ISSN:2302-8149
e-ISSN:2540-9719

Perbandingan *Logistic Regression* dan *Random Forest* menggunakan *Correlation-based Feature Selection* untuk Deteksi *Website Phishing*

Comparison of Logistic Regression and Random Forest using Correlation-based Feature Selection for Phishing Website Detection

[1]Farida, [2]Ali Mustopa*
[1]Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
[2]Sistem Informasi, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
Jalan Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta Indonesia 55283
*e-mail: ali.m@amikom.ac.id

(***received*:** 27 Desember 2021, ***revised*:** 6 Februari 2022, ***accepted*:** 5 November 2022)

Abstrak

Dunia pada saat ini tengah mengalami perkembangan pada bidang teknologi dan komunikasi secara massif, terlebih pada masa pandemi saat ini yang mengharuskan semua proses pembelajaran bahkan bekerja secara daring. Hal ini yang memicu banyaknya kejahatan di dunia internet. Salah satunya yaitu mencuri data pengguna internet melalui sebuah website palsu yang dibangun seperti asli atau disebut juga website phising. Pada penelitian ini, untuk mengatasi maraknya website phising di dunia maya, maka diperlukan suatu model klasifikasi untuk mendeteksi website yang terindikasi phising dengan menggunakan kinerja terbaik dari salah satu algoritma klasifikasi logistic regression dan random forest. Untuk meningkatkan kinerja algoritma klasifikasi dilakukan seleksi fitur menggunkan metode correlation-based feature selection (CFS) untuk menyeleksi atribut yang paling berpengaruh dalam mendeteksi web phising. Berdasarkan hasil uji coba, penerapan algoritma klasifikasi logistic regression dan random forest dalam klasifikasi web phising dihasilkan akurasi sebesar 93,035 % dan 96.834 %, setelah dilakukan seleksi fitur dengan CFS akurasi yang dihasilkan menjadi 92.718 % dan 97.015 %. Dari uji coba terjadi peningkatan akurasi pada random forest sebesar 0,181 % dan terjadi penurunan yang tidak signifikan pada logistic regression. Hasil uji coba membuktikan bahwa seleksi fitur dengan CFS dapat menghilangkan atribut redundan dan dihasilkan akurasi algoritma klasifikasi yang tidak jauh berbeda ketika atribut lengkap dan *Random Forest* memiliki akurasi lebih baik setelah menggunakan CFS .

Kata kunci: website phishing, klasifikasi, logistic regression, random forest, correlation-based feature selection (CFS).

Abstract

The world is currently experiencing mass developments in information technology, especially during the current pandemic, which requires all of us to learn and even work online. They are triggered much crime in the internet world. One of them is stealing internet user data through a fake website built like the original or called a phishing website. In this research , a classification model is needed to detect phishing websites using the best performance from one of the logistic regression and random forest classification algorithms to overcome the rise of phishing websites in cyberspace. Classification performance can be improved using the correlation-based feature selection (CFS) method to select the most influential attribute in detecting web phishing. Based on the test results, applying the logistic regression and random forest classification algorithm in the classification of web phishing resulted in an accuracy of 93.035% and 96.834%. After feature selection with CFS, the accuracy was 92.718% and 97.015%, respectively. On the Testing, There was an increase in accuracy in RandomForest by 0.181% and an insignificant decrease in logistic regression. The test results prove that feature selection with CFS can eliminate redundant attributes and the resulting classification algorithm accuracy is not much different when the details are complete and Random Forest has accuracy better than after using CSF.

