Jurnal E-Komtek
Vol.8 No. 2(2024) pp. 454-465
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Phishing Website Detection Using the Decision Tree Algorithm Method

Stefanus Eko Prasetyo[1], Haeruiddin[2], Arron[3]

[1-3]Department of Information Technology, Universitas Internasional Batam, Indonesia, 29426

2132034.arron@uib.edu

https://doi.org/10.37339/e-komtek.v8i2.2167

Published by Politeknik Piksi Ganesha Indonesia

Artikel Info
Submitted:
06-12-2024
Revised:
26-12-2024
Accepted:
29-12-2024
Online first :
29-12-2024

Abstract

Along with the increasing number of internet users and the rapid development of technology, cyber security threats are becoming more complex, including phishing threats that often cause major losses such as loss of individual or corporate privacy. This study aims to identify phishing websites effectively by applying machine learning algorithms. The dataset used in this study comes from the UCI learning repository developed by the University of Huddersfield. The research methodology includes the stages of problem identification, Cart algorithm collection, validation, and model evaluation. With this method, the study found that the CART algorithm achieved an accuracy level of 90.5% in detecting phishing sites. These results show cyber security, especially in protecting users from phishing threats, this study is expected to contribute to improving data protection and privacy of internet users, as well as encouraging the application of machine learning technology in a more adaptive cyber security system.

Keywords: *Internet users, Dataset, Phishing.*

Abstrak

Seiring dengan meningkatnya jumlah pengguna internet dan pesatnya perkembangan teknologi ancaman keamanan siber pun semakin kompleks, ternasuk ancaman phising yang sering kali menyebabkan kerugiaan besar seperti kehilangan privasi individu atau persuahaan. Penelitian ini bertujuan untuk mengidentifikasi situs web phising secara efektif melalui penerapan algoritma machine learning. Dataset yang digunakan dalam penelitian ini berasal dari repositori pembelajaran UCI yang dikembangkan oleh University of Huddersfield. Metodologi penelitian mencakup tahapan indentifikasi masalahm, pengumpulan algoritma Cart, validasi, dan evaluasi model. Dengan metode ini, penelitian menemukan bahwa algoritma CART mampu mencapai tingkat akurasi 90.5% dalam mendeteksi situs phising. Hasil ini menunjukan keamanan siber, terutama dalam melindungi pengguna dari ancaman phising, penelitian ini di harapkan dapat berkontribusi pada peningkatan proteksi data dan privasi pengguna internet, serta mendorong penerapan teknologi machine learning dalam system keamanan siber yang lebih adaptif

Kata-kata kunci: Pengguna internet, Dataset, Phising.

1. Introduction

In the increasingly advanced digital era, information technology's progress has transformed how humans interact, access information, and perform various activities, including transactions. The internet and electronic devices, such as smartphones and computers, have become inseparable from daily life **[1], [2]**. The accessibility of this technology spans various sectors, from education, health, and entertainment to business and banking **[3]**. Using the internet as a global information medium is highly beneficial in supporting various human activities; however, this technology also opens opportunities for cybercrime. One of the most frequent threats is phishing, which aims to steal users' personal and financial information through various forms of digital manipulation [4].

Phishing is a cyber fraud activity where perpetrators or phishers attempt to obtain sensitive victim data through deceit, such as sending emails resembling trusted sources or creating fake websites almost identical to official ones [5]. For instance, phishers often impersonate banks or popular online services to deceive victims **[6]**. They use links in emails that, when clicked, direct the victim to fake websites designed to steal information, such as credit card numbers, passwords, or other personal information. Phishing evolves alongside technological advances, with increasingly sophisticated methods that make it difficult for victims to distinguish between genuine and fake sites **[7]**.

Behind this complexity, a significant challenge in combating phishing is the perpetrators' ability to continuously change their techniques and tools, making them difficult to detect by traditional security systems [8]. Conventional security approaches, such as using blacklist methods, are often limited in detecting new phishing sites because blacklists only cover already identified sites or IP addresses [9]. Additionally, heuristic methods that rely on specific patterns and rules often produce false positives, where the system detects threats that are not actually harmful **[10]**.

To address the need for a more adaptive approach, machine learning technology emerges as a promising solution for detecting phishing **[11]**. Machine learning enables computers to "learn" from data without being explicitly programmed, allowing them to recognize complex patterns and characteristics in phishing detection. Some popular machine learning algorithms for detecting phishing include Decision Tree, Naïve Bayes, and Random Forest **[12]**. These algorithms have proven capable of processing large datasets and recognizing patterns in phishing sites with high accuracy, enabling efficient phishing threat detection.

