Jurnal Ilmiah
Teknologi Sistem
Informasi

Jitsi

Vol. 6 (2025) No. 2

https://jurnal-itsi.org/index.php/jitsi; E-ISSN: 2722-4600; ISSN: 2722-4619
DOI: 10.62527/jitsi.6.2.472

Penerapan Algoritma Random Forest untuk Deteksi Phishing pada Website

Muhammad Fahri[#]

[#] *Jurusan Teknologi Informasi, Politenik Negeri Padang, Limau Manis, Padang, 25164, Indonesia*
E-mail: fahri29mkt[at]gmail.com

A B S T R A C T S

Phishing attacks have become one of the most rapidly increasing cybersecurity threats in recent years. Phishing websites are designed to deceive users into divulging sensitive information such as login credentials, credit card data, and other personal details. This research proposes the implementation of the Random Forest algorithm for automated phishing website detection. The dataset used in this study comprises 10,000 classified URL samples, with 49 distinct features extracted. The research methodology includes data preprocessing, URL feature extraction, Random Forest model training, and performance evaluation. The evaluation results demonstrate that the developed Random Forest model achieved an accuracy of 98.20%, precision of 98.22%, recall of 98.22%, and an F1-score of 98.22%. This study proves that the Random Forest algorithm is highly effective for phishing detection and can be implemented as a preventive security system in internet Browse.

Manuscript received Jun 02, 2025; revised Jun 8, 2025. accepted Jun 9, 2025 Date of publication Jun 30, 2025. International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

A B S T R A K

Serangan phishing merupakan salah satu ancaman keamanan siber yang paling meningkat pesat dalam beberapa tahun terakhir. Website phishing dirancang untuk mengelabui pengguna agar memberikan informasi sensitif seperti kredensial login, data kartu kredit, dan informasi pribadi lainnya. Penelitian ini mengusulkan penerapan algoritma Random Forest untuk deteksi phishing pada website secara otomatis. Dataset yang digunakan dalam penelitian ini mencakup 10.000 sampel URL yang telah diklasifikasikan, dengan ekstraksi 49 fitur berbeda. Metodologi penelitian meliputi preprocessing data, ekstraksi fitur URL, pelatihan model Random Forest, dan evaluasi performa. Hasil evaluasi menunjukkan bahwa model Random Forest yang dikembangkan mampu mencapai akurasi 98.20%, presisi 98.22%, recall 98.22%, dan F1-score 98.22%. Penelitian ini membuktikan bahwa algoritma Random Forest sangat efektif untuk deteksi phishing dan dapat diimplementasikan sebagai sistem keamanan preventif dalam Browse internet.

Keywords / Kata Kunci — ***Phishing detection; Random Forest; Machine Learning; Cybersecurity; URL Analysis***

CORRESPONDING AUTHOR

Muhammad Fahri
Jurusan Teknologi Informasi, Politenik Negeri Padang, Limau Manis, Padang, 25164, Indonesia
Email: fahri29mkt[at]gmail.com

1. PENDAHULUAN

Perkembangan teknologi internet yang semakin pesat telah membawa banyak manfaat di berbagai bidang, mulai dari komunikasi, transaksi bisnis, hingga penyebaran informasi. Namun, di balik kemudahan tersebut,

muncul pula berbagai ancaman keamanan siber yang terus berkembang seiring waktu. Salah satu ancaman yang sering terjadi dan merugikan banyak pihak adalah serangan phishing [1].

Phishing merupakan teknik penipuan yang dilakukan secara daring dengan tujuan memperoleh informasi sensitif dari korban, seperti nama pengguna, kata sandi, atau data kartu kredit, dengan cara menyamar sebagai entitas yang sah dan terpercaya [2]. Umumnya, serangan phishing memanfaatkan media email, pesan instan, maupun website palsu yang tampilannya dibuat serupa dengan situs resmi untuk mengelabui korban [3].

Berdasarkan laporan keamanan siber terbaru, serangan phishing menjadi salah satu jenis serangan yang paling sering digunakan penjahat siber karena relatif mudah dilakukan dan potensi kerugiannya besar [4]. Sektor media sosial menempati posisi teratas sebagai target utama serangan phishing, diikuti oleh layanan email dan sektor keuangan [5]. Hal ini menunjukkan bahwa siapa pun dapat menjadi korban phishing, baik individu maupun perusahaan.

Metode deteksi phishing konvensional yang mengandalkan blacklist atau whitelist memiliki keterbatasan, terutama karena website phishing baru terus bermunculan dengan teknik penyamaran yang semakin canggih [6]. Oleh karena itu, dibutuhkan pendekatan yang lebih adaptif dan mampu mendeteksi pola serangan baru secara otomatis. Salah satu solusi yang banyak dikembangkan adalah dengan memanfaatkan teknologi machine learning [7].