***Keywords**: website phishing, classification, logistic regression, random forest, correlation-based feature selection (CFS).*

1 Pendahuluan

Dunia pada saat ini tengah mengalami perkembangan pada bidang teknologi dan komunikasi secara massif, terlebih pada masa pandemi saat ini yang mengharuskan semua orang belajar bahkan bekerja secara daring. Semua teknologi berhubungan dengan internet tumbuh berkembang dengan pesat dan semakin disempurnakan dengan kehadiran pandemi. Perkembangan teknologi pasti diikuti dengan peningkatan penggunaan internet. Peningkatan inilah yang memicu banyaknya kejahatan di dunia internet. Salah satunya yaitu pencurian data pengguna internet melalui sebuah website palsu yang dibangun seperti asli atau disebut juga website *phishing*. *Phishing* merupakan salah satu tindakan kejahatan dengan cara membangun halaman website palsu dari website yang menggunakan form login[1]. Dari laporan APWG (*Anti-Phishing Working Group*), mulai pertengahan Maret 2020, penjahat dunia maya meluncurkan berbagai *phishing* bertema COVID-19 dan serangan *malware* terhadap pekerja, fasilitas layanan kesehatan, dan yang baru saja menganggur [2]. Jumlah situs *phishing* yang terdeteki pada kuartal pertama tahun 2020 adalah 165.772, naik dari 162.155 pada kuartal keempat tahun 2019[2]. Setelah dua kali lipat pada tahun 2020, jumlah *phishing* menurun selama kuartal pertama tahun 2021. Namun, Januari 2021 adalah rekor tertinggi APWG, dengan 245.771 serangan yang belum pernah terjadi sebelumnya dalam satu bulan. Sektor lembaga keuangan, webmail, dan media sosial adalah yang paling sering menjadi korban *phishing* di kuartal ini [3]. Berdasarkan survey website adalah sumber ancaman *phishing* terbanyak dan cara pencegahan yang sering dilakukan adalah *self-efficacy*[4].

Untuk mengatasi maraknya *phishing* yang terjadi di dunia maya, maka diperlukan sistem untuk mendeteksi situs *phishing* agar bisa menghindari dan mengurangi kerugian pengguna internet. Dari beberapa penelitian sebelumnya, untuk deteksi web *phising* dan *non-phising* digunakan seleksi fitur dan metode klasifikai. Metode yang digunakan diantaranya yaitu *Random Forest* dan *Logistic Regression*. Pada penelitian Susanto dilakukan deteksi web phising menggunakan metode *Binary Logistic Regression* dan menggunakan seleksi atribut yaitu *Correlation-Based Feature Selection (CFS)* dengan hasil akurasi yang tinggi [5]. *Correlation-Based Analysis* dapat digunakan untuk melakukan atribut analisis relevansi dan menyaring atribut yang secara statistik tidak relevan atau kurang relevan dari proses penambangan deskriptif[6]. Pada penelitian Irawan dilakukan perbandingan algoritma klasifikasi untuk menggidentifikasi web phising salah satunya yaitu *Random Forest* [7]. Pada penelitian ini, nilai akurasi *Random Forest* masih di bawah algoritma *Multilayer Perceptron*. Sedangkan pada penelitian Moedjahedy nilai akurasi *Random Forest* dengan menggunakan seleksi fitur Spearman dan MICe TICe lebih tinggi daripada *Logistic Regression* [8].

Pada penelitian ini menggunakan *Random Forest* dan *Logistic Regression* untuk deteksi web phising meggunakan seleksi fitur berbasis korelasi atau *Correlation-Based Feature Selection* (CFS). Metode ini digunakan peneliti karena *Random Forest* dapat meningkatkan akurasi dan dapat menangani input variabel yang besar, menyeimbangkan error dalam *Unbalanced Dataset* [9]. Sedangkan untuk *Logistic Regression* cocok digunakan untuk memprediksi keanggotaan variabel independen (prediktor) dalam dua group [10]. Sedangkan untuk CFS digunakan untuk menghapus atribut yang tidak relavan dan redundan. Penelitian ini bertujuan untuk membandingkan algoritma *Logistic Regression* dan *Random Forest*, manakah algoritma yang lebih akurat untuk deteksi web *phishing* dengan diterapkannya seleksi fitur berbasis korelasi.

