Comparative Analysis of Neural Network and Bagging Algorithms for Detecting Phishing Sites in Indonesia

Shakira Karin indrawan
Computer Science Program, School of Computer Science
Binus University
Jakarta, Indonesia
shakira.indrawan@binus.ac.id

Eyzar Hermanio
Computer Science Program, School of Computer Science
Binus University
Jakarta, Indonesia
eyzar.hermanio@binus.ac.id

Naila Az Zahra
Computer Science Program, School of Computer Science
Binus University
Jakarta, Indonesia
naila.zahra@binus.ac.id

Abstracts **- Phishing is a cybercrime that involves sharing links with internet users to retrieve personal information. The increased number of internet users in Indonesia has become a breeding ground for internet criminals to carry out their actions. This research was conducted to detect which sites are classified as phishing sites in Indonesia. The data used in this research amounted to 300 records divided into 200 records (67%) for training data and 100 records (33%) for testing data. The test results with 30 features that are used, show that the Bagging approach outperforms the Neural Network method in detecting phishing sites, where Bagging method having higher Precision, Recall, F1-Score, and Accuracy values.**

Keywords *— phishing, neural network, bagging, website, prediction , classification*

I. INTRODUCTION

It is undeniable that internet users are increasing every year [1]. Using the Internet, which makes it easier for humans to carry out activities in various fields of life is one of the reasons for the increasing use of the Internet. The latest data shows that the percentage of internet users in Indonesia has reached 78.19 percent in 2023 or, more precisely, 215,626,157 people, where the total population of Indonesian society is currently 257,773,901 people. These data show that more than half of Indonesia's population uses the Internet to carry out various activities in their lives [2]

The increase in internet users in Indonesia has become a breeding ground for internet criminals to carry out their actions. The ease of finding information and the lack of data security on the internet are loopholes for criminals to intensify their actions. Confidential data such as personal data, passwords, and e-mails from internet users is stolen to generate income [3]. Data theft by internet criminals is usually carried out by exploiting phishing sites. According to data provided by detik.com, phishing cases in Indonesia have touched 34,662 over the last five years, which is worrisome [4].

Phishing is a term that describes cybercrime by sharing links with internet users to retrieve personal data and so on [5]. Meanwhile, a phishing site is designed by internet criminals to be as authentic as possible to the original site. Appearance, content, and URLs are made as authentic as possible to trick the victim into feeling they are accessing the site from a legitimate source [3]. Usually, this phishing attack starts with an email that appears to be sent from an official organization to the victim to follow the attached URL link. If internet users are successfully tricked into entering

the requested information, at that time, the criminals have succeeded in stealing the victim's data [5]

Information obtained from victims of this phishing site is easily used by internet criminals to carry out very detrimental activities to victims. As a result, victims feel that accessing the internet is no longer safe, and many victims experience financial losses. In addition, phishing sites are often used as fraudulent sites on behalf of a legitimate company or organization's website and eventually become a medium for spreading viruses/malware on computers [3]. Of the many problems, this research was conducted to detect which sites are classified as phishing sites to minimize victims of phishing crimes in Indonesia. Several methods can detect phishing websites, including the KNN method, Neural Networks, Naive Bayes, Decision Trees, Bagging, etc. Based on Sunaryono's research entitled "*Penelitian Komparasi Algoritma Klasifikasi dalam Menentukan Website Palsu*" said that Neural Networks are the primary choice for determining phishing websites. Meanwhile, in Febry Eka Purwiantono's research, she said that Bagging was the best algorithm for determining phishing websites. *Neural Network* is a computing system in which architecture and operations are inspired by knowledge of biological nerve cells in the brain [5]. While the Bagging method is an abbreviation method for bootstrap aggregating where this method is a classification algorithm for decision-making using several voices combined into a single prediction [3]. This research compared two methods with superior accuracy in detecting phishing sites from previous studies: the Bagging Method and Artificial Neural Network (ANN). By doing this research, it is expected to get the highest accuracy value in detecting phishing sites.

II. Literature Review

This research is based on several previous studies that also discuss phishing. According to Pungkas Subarkah (2021) in his research entitled "Identifikasi Website Phising Menggunakan Algoritma Classification and Regression Trees (CART)" phishing is threatening and tricky by luring someone to indirectly provide information to the trapper, which has the potential to cause harm [6]. This is in line with the opinion of Febry Eka Purwiantono (2017) that phishing sites cause fear and decrease the trust of internet users in online transactions. Therefore, researchers created a system that can accurately detect phishing sites to minimize losses by using classification methods, Naïve Bayes, Bagging, and Multilayer Perceptron. As a result, the bagging method is the best compared to other algorithms because it excels in False Positive, True Negative, Accuracy, Precision, Recall, and F-Measure values [3]. This bagging method was also used by Alanazi Rayan (2022) in his research entitled "Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique" The results obtained an overall rate of 95% in detecting email spam, where the results were superior when only using the j48 method [7].