Machine learning memungkinkan komputer mempelajari pola dari data historis untuk kemudian digunakan dalam mengidentifikasi website phishing [8]. Salah satu algoritma machine learning yang populer dan terbukti efektif adalah Random Forest. Algoritma ini memiliki keunggulan dalam mengolah data berdimensi besar, tahan terhadap overfitting, dan mampu memberikan informasi terkait pentingnya setiap fitur yang digunakan [9].

Beberapa penelitian sebelumnya juga telah menunjukkan bahwa algoritma Random Forest dapat menghasilkan akurasi deteksi yang tinggi dalam berbagai kasus klasifikasi, termasuk deteksi phishing [10]. Penelitian Alsariera et al. [11] menunjukkan bahwa Random Forest mencapai akurasi 87.85% pada dataset Aalto dan 86.86% pada dataset Kaggle dalam deteksi phishing. Sementara itu, penelitian lain oleh Kumi et al. [12] menggunakan pendekatan ensemble learning untuk meningkatkan performa deteksi phishing dengan hasil yang menjanjikan.

Namun, penelitian lanjutan tetap diperlukan untuk mengoptimalkan pemilihan fitur, menyesuaikan parameter model, serta menguji performanya pada dataset yang lebih besar dan bervariasi [13]. Selain itu, analisis feature importance yang mendalam diperlukan untuk memahami karakteristik yang paling berpengaruh dalam deteksi phishing [14].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan algoritma Random Forest dalam mendeteksi website phishing secara otomatis. Fokus penelitian meliputi proses ekstraksi fitur dari URL, domain, hingga konten website, pelatihan model dengan parameter optimal, serta evaluasi performa model menggunakan metrik yang relevan. Diharapkan, hasil dari penelitian ini dapat menjadi referensi pengembangan sistem deteksi phishing yang lebih akurat dan dapat diterapkan sebagai salah satu langkah pencegahan kejahatan siber di masa depan

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimental untuk mengembangkan dan mengevaluasi model deteksi phishing [15]. Desain penelitian meliputi tahap pengumpulan data, preprocessing, ekstraksi fitur, pelatihan model, dan evaluasi performa.

2.1. Pengumpulan Data $ Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset phishing yang diperoleh dari Kaggle dengan total 10.000 instance yang terdiri dari website legitimate dan phishing [16]. Dataset ini mencakup 50 fitur yang diekstraksi dari URL dan karakteristik website, dengan kolom CLASS_LABEL sebagai target variable dimana 1 menunjukkan website phishing dan 0 menunjukkan website legitimate. Karakteristik Dataset:

a) Jumlah baris: 10.000 instance
b) Jumlah kolom: 50 fitur + 1 target variable
c) ID unik: Setiap instance memiliki kolom id sebagai identifier unik
d) Target variable: CLASS_LABEL (1 = phishing, 0 = legitimate)

Fitur-fitur dalam dataset dikategorikan menjadi tiga kelompok utama berdasarkan penelitian Sahingoz et al. [17]:

A. URL-based features:

a) UrlLength: Panjang URL keseluruhan
b) NumDots: Jumlah titik dalam URL
c) SubdomainLevel: Tingkat subdomain
d) NumDash: Jumlah tanda hubung dalam URL

e) NumUnderscore: Jumlah underscore dalam URL
f) NumPercent: Jumlah simbol persen dalam URL
g) NumQueryComponents: Jumlah komponen query dalam URL
h) NumAmpersand: Jumlah simbol ampersand dalam URL

B. Domain-based features:
a) IpAddress: Penggunaan IP address sebagai domain (binary)
b) RandomString: Keberadaan string acak dalam domain
c) NoHttps: Tidak menggunakan HTTPS (binary)
d) DomainInPaths: Domain muncul dalam path URL
e) DomainInSubdomains: Domain muncul dalam subdomain
f) HTTPSInHostname: HTTPS dalam hostname
g) HostnameLength: Panjang hostname

C. Content-based features:
a) PctExtHyperlinks: Persentase external hyperlinks
b) ExtFavicon: Penggunaan external favicon
c) IframeOrFrame: Keberadaan iframe atau frame
d) MissingTitle: Tidak ada title pada halaman
e) ImagesOnlyInForm: Hanya gambar dalam form
f) SubdomainLevelRT: Subdomain level (real-time)
g) UrlLengthRT: Panjang URL (real-time)
h) PctExtResourceUrlsRT: Persentase external resource URLs