In this research, Decision Tree is chosen as the primary algorithm to be further developed for phishing detection because it can classify data based on rules that are easy to understand and implement **[13]**. This algorithm can categorize data based on specific features or characteristics of phishing sites **[14]**. Furthermore, this study employs filter methods, such as Pearson correlation, to select relevant features to identify phishing sites. This approach improves the model's accuracy and reduces the processing time required by eliminating irrelevant features from the analysis.

To produce a reliable model, this study tests several machine learning algorithms to compare their effectiveness. In addition to Decision Tree, Naïve Bayes and Random Forest are also applied for comparison [15]. Decision Tree is chosen because this algorithm demonstrates high accuracy in phishing detection, reaching 90.5% accuracy with the CART algorithm [16]. With this level of accuracy, the model is expected to contribute to providing a more effective cybersecurity system, particularly in protecting users from phishing threats.

This study has broad objectives, not only to develop an accurate Decision Tree-based phishing detection model but also to evaluate the model's performance compared to other methods and identify the key features that most contribute to phishing site detection **[17]**. Thus, the resulting model will be more adaptive to new phishing threats and capable of detecting new patterns of attacks that may not have existed before [18].

As dependence on digital services increases, cybersecurity becomes increasingly important to maintain user trust and prevent financial, reputational, and other losses due to data theft [19]. Implementing technologies such as machine learning in phishing detection is expected to provide better protection for internet users. This study is expected to significantly contribute to protecting users from the evolving threat of phishing, thereby fostering a safer and more trustworthy digital ecosystem for everyone **[20]**.

2. Method

The method applied in this research is the Classification and Regression Trees (CART) Algorithm **[21]**. CART is a methodology proven to be effective and suitable for data prediction and classification purposes **[22]**. The advantage of the CART Algorithm lies in its non-parametric nature, making it well-suited for numerical data **[23]**. In implementing the CART Algorithm, each row or record in the data is classified based on the values of the predictor variables to achieve the intended target variable [24].

A. Data Collection

The dataset used in this research is sourced from the UCI Machine Learning Repository. The dataset consists of 2,456 website examples with 31 attributes. These attributes cover various aspects of the websites, such as URL length, the presence of special characters, and certificate status.

Table 1. *Dataset* Based on *Class*

No	Klasifikasi	Jumlah Record *Dataset*
1	Bukan *Phishing*	1362
2	*Phishing*	1094
	Jumlah	2456

The following is a brief explanation of some attributes in the dataset:

1) Having_IP_Address: Adanya IP address sebagai domain pada URL (biner) -1 (tidak), 1 (iya).
2) URL_Length: Panjang URL (polinomial) -1 (kurang dari 54 karakter), 0 (antara 54-75 karakter), 1 (lebih dari 75 karakter).
3) Shortining_Service: Penggunaan layanan penyingkatan URL (biner) 0 (tidak), 1 (iya).
4) Having_At_Symbol: Penggunaan simbol "@" pada URL (biner) 0 (tidak), 1 (iya).
5) Double_slash_redirecting: Penggunaan simbol "//" untuk mengalihkan website (biner) 0 (tidak), 1 (iya).
6) Prefix_Suffix: Penggunaan simbol "-" pada domain dalam URL (biner) -1 (tidak), 1 (iya).
7) Having_Sub_Domain: Penggunaan subdomain (polinomial) -1 (tidak punya), 0 (1 subdomain), 1 (lebih dari 1 subdomain).
8) SSLfinal_State: Status sertifikat SSL (polinomial) -1 (tepercaya), 0 (tidak tepercaya), 1 (tidak ada).
9) Domain_registration_length: Masa berlaku domain (biner) 0 (lebih dari 1 tahun), 1 (kurang dari 1 tahun).
10) Favicon: Penggunaan favicon dari link eksternal (biner) 0 (tidak), 1 (iya).
11) Port: Penggunaan port tertentu seperti 21, 22, 23, 445 (biner) 0 (tidak), 1 (iya).
12) HTTPS_token: Penggunaan HTTPS dalam bagian domain pada URL (biner) 0 (tidak), 1 (iya).
13) Request_URL: Persentase permintaan URL eksternal (polinomial) -1 (kurang dari 22%), 0 (22%-61%), 1 (lebih dari 61%).