2 Tinjauan Literatur

Untuk mendeteksi web phising terdapat banyak metode yang dapat digunakan. Sudah banyak peneliti yang membahas tentang web *phishing* diantaranya pada penelitian Susanto membahas tentang mendeteksi web *phishing* menggunakan binary logistic regression dan menggunakan *Correlation-Based Feature Selection* (CFS) untuk seleksi atribut. Penerapan *binary logistic regression* dalam mendeteksi web *phishing* memperoleh hasil akurasi 93,99% sebelum seleksi atribut dan 93,20% setelah seleksi atribut [5].

Penelitian terkait web *phishing* juga sempat dilakukan pada penelitian Irawan, berbeda dari sebelumnya, penelitian ini membahas identifikasi web phising dengan membandingkan algoritma

 ISSN:2302-8149
e-ISSN:2540-9719

klasifikasi. Algoritma yang dibandingkan yaitu *Support Vector Machine, Decision Tree, Random Forest*, dan *Multilayer Perceptron*. Hasil penelitian ini menunjukan tingkat akurasi *Support Vector Machine* 84.45% dan nilai AUC 0.863, tingkat akurasi *Decision Tree* 85.20% dan nilai AUC 0.92, tingkat akurasi *Random Forest* 85% dan nilai AUC 0.973, dan *Multilayer Perceptron* memiliki performa terbaik dengan tingkat akurasi mencapai 93.15% dan nilai AUC 0.976[7].

Penelitian terkait juga sempat dilakukan pada penelitian Sunge, peneliti membahas tentang mengoptimasi dari algoritma C4.5 dengan menggunakan seleksi fitur algoritma genetika untuk memprediksi web *phishing*. *Naive Bayes* sebagai classifier yang dapat digunakan untuk memprediksi sesuatu berdasarkan data termasuk untuk memprediksi apakah situs tersebut termasuk situs phising atau non-phising[11]. Algoritma genetika digunakan untuk optimasi dari hasil data *training* dan *testing* pada algoritma C4.5. Algoritma C5.4 mempunyai nilai kinerja baik akurasi, sensitifity, presisi maupun f-measure yang paling tinggi[12]. Penerapan Algoritma C4.5 dalam identifikasi web phishing memperoleh hasil 83.81% dan naik setelah menggunakan seleksi fitur algoritma gentika menjadi 86.40%[13].

Pada penelitian ini menggunakan *random forest* dan *logistic regression* untuk deteksi web phising meggunakan seleksi fitur berbasis korelasi atau *Correlation-Based Feature Selection* (CFS).

3 Metode Penelitian

a. Seleksi fitur

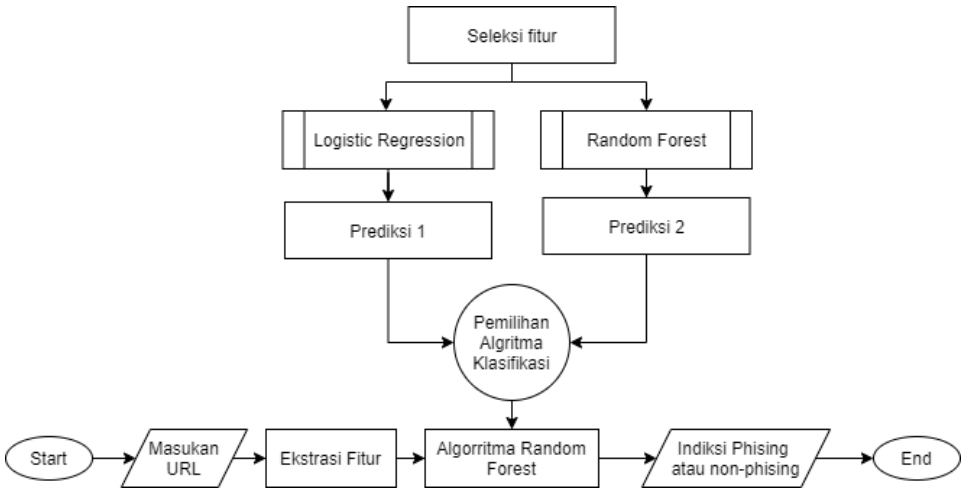
Pada penelitian ini menggunakan dataset yang diunduh dari UCI *Machine Learning Repository*. Untuk data situs *phishing* didapatkan dari *PhishTank* dan untuk situs *non-phishing* didapatkan dari Moz Top 500 dan dari Alexa. Pada penelitian ini, tahap yang dilakukan yaitu mengkonversi format dataset dari arff (*Attribute Relation File Format*) menjadi csv (*Comma Seperated Values*) yang didapat dari UCI *Machine Learning Repository*. Kemudian dilakukaan seleksi dan penghapusan atribut yang kosong atau tidak digunakan pada dataset. Selanjutnya mengekstrak fitur yang digunakan untuk deteksi web phising. Pada penelitian ini menggunakan Correlation based Featured Selection (CFS) dalam seleksi fitur