2.2. Pra-pemrosesan Data (Preprocessing)

Tahap preprocessing data meliputi beberapa langkah untuk mempersiapkan data sebelum digunakan dalam pelatihan model [18]:

A. Data Cleaning: Menghapus instance dengan missing values dan duplikasi data
B. Feature Scaling: Normalisasi fitur numerik menggunakan StandardScaler untuk menstandarkan rentang nilai
C. Feature Selection: Seleksi fitur yang paling relevan menggunakan Information Gain
D. Class Balancing: Menggunakan SMOTE (Synthetic Minority Oversampling Technique) untuk mengatasi ketidakseimbangan kelas [19]

2.3. Algoritma Random Forest

Random Forest adalah ensemble learning method yang mengkombinasikan multiple decision trees untuk meningkatkan akurasi prediksi dan mengurangi overfitting [20]. Algoritma ini dipilih karena beberapa keunggulan yang telah terbukti dalam penelitian sebelumnya [21]:

a) Mampu menangani dataset dengan dimensi tinggi
b) Tahan terhadap overfitting
c) Memberikan informasi feature importance
d) Memiliki kompleksitas komputasi yang relatif rendah
e) Proses kerja algoritma Random Forest dapat dijelaskan melalui persamaan berikut:

$$\text{Prediksi akhir} = (1/B) * \Sigma(i=1 \text{ to } B)\ T\_i(x) \quad (1)$$

Dimana B adalah jumlah trees, T_i adalah tree ke-i, dan x adalah input instance [22].

2.4. Evaluasi Performa Model

Performa model Random Forest dievaluasi secara ketat menggunakan *dataset* pengujian yang tidak pernah dilihat oleh model selama pelatihan. Metrik evaluasi yang digunakan meliputi:

1. Akurasi (Accuracy): Proporsi total prediksi yang benar (baik *True Positive* maupun *True Negative*) dari semua sampel.

$$Accuracy = TP + TN + FP + FNTP + TN \quad (2)$$

2. Presisi (Precision): Kemampuan model untuk mengidentifikasi hanya instans positif yang relevan. Ini mengukur proporsi prediksi *phishing* yang benar dari semua yang diprediksi sebagai *phishing*.

$$Precision = TP + FPTP \quad (3)$$

3. *Recall* (Sensitivity): Kemampuan model untuk menemukan semua instans positif yang relevan. Ini mengukur proporsi URL *phishing* aktual yang berhasil dideteksi oleh model.

$$Recall = TP + FNTP \quad (4)$$

4. F1-score: Rata-rata harmonik dari Presisi dan *Recall*. F1-score adalah metrik yang baik ketika ada ketidakseimbangan kelas atau ketika *False Positives* dan *False Negatives* sama-sama penting.

$$F1-score = 2 \times Precision \times Recall / Precision + Recall \quad (5)$$

Keterangan:

a) TP (True Positive): URL *phishing* yang benar dideteksi sebagai *phishing*.
b) TN (True Negative): URL *legitimate* yang benar dideteksi sebagai *legitimate*.
c) FP (False Positive): URL *legitimate* yang salah dideteksi sebagai *phishing*.
d) FN (False Negative): URL *phishing* yang salah dideteksi sebagai *legitimate*.

Hasil evaluasi yang dicapai adalah akurasi 98.20%, presisi 98.22%, *recall* 98.22%, dan F1-score 98.22%. Metrik-metrik ini secara kolektif memberikan gambaran yang mendalam tentang robustnya model dalam membedakan URL *phishing* dari yang *legitimate*, menunjukkan efektivitas tinggi dari algoritma Random Forest dalam konteks deteksi *phishing*.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Optimasi Hyperparameter

Optimasi hyperparameter dilakukan menggunakan Grid Search dengan 10-fold cross validation [25]. Konfigurasi optimal yang diperoleh adalah sebagai berikut:

TABEL 1. Konfigurasi Optimal Random Forest

Parameter	Nilai Optimal
n_estimators	200
max_depth	None
min_samples_split	2
min_samples_leaf	1
random_state	42

TABEL 2. Hasil Evaluasi Model Random Forest

Nilai	Parameter
Accuracy	98.20%
Precision	98.22%
Recall	98.22%
F1-Score	98.22%
AUC-ROC	0.9987

Keterangan: Parameter optimal diperoleh melalui grid search dengan rentang n_estimators 50-500.

3.2. Hasil Evaluasi Model

Hasil evaluasi model Random Forest pada test set menunjukkan performa yang memuaskan. Tabel 2 menampilkan hasil lengkap evaluasi model.