14) URL_of_Anchor: Persentase penggunaan tag <a> yang mengarah ke domain berbeda (polinomial) -1 (kurang dari 31%), 0 (31%-67%), 1 (lebih dari 67%).
15) Links_in_tags: Persentase penggunaan tag <link>, <meta>, dan <script> yang mengarah ke domain berbeda (polinomial) -1 (kurang dari 17%), 0 (17%-81%), 1 (lebih dari 81%).
16) SFH: Domain pemrosesan Server Form Handler (polinomial) -1 (pada domain yang sama), 0 (pada domain yang berbeda), 1 (kosong).
17) Submitting_to_email: Penggunaan fungsi "mail()" atau "mailto" dalam PHP untuk mengirim informasi user (biner) 0 (tidak), 1 (iya).
18) Abnormal_URL: Kecocokan website dengan catatannya yang ditunjukkan pada basis data WHOIS (biner) -1 (cocok), 1 (tidak cocok).
19) Redirect: Jumlah pengalihan website yang dilakukan (polinomial) -1 (kurang dari 2 kali), 0 (2-4 kali), 1 (lebih dari 4 kali).
20) on_mouseover: Perubahan status bar ketika event onMouseOver aktif (biner) 0 (tidak), 1 (iya).
21) RightClick: Keadaan klik kanan pada website (biner) 0 (diaktifkan), 1 (dinonaktifkan).
22) popUpWindow: Penggunaan pop-up window untuk meminta user mengisi data mereka (biner) 0 (tidak), 1 (iya).
23) Iframe: Penggunaan fungsi iframe (biner) 0 (tidak), 1 (iya).
24) age_of_domain: Umur domain (biner) -1 (lebih dari atau sama dengan 6 bulan), 1 (kurang dari 6 bulan).
25) DNSRecord: Adanya catatan DNS pada domain (biner) -1 (ada), 1 (tidak ada).
26) web_traffic: Rank lalu lintas website dalam basis data Alexa (polinomial) -1 (di atas 100.000), 0 (di bawah 100.000), 1 (tidak terdaftar).
27) Page_Rank: Nilai PageRank website (biner) -1 (lebih dari atau sama dengan 0.2), 1 (kurang dari 0.2).
28) Google_Index: Adanya website dalam indeks pencarian Google (biner) -1 (iya), 1 (tidak).
29) Links_pointing_to_page: Jumlah link eksternal yang menunjuk ke website (polinomial) -1 (lebih dari 2), 0 (1 atau 2), 1 (tidak ada).
30) Statistical_report: Host berasal dari Top Phishing IPs atau Top Phishing Domains yang dibuat oleh beberapa pihak seperti StopBadware dan PhishTank (biner) -1 (tidak), 1 (iya).
31) Result (Label): Hasil identifikasi website (biner) -1 (bukan phishing), 1 (phishing).

B. Pre-Processing Stage

Pre-processing is an essential step to ensure that the data is ready for use in the machine learning model. The pre-processing steps used are as follows:

- Handling Missing Values: Missing values are identified and filled using imputation methods such as mean/mode/median, or removed if the percentage of missing values is very small.
- Data Normalization: Data is normalized using techniques such as Min-Max Scaling or Z-score Normalization to ensure that each attribute is on the same scale.
- Data Splitting: The dataset is divided into two subsets: training data (80%) and testing data (20%) using stratified sampling to ensure the same class distribution in both subsets.

C. Implementation of the Decision Tree Algorithm

Decision Tree is a popular and effective classification method for various types of data. To ensure that this method can be reproduced

1. Modeling

The Decision Tree algorithm is implemented using Scikit-learn in Python. This library is chosen for its powerful capabilities and comprehensive documentation.