b. Uji coba

Uji coba dilakukan dengan membandingkan hasil *training* dan *testing* antara algoritma *random forest* dan *logistic regression* yang telah diterapkan seleksi fitur berbasis korelasi (CFS) dan belum diterapkan CFS. Uji coba ini dilakukan untuk menentukan algoritma mana yang terbaik yang bisa digunakan untuk mendeteksi web phishing. Uji coba ini meliputi perhitungan *accuracy*, *precision*, *recall*, dan *f-measure*.

c. Implementasi Sistem

Pada tahap ini dilakukan implementasi algoritma ke sistem yang dibuat. Algoritma yang diterapkan yaitu algoritma yang memiliki nilai akurasi terbaik. Sistem dirancang menggunakan framework Flask dari Bahasa Python, seperti pada Gambar 1 berikut ini;



Gambar 1. Perancangan Sistem

 ISSN:2302-8149
e-ISSN:2540-9719

e. Uji Coba Sistem Deteksi Website Phising

Pada tahap ini dilakukan pengujia terkait system deteksi website Phising menggunakan sampel website yang telah dituntukan. Dari hasil tersebut diharapkan dapat terlihat keberhasilan penggunaan metode yang digunakan.

4 Hasil dan Pembahasan

a. Seleksi Fitur

Teknik seleksi fitur Correlation based Featured Selection (CFS) digunakan untuk melakukan pemilihan fitur berdasarkan korelasi antar fitur, sehingga dapat meningkatkan performa dari sistem temu kembali[14]. Pada penelitian ini setelah dilakukan seleksi menggunakan CFS menyisakan 25 fitur dari yang sebelumnya berjumlah 30 dan selarar dengan penentuan fitur yang dilakukan oleh salim[15]. Fitur-fitur setelah dilakukan seleksi dapat dilihat pada Tabel 1 berikut.

Tabel 1. Hasil Seleksi Fitur

No	Fitur
0	having_IP_Address
1	URL_Length
2	Shortining_Service
3	having_At_Symbol
5	Prefix_Suffix
6	having_Sub_Domain
7	SSLfinal_State
8	Domain_registeration_length
9	Favicon
12	Request_URL
13	URL_of_Anchor
14	Links_in_tags
15	SFH
17	Abnormal_URL
18	Redirect
19	on_mouseover
20	RightClick
22	Iframe
23	age_of_domain
24	DNSRecord
25	web_traffic
26	Page_Rank
27	Google_Index
28	Links_pointing_to_page
29	Statistical_report

b. Uji Coba *Logistic Regression*

Di bawah ini adalah hasil uji coba kinerja algoritma *Logistic Regression* (LR) sebelum dan sesudah diterapkan seleksi fitur berbasis korelasi (CFS) bila dilihat dari *precision, recall*, dan *f-measure pada* Tabel 2 dan 3.