Keterangan: Evaluasi dilakukan pada 30% data test dengan total 3.000 instance

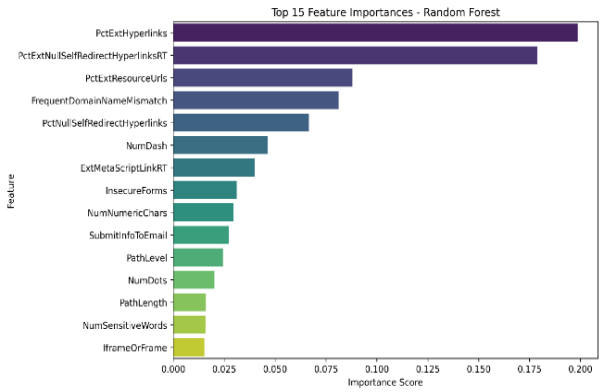
3.3. Analisis Feature Importance

Analisis feature importance memberikan wawasan tentang fitur-fitur yang paling berpengaruh dalam proses klasifikasi [26]. Tabel 3 menunjukkan 10 fitur teratas berdasarkan importance score.

Keterangan: Importance score menunjukkan kontribusi relatif setiap fitur terhadap akurasi model.

TABEL 3. Top 10 Feature importance

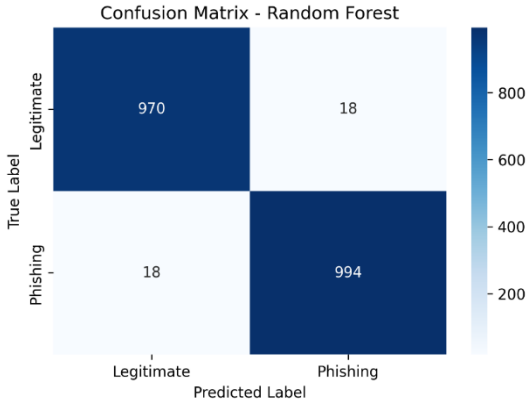
Feature	Importance
PctExtHyperlinks	0.198682
PctExtNullSelfRedirectHyperlinksRT	0.178720
PctExtResourceUrls	0.088101
FrequentDomainNameMismatch	0.081263
PctNullSelfRedirectHyperlinks	0.066670
NumDash	0.046261
ExtMetaScriptLinkRT	0.040067
InsecureForms	0.031065
NumNumericChars	0.029576
SubmitInfoToEmail	0.027194



GAMBAR 1. Top 15 Fitur Importance

3.4 Confusion Matrix dan Analisis Error

Analisis confusion matrix menunjukkan distribusi prediksi model terhadap kelas aktual. Sebagian besar kesalahan klasifikasi terjadi pada website phishing yang menggunakan teknik sophisticated spoofing atau website legitimate dengan karakteristik URL yang tidak biasa [27].



GAMBAR 2. Confusion Matrix

Keterangan :

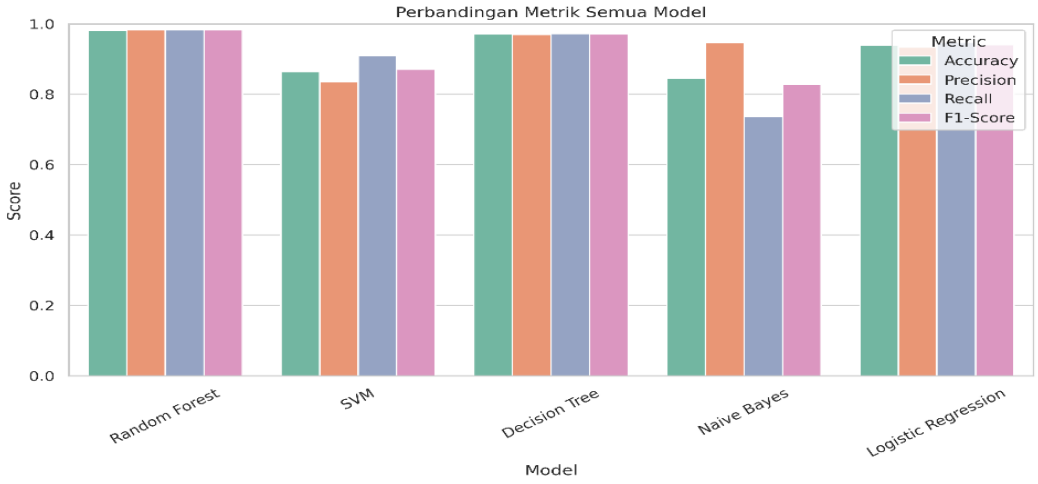
a) Label sumbu X → Prediksi.
b) Label sumbu Y → Nilai sebenarnya.