2.Parameter Setting

- Criterion: Gini Impurity atau Entropy digunakan sebagai kriteria pemisahan.
- Max Depth: Menentukan kedalaman maksimum pohon untuk menghindari overfitting. Misalnya, max_depth=10.
- Min Samples Split: Jumlah minimum sampel yang diperlukan untuk memisahkan node internal, misalnya, min_samples_split=2.
- Min Samples Leaf**:** Jumlah minimum sampel yang diperlukan di node daun, misalnya, min_samples_leaf=1

3.Training

- Model dilatih menggunalkan data dengan parameter yang telah di tentukan
- ‘fron sklearb.tree inport DecisionTreeClassifier’
- ‘model.fit(X_train, y_train)’

4. Cross-Validation:

- Teknik k-fold cross-validation (misalnya, 10-fold) digunakan untuk mengevaluasi model dan memastikan bahwa model tidak overfitting.
- ‘From sklearn.model_selection import cross_val_score’
- ‘scores = cross_val_score(model, X_train, y_train, cv=10)’

D. Confusion Matrix:

The performance evaluation tool used in machine learning classification models measures the model's performance. This is a table that illustrates the classification algorithm's performance by showing the number of correct and incorrect predictions made by the model compared to the actual values of the tested data.

Table 2. *Confusion Matrix*

Actual\Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

True Positive (TP): Correct prediction for the positive class (a phishing website detected as phishing).

True Negative (TN): Correct prediction for the negative class (a non-phishing website detected as non-phishing).

False Positive (FP): Incorrect prediction for the positive class (a non-phishing website detected as phishing).

False Negative (FN): Incorrect prediction for the negative class (a phishing website detected as non-phishing).

From the confusion matrix, we can calculate several important performance metrics.

1. Precisoin

The proportion of true positive predictions. Precision =

$$\frac{TP}{TP+FP}$$

2. Recall

The proportion of true positive predictions. Recall =

$$\frac{TP}{TP+FN}$$

3. F-Meassure

The proportion of true positive predictions. Recall =

$$\text{F1-Score} = 2\ X\ \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

4. Accuracy

The proportion of true positive predictions. Accuracy =

$$\frac{TP+TN}{TP+FN+FP+FN}$$

E. Calculation of the Confusion Matrix

- Confusion Matrix Class 1

True Positive (TP) = 1046

False Negative (FN) = 48

False Positive (FP) = 68

True Negative (TN) = 1294

- B. Confusion Matrix Class -1

True Positive (TP) = 1294

False Negative (FN) = 68

False Positive (FP) = 48

True Negative (TN) = 1046

Total accuracy is calculated by combining all relevant values. The total True Positive (TP) and True Negative (TN) are calculated as the sum of 1046 and 1294, resulting in 2340. Meanwhile, the total False Negative (FN) and False Positive (FP) are calculated as the sum of 48, 68, 68, and 48, resulting in 232.

The total accuracy is then calculated by dividing the total TP and TN by the total sum of TP, TN, FN, and FP.

$$\text{Akurasi} = \frac{2340}{2340+232} = \frac{2340}{2572} = 0.910$$

The final results of this calculation are presented in a table that summarizes the precision, recall, F-Measure, and accuracy values of the tested classification model:

Table 5. The Accuracy value *of the phishing website dataset*

Class	Precision	Recall	F-Measure	Akurasi
1	0,939	0.956	0.947	0.910
-1	0.964	0.950	0.957	0.910
Weighted Avg	0.953	0.953	0.953	0.910

This table shows that the precision, recall, and F-Measure values are calculated from the phishing website dataset using the CART Algorithm and analyzed with WEKA 3.9 software. The results indicate that the model performs well in classifying the data, with an overall accuracy of 91%.

The confusion matrix is a critical tool in evaluating classification models as it provides a direct overview of prediction accuracy. By containing information on TP, TN, FP, and FN, the confusion matrix allows for in-depth analysis of the model's performance, error identification,

and threshold setting optimization. This is crucial in improving the accuracy and relevance of the model in various practical applications such as cybersecurity and healthcare.

F. Validation and Evaluation

The validation and evaluation stage aims to measure the accuracy of the generated model. The techniques used in this research include the confusion matrix and 10-fold cross-validation.

G. Conclusion

This research focuses on detecting phishing websites using the Decision Tree algorithm, specifically the CART Algorithm, with analysis conducted using WEKA 3.9 software. Based on the results obtained, this study emphasizes that the Decision Tree algorithm, particularly the CART Algorithm, is an effective and reliable tool for detecting phishing websites. With high accuracy results and other evaluation metrics showing strong performance, this model can be widely used to enhance internet security and protect users from phishing threats.

3. Result

Phishing website detection is a critical issue in cybersecurity, as phishing attacks can lead to financial losses and personal data theft [25]. In this study, the CART (Classification and Regression Trees) Algorithm is applied to develop a classification model that can detect phishing websites with a high level of accuracy. Below is a more detailed discussion of the results obtained and their implications.