Tabel 2. Kinerja Algosritma LR Sebelum Seleksi Fitur

Kelas	Precision	Recall	F-Measure
Phishing	0,9284	0,9132	0,9207
Non-phishing	0,9318	0,9439	0,9378

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 3. Kinerja Algosritma LR Sesudah Seleksi Fitur

Kelas	Precision	Recall	F-Measure
Phishing	0,9261	0,9081	0,9170
Non-phishing	0,9280	0,9423	0,9351

Berdasarkan Tabel 2 dan 3, kinerja terbaik dari algoritma *Logistic Regression* diperoleh sebelum dilakukan seleksi fitur, namun nilai *precision*, *recall*, dan *f-measure* mengalami penurunan yang tidak signifikan bahkan cenderung stabil.

c. Uji Coba *Random Forest*

Dibawah ini adalah hasil uji coba kinerja algoritma *Random Forest* (RF) sebelum dan sesudah diterapkan seleksi fitur berbasis korelasi (CFS) bila dilihat dari *precision*, *recall*, dan *f-measure* pada Tabel 4 dan 5.

Tabel 4. Kinerja Algosritma RF Sebelum Seleksi Fitur

Kelas	Precision	Recall	F-Measure
Phishing	0,9652	0,9632	0,9642
Non-phishing	0,9708	0,9723	0,9715

Tabel 5. Kinerja Algosritma RR Sesudah Seleksi Fitur

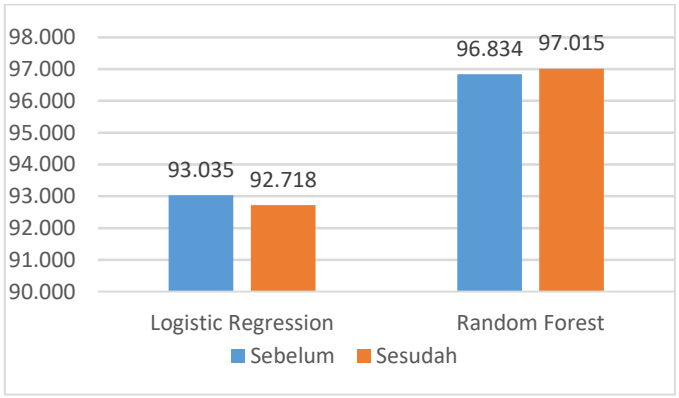
Kelas	Precision	Recall	F-Measure
Phishing	0,9711	0,9612	0,9647
Non-phishing	0,9693	0,9772	0,9733

Berdasarkan Tabel 4 dan 5, kinerja terbaik dari algoritma random forest diperoleh sesudah diterapkannya CFS dengan nilai *precision*, *recall*, dan *f-measure* lebih tinggi yang artinya lebih efektif dalam mendeteksi web phising. Berikut merupakan tabel hasil kinerja serta perbandingan algortima *Logistic Regression* (LR) dan *Random Forest* (RF) sebelum dan sesudah diterapkannya CFS. Berdasarkan Tabel 6 dapat dilihat bahwa kinerja Random Forest memiliki performa yang lebih baik dibandingkan logistic Regression.

Dari hasil pengujian kinerja akurasi pada Gambar 2, *Logistic Regression* mengalami penurunan sebesar 0,317% setelah diterapkannya seleksi fitur CFS. Sedangkan *Random Forest* mengalami peningkatan sebesar 0,181% setelah diterapkannya seleksi fitur CFS.