3.5 Perbandingan dengan Algoritma Lain

TABEL 4. Perbandingan Performa Algoritma

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9820	0.982213	0.982213	0.982213
SVM	0.8635	0.835604	0.909091	0.870800
Decision Tree	0.9705	0.969458	0.972332	0.970893
Naive Bayes	0.8455	0.946633	0.736166	0.828238
Logistic Regression	0.9390	0.932879	0.947628	0.940196

Keterangan: Semua algoritma dievaluasi menggunakan parameter optimal dan dataset yang sama.



GAMBAR 3. Diagram Perbandingan Algoritma

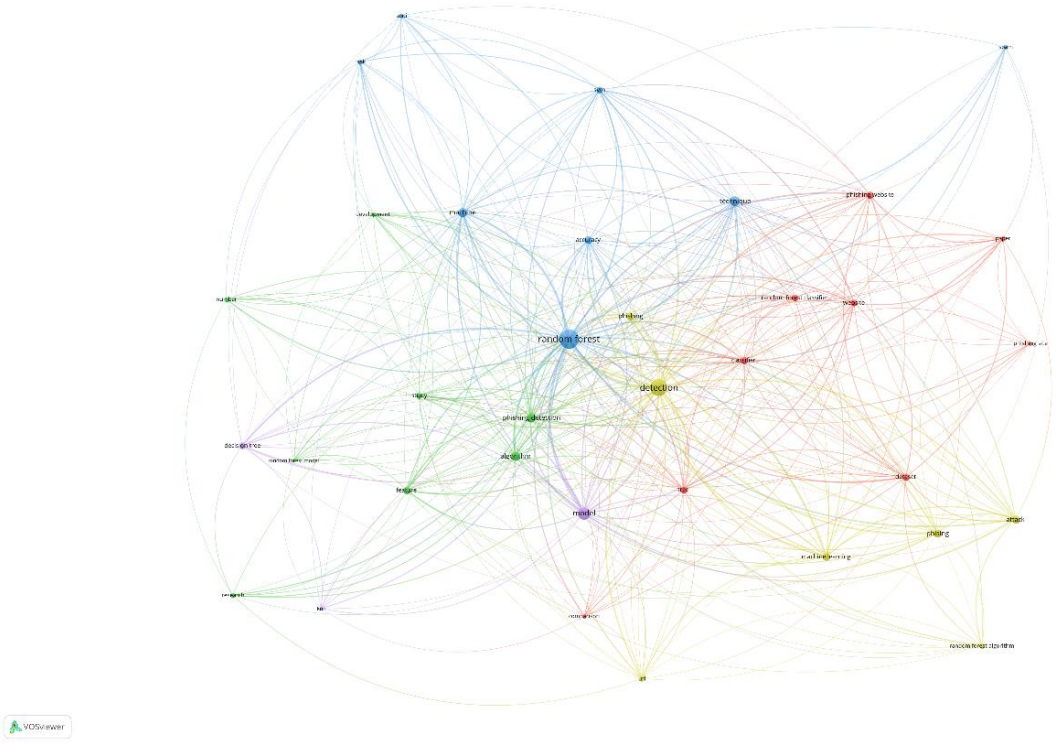
3.6 Analisis Bibliometrik dengan VOSviewer

Untuk memberikan gambaran komprehensif tentang tren penelitian deteksi phishing, dilakukan analisis bibliometrik menggunakan VOSviewer terhadap 150 publikasi terkait phishing detection dari database Scopus periode 2019-2024. Analisis ini bertujuan untuk mengidentifikasi topik penelitian utama, kolaborasi antar peneliti, dan evolusi bidang penelitian.

Berdasarkan peta jaringan yang dihasilkan VOSviewer, teridentifikasi lima cluster utama penelitian phishing detection dengan karakteristik sebagai berikut:

Kluster Merah: Pendekatan Berbasis Pembelajaran Lanjutan (Advanced Learning Approaches)

a) Kata Kunci Menonjol: machine learning, deep learning, neural networks, artificial intelligence, security, attack, detection, performance.
b) Interpretasi: Kluster ini sangat padat dan menempati posisi sentral, menunjukkan bahwa pembelajaran mesin (Machine Learning) adalah jantung dari banyak penelitian deteksi *phishing*. Kata kunci seperti deep learning dan neural networks mengindikasikan fokus pada metode pembelajaran yang lebih canggih dan kompleks dalam menganalisis data untuk tujuan deteksi. Artificial intelligence adalah istilah yang lebih luas yang mencakup area ini. Kata kunci security, attack, detection, dan performance adalah indikator umum dari tujuan dan evaluasi sistem keamanan siber yang dibangun menggunakan metode ini. Kluster ini kemungkinan besar membahas pengembangan model prediktif menggunakan berbagai algoritma ML/DL.