1. Total Accuracy

The developed model achieved an overall accuracy of 91% [25]. This means that the model is able to classify 91% of all data correctly, both as phishing and non-phishing. This high level of accuracy indicates that the CART Algorithm is highly effective for binary classification tasks in the context of phishing detection.

2. Precision dan Recall

The high precision and recall values for both classes (phishing and non-phishing) show that the model is not only accurate in its predictions but also consistent in detecting phishing cases. The precision for class "1" (phishing) is 0.939, meaning that 93.9% of all positive predictions made by the model are truly phishing websites [12]. The recall for class "1" is 0.956, indicating that the model successfully detected 95.6% of all actual phishing websites in the dataset [13]. Both metrics are crucial in the context of cybersecurity, as failure to detect phishing websites (false negatives) or misidentifying safe websites as phishing (false positives) can have serious consequences [14].

3. F-Measure

The F-Measure value, which is the harmonic mean of precision and recall, is also high for both classes (0.947 for class "1" and 0.957 for class "-1") [25]. This indicates that the model has a good balance between precision and recall, demonstrating consistent and reliable performance.

4. Conclusion

This research aims to develop a phishing detection model based on machine learning algorithms, with a primary focus on the Decision Tree algorithm. The resulting model demonstrates an adaptive ability to detect phishing with a high accuracy rate of 90.5%. This result is supported by feature selection using the Pearson correlation method, which helps identify the most relevant features in distinguishing phishing sites. These findings indicate that the Decision Tree algorithm has great potential for application in more effective cybersecurity systems, especially in addressing evolving threat patterns.

This study provides a significant contribution to the field of cybersecurity, particularly in utilizing machine learning technology to enhance user protection against phishing threats. With its adaptive approach, the model is capable of meeting the need for a smarter and more responsive detection system to address the evolution of phishing techniques. For further development, it is recommended that similar studies include analysis of larger datasets and the use of more complex algorithms to improve detection performance. Additionally, integrating the model with real-time-based security systems could be a strategic step to enhance the reliability of user protection in the digital age.

References

[1] A. Arifin and H. Rahmawati, "SISTEM PAKAR DIAGNOSA PHISING DENGAN METODE CERTAINTY FACTOR BERBASIS WEB."

[2] S. Sekolah, T. Manajemen Informatika, D. Komputer, and W. Utama, "PENELITIAN KOMPARASI ALGORITMA KLASIFIKASI DALAM MENENTUKAN WEBSITE PALSU," vol. 1, no. 1, pp. 2598–294, 2017, [Online]. Available: http://idsirtii.or.id

[3] D. R. K. Saputra, Y. V. Via, and A. N. Sihananto, "DETEKSI ANOMALI MENGGUNAKAN ENSEMBLE LEARNING DAN RANDOM OVERSAMPLING PADA PENIPUAN TRANSAKSI KEUANGAN," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4910.

[4] A. Susilo Yuda Irawan, N. Heryana, H. Siti Hopipah, D. Rahma Putri, and J. Hs Ronggo Waluyo Puseurjaya Telukjambe Timur Karawang Jawa Barat, "Identifikasi Website Phishing dengan Perbandingan Algoritma Klasifikasi," 2021. [Online]. Available: www.phishtank.com