Tabel 6. Hasil Kinerja Akurasi LR dan RF Sebelum dan Sesudah Diterapkan CFS

Pengujian	Logistic Regression	Random Forest
Sebelum	93,035 %	96,834%
Sesudah	92,718 %	97,015 %



Gambar 2. Perbandingan Akurasi Sebelum dan Sesudah Diterapkannya CFS

 ISSN:2302-8149
e-ISSN:2540-9719

d. Sistem Deteksi *Web Phishing*

Dari hasil uji coba algoritma *Logistic Regression* dan *Random Forest* sebelum dan sesudah diterapkannya seleksi fitur CFS pada penelitian ini, diputuskan bahwa algoritma *Random Forest* menjadi algoritma pada sistem deteksi web phising karena memiliki nilai akurasi (*accuracy*), *precision, recall* dan *f-measure* terbaik dibandingkan algoritma *Logistic Regression* sebelum dan sesudah diterapkannya seleksi fitur CFS. Sistem deteksi dibangun dengan bahasa pemrograman Python menggunakan *framework flask* yang hanya memiliki satu antarmuka yaitu *form* masukkan URL seperti pada Gambar 3.

Gambar 3. Interface dari Sistem Deteksi Web Phising

Untuk melakukan deteksi, masukkan URL web pada form. Setelah dicari maka akan muncul kalimat "Bukan website phising" jika web dideteksi sebagai non-phising seperti pada Gambar 4, namun jika kalimat yang muncul berupa "Terindikasi sebagai web phising" artinya web dideteksi sebagai web phising seperti pada Gambar 5.

Gambar 4. Hasil Indikasi Bukan Web Phising

Gambar 5. Hasil Indikasi Web Phising

e. Uji Coba Sistem Deteksi Website Phising

Hasil uji coba sistem deteksi website *phishing* dari 20 sampel URL yang diuji, algoritma *random forest* dengan seleksi fitur CFS menghasilkan 17 prediksi benar antara website *phising* dan *non-phishing*. Sistem mampu memprediksi 10 web *non-phising* dengan benar dan 7 web *phishing* dengan benar, sementara 3 web *phishing* salah diprediksi sebagai *non-phishing*. Ada beberapa faktor yang mempengaruhi dalam memprediksi salah seperti dataset training, akurasi algortima, fitur pada URL, dan ketika melakukan uji coba deteksi. Berikut adalah hasil deteksi web *phishing* pada Tabel 7.

ISSN:2302-8149
e-ISSN:2540-9719

Tabel 7. Hasil Deteksi Data Sampel URL Web Phising dan Non-Phising

URL	Realita	Hasil Deteksi
https://apple.com/	Non-Phising	Non-Phising
https://paypal.com	Non-Phising	Non-Phising
https://mozilla.org/	Non-Phising	Non-Phising
http://linkedin.com/	Non-Phising	Non-Phising
https://www.adobe.com/	Non-Phising	Non-Phising
https://baidu.com/	Non-Phising	Non-Phising
http://amazon.com/	Non-Phising	Non-Phising
https://bit.ly/	Non-Phising	Non-Phising
https://google.com/	Non-Phising	Non-Phising
https://ebay.com/	Non-Phising	Non-Phising
https://anonzon.kqrz03.cn/	Phishing	Phising
https://amazom.dnwnab.shop/	Phising	Phising
https://fbcontent.com	Phising	Phising
https://y-amazon.top	Phising	Phising
https://freenltro.ru	Phising	Phising
https://lnkd.in/gVnPDt-w	Phising	Non-Phising
https://yahoomailerpotr.weebly.com/	Phising	Phising
https://www.ingareaclientesprivada.com/	Phising	Phising
https://goo.su	Phising	Non-Phising
https://unecred.com	Phising	Non-Phising

5 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dapat diambil kesimpulan bahwa penerapan seleksi fitur CFS pada peneltian ini, mempengaruhi kinerja algoritma *Logistic Regression* dan *Random Forest* seperti *accuracy*, *precision*, *recall*, dan *f-measure*. Hasil uji coba dengan menggunakan CFS mengalami kenaikan pada *Random Forest* dari nilai akurasi 96.834 % menjadi 97.015 %. Sedangkan untuk *Logistic Regression* mengalami penurunan yang tidak signifikan pada kinerja secara keseluruhan, dengan akurasi sebesar 93,035 % menjadi 92.718 %. Dari kedua algoritma tersebut lebih akurat algoritma random forest setelah diterapkan seleksi fitur CFS. Sistem deteksi website phising dengan implementasi algoritma *Random Forest* dengan menggunakan seleksi fitur CFS dilakukan uji coba deteksi menggunakan 20 sampel URL menghasilkan 17 prediksi benar terhadap prediksi phising dan non-phising.