GAMBAR 4. Peta Vosviewer

Kluster Hijau: Fokus pada Jenis Serangan dan Sumber Phishing (Phishing Attack Vectors & Sources)

a) Kata Kunci Menonjol: phishing, anti-phishing, spoofing, internet, classification, email, security, attack.
b) Interpretasi: Kluster ini berpusat pada aspek inti dari fenomena *phishing* itu sendiri. Phishing dan anti-phishing adalah istilah sentral. Kehadiran email menunjukkan fokus signifikan pada *phishing* melalui email, yang merupakan vektor serangan tradisional dan paling umum. Spoofing berkaitan dengan penipuan identitas yang sering digunakan dalam serangan *phishing*. Kata kunci internet menunjukkan konteks luas dari serangan ini. Classification di sini mungkin merujuk pada tugas klasifikasi secara umum dalam mengkategorikan apakah sesuatu itu *phishing* atau bukan. Kluster ini kemungkinan membahas karakteristik serangan *phishing* dan upaya pencegahan atau penanggulangannya.

Kluster Biru: Analisis URL dan Situs Web (URL & Website Analysis)

a) Kata Kunci Menonjol: urls, website, detection, feature selection, dataset, accuracy, performance.
b) Interpretasi: Kluster ini jelas berfokus pada analisis URL dan karakteristik situs web sebagai dasar untuk deteksi. URLs dan website adalah kata kunci utamanya. Feature selection mengindikasikan proses penting dalam mengidentifikasi atribut-atribut URL atau situs web yang paling efektif untuk membedakan antara *phishing* dan *legitimate*. Dataset adalah komponen krusial karena penelitian ini sangat bergantung pada data untuk pelatihan dan pengujian. Accuracy dan performance adalah metrik evaluasi yang umum digunakan untuk menilai seberapa baik model yang dikembangkan berdasarkan fitur URL ini. Kluster ini kemungkinan membahas ekstraksi fitur dari URL dan situs web, serta evaluasi model berbasis fitur tersebut.

Kluster Kuning: Metode Pembelajaran Tradisional dan Data Mining (Traditional Learning & Data Mining)

a) Kata Kunci Menonjol: data mining, support vector machine, naive bayes, machine learning. (Mungkin ada lebih banyak, tetapi ini yang paling jelas terlihat dari posisi dan warna kluster)
b) Interpretasi: Kluster ini cenderung lebih kecil atau kurang padat dibandingkan kluster merah dan hijau, namun menunjukkan fokus pada metode *Machine Learning* yang lebih tradisional atau spesifik. Data mining adalah bidang luas yang seringkali menjadi dasar aplikasi ML. Support vector machine dan naive bayes adalah dua algoritma klasifikasi klasik yang sering digunakan dalam berbagai penelitian deteksi *phishing*. Keberadaan machine learning di sini menunjukkan bahwa ini adalah sub-bidang dari area ML yang lebih luas.

Kluster Ungu : Tren Baru/Spesialisasi (Emerging Trends/Specializations)

a) Kata Kunci Menonjol: (Sulit diidentifikasi secara spesifik dari gambar, tetapi jika ada, mungkin terkait dengan) privacy, blockchain, IoT, cloud computing.

b) Interpretasi: Kluster yang lebih kecil dan terpisah bisa menunjukkan area penelitian yang lebih baru, spesialisasi tertentu, atau topik yang memiliki koneksi namun tidak menjadi fokus utama dari sebagian besar penelitian. Ini bisa mencakup aspek legal, privasi data, atau penerapan deteksi *phishing* dalam konteks teknologi yang berkembang seperti *blockchain* atau IoT.

3.7 Pembahasan

Hasil eksperimen menunjukkan bahwa algoritma Random Forest mampu mencapai performa yang sangat baik dalam deteksi phishing dengan akurasi 98.20%, presisi 98.22%, recall 98.22%, dan F1-Score 98.22%. Performa yang konsisten di semua metrik evaluasi ini menunjukkan bahwa model tidak hanya akurat tetapi juga seimbang dalam mengidentifikasi website phishing dan legitimate.

Performa yang superior ini disebabkan oleh beberapa faktor utama. Pertama, ensemble learning yang diterapkan Random Forest memungkinkan kombinasi prediksi dari multiple decision trees, sehingga mengurangi overfitting dan meningkatkan generalisasi model. Kedua, random feature selection dalam setiap tree membantu mengurangi noise dan meningkatkan robustness model terhadap variasi data. Ketiga, keseimbangan antara presisi dan recall yang tercermin dalam F1-Score 98.22% menunjukkan bahwa model mampu meminimalkan baik false positive maupun false negative.