[5] "Klasifikasi Malicious URL Menggunakan Algoritma Improved Random Forest dan Random Forest Berbasis Web," *Jurnal Sains dan Informatika*, vol. 9, no. 1, pp. 8–14, Apr. 2023, doi: 10.22216/jsi.v9i1.1378.
[6] S. Diantika, "PENERAPAN TEKNIK RANDOM OVERSAMPLING UNTUK MENGATASI IMBALANCE CLASS DALAM KLASIFIKASI WEBSITE PHISHING MENGGUNAKAN ALGORITMA LIGHTGBM," 2023.
[7] F. Bela Fransiska and F. B. Tobing, "Securing Indonesia Cyber Space: Strategies for Cyber Security in the Digital Era," *Cyber Security in the Digital Era. JSSP*, vol. 7, no. 1, pp. 50–62, 2023.
[8] A. H. Nasrullah, "IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI PRODUK LARIS," vol. 7, no. 2, 2021, [Online]. Available: http://ejournal.fikom-unasman.ac.id
[9] "Memperkuat Pertahanan Siber Guna Meningkatkan Ketahanan Nasional."
[10] B. R. Sanjaya *et al.*, "PENGEMBANGAN CYBER SECURITY DALAM MENGHADAPI CYBER WARFARE DI INDONESIA," 2022.
[11] A. Maria De Liguori Sakunab and A. Silvia, "Halaman: 195-208 Terakreditasi Peringkat 5 (SINTA 5) sesuai SK RISTEKDIKTI Nomor," 2021. [Online]. Available: http://ejurnal.ubharajaya.ac.id/index.php/JKI
[12] "Simorangkir et al. - 2023 - CYBER SECURITY DALAM STUDI KEAMANAN NASIONAL POLI".
[13] K. Marcello Jonathan, B. Mulyawan, and N. Jaya Perdana, "Jurnal Ilmu Komputer dan Sistem Informasi PERBANDINGAN KINERJA ALGORITMA NAÏVE BAYES DAN C4.5 UNTUK MENDETEKSI PENGELABUAN UNIFORM RESOURCE LOCATOR (PHISHING URL)."
[14] D. Napitupulu and M. Kom, "Kajian Peran Cyber Law Dalam Memperkuat Keamanan Sistem Informasi Nasional."
[15] A. Razzaq, M. Aditya, A. Widya, O. Kuncoro Putri, D. L. Musthofa, and P. Widodo, "Serangan Hacking Tools sebagai Ancaman Siber dalam Sistem Pertahanan Negara (Studi Kasus: Predator)," *Global Political Studies Journal*, vol. 6, 2022, doi: 10.34010/gpsjournal.v6i1.
[16] A.- Husaini, I. Hariyanti, and A. R. Raharja, "Perbandingan Algoritma Decision Tree dan Naive Bayes dalam Klasifikasi Data Pengaruh Media Sosial dan Jam Tidur Terhadap Prestasi Akademik Siswa," *Technologia : Jurnal Ilmiah*, vol. 15, no. 2, p. 332, Apr. 2024, doi: 10.31602/tji.v15i2.14381.
[17] A. F. Mahmud and S. Wirawan, "Sistemasi: Jurnal Sistem Informasi Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi Phishing Website Detection using Machine Learning Classification Method." [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[18] A. Ferdita Nugraha, R. Faticha, A. Aziza, and Y. Pristyanto, "Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing," vol. 7, no. 1.
[19] R. P. Ramadhan and T. Desyani, "Implementasi Algoritma J48 Untuk Identifikasi Website Phising," *Teknik dan Multimedia*, vol. 1, no. 2, 2023.
[20] A. F. Rosy, "Kerjasama Internasional Indonesia: Memperkuat Keamanan Nasional di Bidang Keamanan Siber," *Journal of Government Science (GovSci) : Jurnal Ilmu Pemerintahan*, vol. 1, no. 2, pp. 118–129, Jul. 2020, doi: 10.54144/govsci.v1i2.12.
[21] "29819".
[22] D. Ciri Link Phishing Menggunakan Algoritma Random Forest Untuk Meningkatkan Keamanan Cyber, Y. Purnomo, C. Herdian, A. Radiatul Kamila, and S. Virginia, "Techno

Xplore Jurnal Ilmu Komputer dan Teknologi Informasi." [Online]. Available: https://www.kaggle.com/

[23] V. Aprelia Windarni, A. Ferdita Nugraha, S. Tri Atmaja Ramadhani, D. Anisa Istiqomah, F. Mahananing Puri, and A. Setiawan, "DETEKSI WEBSITE PHISHING MENGGUNAKAN TEKNIK FILTER PADA MODEL MACHINE LEARNING," 2023.

[24] Y. W, Y. B. Fitriana, S. Esabela, and F. Hamdani, "Deteksi Serangan Malware Pada Web Aplikasi Menggunakan Metode Malware Analis Dinamis dan Statis," *Digital Transformation Technology*, vol. 4, no. 1, pp. 461–470, Jul. 2024, doi: 10.47709/digitech.v4i1.4270.

[25] A. T. Zy, A. T. Sasongko, and A. Z. Kamalia, "Penerapan Naïve Bayes Classifier, Support Vector Machine, dan Decision Tree untuk Meningkatkan Deteksi Ancaman Keamanan Jaringan," *Media Online)*, vol. 4, no. 1, pp. 610–617, 2023, doi: 10.30865/klik.v4i1.1134.