Referensi

[1] H. Anwari and Java Creativity, *Website Hantu*. Jakarta: Elex Media Komputindo, 2011.
[2] APWG, “Phishing Activity Trends Report, 4th Quarter 2019,” 2019.
[3] APWG, “Phishing activity trends report 2nd Quarter 2021,” 2021.
[4] M. H. Wibowo and N. Fatimah, “Ancaman Phishing terhadap Pengguna Sosial Media dalam Dunia Cyber Crime,” *J. Educ. Inf. Comun. Technol.*, vol. 1, no. 1, pp. 1–5, 2017.
[5] B. M. Susanto, “Binary Logistic Regression untuk Mendeteksi Website Phising menggunakan Correlation-Based Feature Selection,” *J. Teknol. Inf. dan Terap.* , vol. 2, no. 2, pp. 255–260, Mar. 2019.
[6] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques*, 3rd ed. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.
[7] A. S. Y. Irawan, N. Heryana, H. S. Hopipah, D. Rahma, and others, “Identifikasi Website Phishing dengan Perbandingan Algoritma Klasifikasi,” *Syntax J. Inform.*, vol. 10, no. 01, pp. 57–67, 2021.
[8] J. H. Moedjahedy, A. Setyanto, and K. Aryasa, “Analisis Perbandingan Korelasi Spearman dan Maximal Information Coefficient dalam Seleksi Fitur Website Phishing menggunakan Algoritma Machine Learning,” *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 12, no. 2, p. 107, Mar. 2021, doi: 10.22303/csrid.12.2.2020.107-116.

[9] F. A. Kurniawan, Adiwijawa, and A. P. Kurniati, "Analisis dan Implementasi Random Forest dan Classification dan Regression Tree (Cart) untuk Klasifikasi pada Misuse Intrusion Detection System," Telkom University, 2012.
[10] Y. S. Nugroho and N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen terhadap Mobil menggunakan Metode Random Forest," *J. Tek. Elektro*, vol. 9, no. 1, pp. 24–29, 2017.
[11] F. E. Purwiantono and A. Tjahyanto, "Model Klasifikasi untuk Deteksi Situs Phising di Indonesia," *researchgate*, pp. 1–11, 2017.
[12] E. Nanda, Istikomah, N. A. Amari, and Y. Pristyanto, "Perbandingan Klasifikasi Algoritma K-NN, Neural Network, Naïve Bayes, C 4.5 untuk Mendeteksi Web Phising," *FAHMA*, vol. 16, no. 3, pp. 33–42, Sep. 2018.
[13] A. S. Sunge, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradig. - J. Komput. dan Inform.*, vol. 20, no. 2, pp. 27–32, Dec. 2018.
[14] Y. A. Sari, R. K. Dewi, and C. Fatichah, "Seleksi Fitur menggunakan Ekstraksi Fitur Bentuk, Warna, dan Tekstur dalam Sistem Temu Kembali Citra Daun," *JUTI J. Ilm. Teknol. Inf.*, vol. 12, no. 1, p. 1, Jan. 2014, doi: 10.12962/j24068535.v12i1.a39.
[15] T. Salim and Y. C. Giap, "Data Mining Identifikasi Website Phising menggunakan Algoritma C4.5," *TAM (Technology Accept. Model.*, vol. 8, no. 2, 2017.