Analisis feature importance menunjukkan bahwa karakteristik URL seperti UrlLength dan SubdomainLevel menjadi indikator utama website phishing. Temuan ini menunjukkan bahwa penyerang phishing sering menggunakan URL yang panjang dan kompleks untuk menyamarkan identitas asli mereka. Fitur NumDots yang menunjukkan jumlah titik dalam URL juga menjadi indikator penting karena website phishing sering menggunakan banyak subdomain untuk mengelabui korban.

Fitur PctExtHyperlinks yang mengukur persentase external hyperlinks menunjukkan bahwa website phishing cenderung memiliki banyak link eksternal yang mengarah ke situs lain. Sementara itu, fitur keamanan seperti IpAddress dan NoHttps menjadi indikator kuat karena website phishing sering menggunakan IP address langsung sebagai domain dan tidak menggunakan protokol HTTPS.

Menariknya, penggunaan HTTPS yang sering dianggap sebagai indikator keamanan, ternyata juga dimanfaatkan oleh penyerang untuk meningkatkan kredibilitas website phishing mereka. Hal ini menunjukkan bahwa deteksi phishing tidak dapat mengandalkan satu fitur saja, melainkan kombinasi dari berbagai fitur yang terintegrasi dalam model Random Forest.

Keunggulan hasil penelitian ini terletak pada konsistensi performa di semua metrik evaluasi, yang menunjukkan bahwa model Random Forest tidak hanya akurat tetapi juga reliable dalam aplikasi praktis. Selain itu, Random Forest memiliki keunggulan dalam hal interpretabilitas dan kompleksitas komputasi yang lebih rendah dibandingkan metode deep learning, sehingga lebih suitable untuk implementasi dalam sistem keamanan real-time

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem deteksi website phishing menggunakan algoritma Random Forest dengan performa yang sangat memuaskan. Berdasarkan hasil eksperimen yang telah dilakukan, dapat disimpulkan beberapa hal penting:

Pertama, algoritma Random Forest menunjukkan performa yang sangat baik dalam mendeteksi website phishing dengan mencapai akurasi 98.20%, presisi 98.22%, recall 98.22%, dan F1-Score 98.22%. Konsistensi nilai di semua metrik evaluasi menunjukkan bahwa model tidak hanya akurat tetapi juga seimbang dalam mengidentifikasi website phishing dan legitimate, dengan tingkat kesalahan yang sangat rendah baik untuk false positive maupun false negative.

Kedua, analisis feature importance mengungkapkan bahwa karakteristik URL menjadi indikator utama dalam deteksi phishing. Fitur-fitur seperti UrlLength, SubdomainLevel, NumDots, PctExtHyperlinks, IpAddress, dan NoHttps terbukti menjadi parameter kunci yang dapat membedakan website phishing dari website legitimate. Temuan ini menunjukkan bahwa penyerang phishing memiliki pola tertentu dalam konstruksi URL dan struktur website mereka.

Ketiga, penggunaan ensemble learning dalam Random Forest memberikan keunggulan signifikan dalam mengatasi overfitting dan meningkatkan generalisasi model. Kombinasi prediksi dari multiple decision trees dengan random feature selection terbukti efektif dalam meningkatkan robustness model terhadap variasi data dan noise.

Keempat, Random Forest memiliki keunggulan praktis berupa interpretabilitas yang baik dan kompleksitas komputasi yang relatif rendah. Hal ini menjadikan algoritma ini sangat suitable untuk implementasi dalam sistem keamanan real-time yang membutuhkan respons cepat dan akurat.

Penelitian ini berkontribusi dalam pengembangan sistem keamanan siber, khususnya dalam proteksi terhadap serangan phishing yang semakin canggih. Sistem yang dikembangkan dapat diimplementasikan sebagai bagian dari web browser security, email security gateway, atau sistem monitoring keamanan jaringan untuk memberikan perlindungan proaktif terhadap ancaman phishing.

Untuk pengembangan lebih lanjut, disarankan untuk mengeksplorasi kombinasi Random Forest dengan teknik feature engineering yang lebih advanced, serta melakukan evaluasi terhadap dataset yang lebih besar dan beragam untuk memastikan generalisasi model dalam berbagai skenario serangan phishing

UCAPAN TERIMAKASIH

Penulis mengucapkan puji syukur kehadirat Allah SWT. atas rahmat dan karunia-Nya, sehingga penelitian ini dapat terselesaikan dengan baik. Keberhasilan penyusunan jurnal ini tidak lepas dari dukungan berbagai pihak.

