat beragam. Teknik yang digunakan oleh penyerangan pun semakin beragam dan kompleks. Berbagai serangan tersebut melibatkan malicious software atau yang bisa disebut malware yang merupakan suatu program jahat. Ancaman malware dan penyebarannya bisa melalui berbagai cara, yaitu cara yang sering menyisipkan di sebuah aplikasi ataupun file tertentu. [4] Malware dapat menyebar dengan cepat di jaringan tanpa campur tangan dari pengguna. Sistem pendeteksi malware masih menjadi masalah, karena malware baru yang selalu berevolusi dengan menggunakan teknik yang berbeda untuk menghindari metode pendeteksian, maka dari itu diperlukan pengembangan teknik deteksi malware yang dapat mendeteksi malware secara akurat. Sasaran utama dari malware adalah untuk memata-matai seseorang, mencuri informasi atau data pribadi orang lain seperti m-banking, membobol security program dan lain-lain. Pada umumnya, sebuah malware diciptakan untuk merusak atau membobol suatu software atau sistem operasi melalui script yang dirahasiakan, dalam arti lain disisipkan secara tersembunyi oleh penyerang. Perkembangan malware semakin pesat mengharuskan pengguna komputer semakin waspada agar informasi pribadi ataupun file yang penting tidak diambil oleh orang yang tidak berhak. [5] Demikian juga bagi para pelaku bisnis baik perusahaan maupun perorangan yang bergantung dengan sistem komputer untuk menjaga agar datanya tetap aman, dan malware tidak dapat mencuri atau merusak data yang dimiliki [6].

Internet menjadi populer sehingga mengubah cara berpikir dan standar hidup orang banyak. Salah satu faktor pendorong dari banyaknya malware berevolusi yaitu pesatnya pertumbuhan e-commerce, persoalan komputer yang tidak aman dan penetrasi internet yang terus meningkat. Sehingga membutuhkan suatu teknik baru untuk mendeteksi malware seperti machine learning. Teknologi machine learning dapat mendeteksi malware dengan mempelajari perilaku malware dan executable yang berbahaya. [7] Salah satu cara untuk deteksi malicious website yaitu dengan “exploit” application layer. Application layer adalah layer tertinggi dalam Open System Interconnection (OSI) layer. Layer ini berfokus pada process-to-process communication melewati IP network dan menyediakan layanan interface komunikasi dan user services. Application layer menyediakan banyak layanan diantaranya: file transfer, network data sharing, web surfing, web chat, dan email clients. [8] Untuk membedakan malicious website dan yang bukan maka digunakan metode klasifikasi. Klasifikasi merupakan proses untuk mengambil inputdan akan dimasukkan ke dalam kelas yang sudah ada. Metode ini dapat digunakan untuk memprediksi data baru.

METODE PENELITIAN



Gambar 1. Framework Penelitian ADASYN

Metode penelitian yang digunakan dalam studi ini melibatkan tahapan pre-processing data untuk mengolah dataset awal yang terdiri dari 2 fitur, yaitu "url" (domain utama) dan "type" (phishing, benign, defacement). Selanjutnya, dataset tersebut diperluas menjadi 11 fitur yang lebih informatif untuk meningkatkan performa dan akurasi analisis. Fitur yang ditambahkan meliputi "url_type", "url_len", "letters_count", "digits_count", "special_chars_count", "shortened", "abnormal_url", "secure_http", "have_ip", "url_region", dan "root_domain". Pertama, fitur "url_type" ditetapkan sebagai kelas atau label yang akan diprediksi dalam proses klasifikasi. Nilai pada fitur ini mencakup tiga kategori, yaitu "phishing", "benign", dan "defacement", yang merepresentasikan tipe dari setiap URL.