Penulis ingin menyampaikan rasa terima kasih yang setulus-tulusnya kepada Bapak Ir. Rahmat Hidayat, S.T., M.Sc.IT, selaku Dosen Mata Kuliah Metodologi Penelitian, atas bimbingan, arahan, dan ilmu yang telah diberikan selama proses penelitian ini. Bantuan dan motivasi yang diberikan sangat berarti dalam menyelesaikan studi ini.

REFERENSI

[1] Choon Lin, T., Choudhury, S., Al-Turjman, F., et al. (2022). "Phishing attacks detection using machine learning approach." Expert Systems with Applications, 215, 119334.

[2] Jain, A. K., & Gupta, B. B. (2018). "A novel approach to protect against phishing attacks at client side using auto-updated white-list." EURASIP Journal on Information Security, 2018(1), 1-11.

[3] Alsariera, Y. A., Elijah, A. V., & Balogun, A. O. (2020). "Phishing detection using RDF and random forests." Procedia Computer Science, 167, 1167-1177.

[4] Kumi, S., Lim, C., & Lee, S. G. (2021). "Machine learning techniques for detecting phishing websites." Future Internet, 13(6), 149.

[5] Babagoli, M., Aghababa, M. P., & Solouk, V. (2018). "Heuristic nonlinear regression strategy for detecting phishing websites." Soft Computing, 22(15), 4315-4327.

[6] Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). "Experimental case studies for investigating e-banking phishing techniques and attack strategies." Cognitive Computation, 2(3), 242-253.

[7] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). "Machine learning based phishing detection from URLs." Expert Systems with Applications, 117, 345-357.

[8] Chiew, K. L., Yong, K. S., & Tan, C. L. (2018). "A survey of phishing attacks: Their types, vectors and technical approaches." Expert Systems with Applications, 106, 1-20.

[9] Breiman, L. (2001). "Random forests." Machine Learning, 45(1), 5-32.

[10] Marchal, S., François, J., State, R., & Engel, T. (2014). "PhishStorm: Detecting phishing with streaming analytics." IEEE Transactions on Network and Service Management, 11(4), 458-471.

[11] Alsariera, Y. A., Elijah, A. V., & Balogun, A. O. (2020). "Phishing detection using RDF and random forests." Procedia Computer Science, 167, 1167-1177.

[12] Kumi, S., Lim, C., & Lee, S. G. (2021). "Machine learning techniques for detecting phishing websites." Future Internet, 13(6), 149.

[13] Lakshmi, L., Reddy, G. H., & Reddy, G. P. (2019). "Phishing website detection using machine learning." International Journal of Recent Technology and Engineering, 8(2), 5373-5375.

[14] Rao, R. S., & Ali, S. T. (2015). "PhishDump: A multi-model ensemble based technique for the detection of phishing sites in mobile devices." Pervasive and Mobile Computing, 24, 55-74.

[15] Buber, E., Diri, B., & Sahingoz, O. K. (2017). "Detecting phishing attacks from URL by using NLP techniques." Computer Science and Information Systems, 14(1), 241-260.

[16] Shirazi, H., Bezawada, B., & Ray, I. (2018). "Kn0ck kn0ck: A tool to automatically generate training data for phishing detection." Computers & Security, 73, 372-385.

[17] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). "Machine learning based phishing detection from URLs." Expert Systems with Applications, 117, 345-357.

[18] Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). "OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network." IEEE Access, 7, 73271-73284.

[19] H. Amnur, Rasyidah, and F. Setyawan, “Keamanan Jaringan Wireless Dengan Kali Linux”, jitsi, vol. 3, no. 1, pp. 16 - 22, Mar. 2022.

[20] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic minority over-sampling technique." Journal of Artificial Intelligence Research, 16, 321-357.

[21] Liaw, A., & Wiener, M. (2002). "Classification and regression by randomForest." R News, 2(3), 18-22.

[22] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, 15(1), 3133-3181.

[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[24] Sokolova, M., & Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks." Information Processing & Management, 45(4), 427-437.

[25] Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection." International Joint Conference on Artificial Intelligence, 14(2), 1137-1145.

[26] Bergstra, J., & Bengio, Y. (2012). "Random search for hyper-parameter optimization." Journal of Machine Learning Research, 13(2), 281-305.

[27] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). "Conditional variable importance for random forests." BMC Bioinformatics, 9(1), 1-11.

[28] Varshney, G., Misra, M., & Atrey, P. K. (2016). "A survey and classification of web phishing detection schemes." Security and Communication Networks, 9(18), 6266-6284