Selain menggunakan metode pre-processing untuk meningkatkan kualitas dataset, dalam penelitian ini juga diterapkan teknik ADASYN (Adaptive Synthetic Sampling) untuk menangani ketidakseimbangan kelas dalam data. Sebelumnya, dataset awal dibagi menjadi empat tipe kelas dengan jumlah sampel masing-masing: "phishing" (2) sebanyak 433,419 sampel, "benign" (3) sebanyak 428,030 sampel, "defacement" (1) sebanyak 427,420 sampel, dan "others" (0) sebanyak 427,380 sampel. ADASYN merupakan salah satu teknik oversampling yang adaptif, yang bertujuan untuk menyeimbangkan distribusi sampel di dalam setiap kelas dengan cara menciptakan sampel sintetis pada kelas minoritas (minority class). Dengan penerapan ADASYN, jumlah sampel pada kelas "phishing", "benign", dan "defacement" akan meningkat secara adaptif hingga mendekati jumlah sampel pada kelas mayoritas, yaitu "others". Setelah melakukan pengolahan data selanjutnya melakukan eksperimen dengan menggunakan algoritma random forest, decision tree dan naïve bayes dan mendapatkan hasil sebagai berikut.

Tabel 1. Hasil Akurasi

Algoritma	Accuracy	Precision	Recall	F1
Random Forest	99%	99%	99%	99%
Decision Tree	98%	98%	99%	98%
Naïve Bayes	96%	96%	96%	96%

Hasil evaluasi model menunjukkan adanya perbedaan performa yang signifikan di antara ketiga algoritma klasifikasi. Model Random Forest mencapai akurasi tertinggi, yaitu sebesar 99%, dan juga memiliki nilai presisi, recall, dan F1-Score yang sangat baik. Hal ini menandakan kemampuan model dalam mengklasifikasikan data dengan akurat dan dapat mengenali kelas minoritas dengan baik. Di sisi lain, model Decision Tree memiliki akurasi yang sedikit lebih rendah, yakni sebesar 98%, dan memiliki nilai presisi, recall, dan F1-Score yang cukup baik, namun sedikit lebih rendah dibandingkan dengan model Random Forest. Sementara itu, model Naive Bayes menunjukkan performa yang lebih rendah dengan akurasi sebesar 96% dan nilai presisi, recall, dan F1-Score yang lebih rendah dari dua model lainnya. Berdasarkan hasil ini, dapat disimpulkan bahwa Random Forest adalah metode algoritma klasifikasi yang paling sesuai untuk mengklasifikasikan tipe URL dalam dataset yang telah diolah. Model ini mampu memberikan prediksi yang akurat dan konsisten, serta memiliki kemampuan yang baik dalam menghadapi masalah ketidakseimbangan kelas. Namun, keputusan akhir dalam memilih algoritma harus mempertimbangkan tujuan analisis data dan karakteristik dari dataset yang lebih luas.

HASIL DAN PEMBAHASAN

Table 2. Table Hasil Eksperimen

Algoritma	Precision	Akurasi
Random Forest	99%	99%
Decision Tree	98%	98%
Naïve Bayes	96%	96%

Random Forest mengungguli metode lain dengan mencapai akurasi yang hampir sempurna, yaitu 99%. Sementara itu, Decision Tree juga memberikan kinerja yang baik dengan akurasi sebesar 98%. Di sisi lain, Naive Bayes menghasilkan akurasi sebesar 96%. Meskipun Naive Bayes adalah metode yang sederhana dan komputasionalnya cepat, hasilnya sedikit lebih rendah dibandingkan dengan metode lain dalam eksperimen ini. Naive Bayes mengasumsikan independensi antara fitur, yang mungkin tidak sepenuhnya memenuhi kondisi dalam dataset ini.

KESIMPULAN

Kesimpulan dari penelitian ini adalah bahwa metode Random Forest menunjukkan kinerja yang paling baik dalam mengklasifikasikan tipe URL dalam dataset yang telah diolah. Dengan akurasi hampir sempurna mencapai 99%, Random Forest mampu memberikan prediksi yang akurat dan konsisten, serta memiliki kemampuan yang baik dalam menghadapi masalah ketidakseimbangan kelas. Meskipun metode Decision Tree juga memberikan kinerja yang baik dengan akurasi sebesar 98%, hasilnya sedikit lebih rendah dibandingkan dengan Random Forest. Di sisi lain, metode Naive Bayes menunjukkan performa yang lebih rendah dengan akurasi sebesar 96% dan nilai presisi, recall, serta F1-Score yang lebih rendah dari dua metode lainnya. Hal ini menandakan bahwa Naive Bayes, meskipun sederhana dan komputasionalnya cepat, mungkin tidak sepenuhnya cocok untuk karakteristik dataset yang kompleks seperti yang digunakan dalam penelitian ini. Dalam rangka meningkatkan kehandalan dan generalisasi dari model yang dihasilkan, penelitian selanjutnya dapat memperluas eksperimen ini dengan melibatkan lebih banyak metode klasifikasi dan teknik pre-processing data lainnya. Dengan demikian, dapat diharapkan pengembangan metode klasifikasi yang lebih efisien dan akurat untuk menghadapi tantangan dalam analisis data yang semakin kompleks dan beragam.

REFERENSI

[1] O. Adiputra and E. Setiawan, “Jurnal Sains dan Informatika,” J. Sains dan Inform., vol. 4, no. 1, pp. 8–14, 2018, doi: 10.22216/jsi.v4i1.

[2] D. Stevanovic, N. Vlajic, and A. An, “Unsupervised clustering of web sessions to detect malicious and non-malicious website users,” Procedia Comput. Sci., vol. 5, pp. 123–131, 2011, doi: 10.1016/j.procs.2011.07.018.

[3] M. Jain and P. Bajaj, “Techniques in Detection and Analyzing Malware Executables: A Review,” Int. J. Comput. Sci. Mob. Comput., vol. 35, no. 5, pp. 930–935, 2014.

[4] R. Adenansi and L. A. Novarina, “Malware dynamic,” J. Educ. Inf. Commun. Technol., vol. 1, no. 1, pp. 37–43, 2017.

[5] D. Ashit, “Detection of Malware and Malicious Executables Using E-Birch Algorithm,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 1, pp. 124–126, 2016, doi: 10.14569/ijacsa.2016.070118.

[6] “Technopedia”, [Online]. Available: https://www.techopedia.com/definition/6006/application-layer

[7] A. Altaher, “Phishing Websites Classification using Hybrid SVM and KNN Approach,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, pp. 90–95, 2017, doi: 10.14569/ijacsa.2017.080611.

[8] N. Nursapdahi, A. S. Fitrani, and ..., “Studi Analisa Serangan Sql Injection,” Pros. SEMNAS …, pp. 185–190, 2022, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/2474

[9] I. A. Nikmatun and I. Waspada, “Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor,” J. SIMETRIS, vol. 10, no. 2, pp. 421–432, 2019.

[10] M. Vishwakarma and N. Kesswani, “A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection,” Decis. Anal. J., vol. 7, no. April, p. 100233, 2023, doi: 10.1016/j.dajour.2023.100233.

[11] W. Elghazel et al., “Random forests for industrial device functioning diagnostics using wireless sensor networks,” IEEE Aerosp. Conf. Proc., vol. 2015-June, 2015, doi: 10.1109/AERO.2015.7119275.

[12] X. Ma et al., “Predicting the utilization factor of blasthole in rock roadways by random forest,” Undergr. Sp., vol. 11, pp. 232–245, 2023, doi: 10.1016/j.undsp.2023.01.006.

[13] P. Yhoga et al., “Media Sosial Menggunakan Metode Decision Tree Untuk Badan Pusat,” 2017.

[14] H. Amalia, R. Rahmadanti, A. Syaiin, and S. Salsabila, “Prediksi Resiko Kesehatan Ibu Hamil Dengan Menggunakan Metode Decision Tree,” vol. 11, no. 1, pp. 48–53, 2023.

[15] B. A. Qowy, F. Hanafi, M. A. Riandi, and A. Nuraminah, “Jurnal Teknik Informatika dan Elektro Penerapan Pemilihan Model Dinamis Algoritma Behaviour Tree Decision dalam Third Person Game pada Musuh Non-Playable Charater,” vol. 3, no. 1, pp. 32–37, 2021.