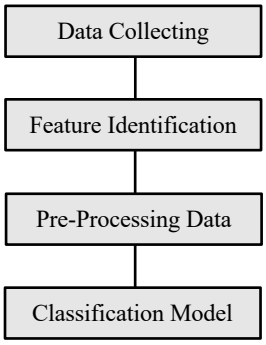
The bagging method is widely used by several researchers, including Winalia Agwil and friends, who examined the accuracy of student graduation time. The study concluded that the bagging cart method could improve classification performance, seeing from the increase in total accuracy value and sensitivity value [8]. In addition, another study with a similar method was also conducted by Tyas Setiyorini and friends, who examined the application of the bagging method to reduce data noise in neural networks for estimating concrete compressive strength. The researchers found that the bagging method has a better performance or accuracy level of concrete compressive strength estimation than the individual neural network method [9].

In contrast to research by Aswan Supriyadi Sunge (2022), which predicts phishing in website security with machine learning comparisons, researchers do not use the bagging method but the Decision Tree, Naïve Bayes, Multilayer Perceptron (Neural Network), K-Nearest Neighbor, and Support Vector Machine methods. Some of these methods turned out to produce a higher level of accuracy using the Neural Network method [5]. Similarly, in a study entitled "Klasifikasi Situs Phishing dengan Menggunakan Neural Network dan K-Nearest Neighbor", researchers used the Neural Network method with a dataset of 11,055 sites, producing the best accuracy compared to other methods [10]. Similar research by Sunaryono (2017), which compares classification algorithms in determining fake websites with the decision tree method (C4.5), Naive Bayes, K-Nearest Neighbor, Support Vector Machine, and Neural Network using 2,456 data with 30 variables, concluded that the Support Vector Machine and Neural Network methods have the highest accuracy [11].

Research related to the Neural Network method was also conducted by Eko Prasetiyo Rohmawan (2018), who predicted student graduation on time with the decision tree method and Artificial Neural Network, resulting in an accuracy rate of 79.74% [12]. In addition, there is also research on the classification of public opinion regarding covid-19 by Euis Saraswati and friends using the same method and producing an accuracy value of 88.62%, precision of 91.5%, and recall of 95.73%. According to him, the method is good enough for text-mining classification [13].

III. RESEARCH METHOD

The methodology in this study consist of several stages. The following is an overview of the methodological stages carried out:



Step Of Research Method

A. Data Collecting

In this research, data searches in the form of URLs of phishing and non-phishing sites via the internet, email, and references from previous research. The source of the internet comes from [14]. In this study only used 500 sites, each consisting of 375 (75%) training data and 125 (25%) testing data. Below are examples of phishing and non-phishing sites that were successfully obtained:

Phishing Sites
http://shadetreetechnology.com/V4/validation/a111aedc8ae390eabcfa130e041a10a4
https://support-appleld.com.secureupdate.duilawyeryork.com/ap/89e6a3b4b063b8d/?cmd=_update&dispatch=89e6a3b4b063b8d1b&locale=_
http://appleid.apple.com-app.es/
http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531
http://html.house/l7ceeid6.html
http://wave.progressfilm.co.uk/time3/?logon=myposte
http://beta.kenaidanceta.com/postamok/d39a2/source
http://www.ktplasmachinery.com/cs/
http://batvrms.net/deliver/D2017HL/u.php
https://polarklimatsgserver.blogspot.com/

List of phishing sites table

Non-phishing Sites
http://www.crestonwood.com/router.php https://www.missfiga.com/
http://rgipt.ac.in
http://www.iracing.com/tracks/gateway-motorsports-park/
http://www.mutuo.it
http://vamoaestudiarmedicina.blogspot.com/
https://parade.com/425836/joshwigler/the-amazing-race-host-phil-keoghan-previews-the-season-27-premiere/
https://www.astrologyonline.eu/Astro_MemoNew/Profilo.asp
https://www.lifewire.com/tcp-port-21-818146
https://technofizi.net/top-best-mp3-downloader-app-for-android-free-music-download/
https://www.missfiga.com/

List of non-phishing sites table

B. Feature identification

Before classifying, the step that needs to be done in this research is to determine the features. The determination/selection of features is intended to obtain features that support the detection of phishing sites so that the resulting research is more accurate. These features are selected based on several literature studies conducted in this research.

In this study, 30 features will be used; here's an explanation of some of these features:

1. Contains an IP address

   In some previous studies, it was said that a phishing site usually uses an IP address in the hostname section of the site where the IP address is used to steal someone's personal information [10]. If the site contains an IP address, it will be given a value of 1. But if not, it will be given a value of -1.

2. Contains the '@' symbol

   If a site contains the '@' symbol, the symbol causes the browser to ignore anything before the '@' symbol. Therefore, if a site has an '@' symbol, the site will be classified as a phishing site [10]. If the site contains the '@' symbol, it will be given a value of 1. But if not, it will be given a value of -1.

3. Prefixes and suffixes

   There are four types of affixes, namely affixes in front (prefixes), affixes in the middle (infixes), affixes at the back (suffixes), and affixes in front and at the back (confixes). Phishing sites generally use two domain suffixes, whereas internet users usually only pay attention to the first suffix and ignore the others. [15]. This research will only use two types of affixes (prefixes and suffixes) to detect phishing sites.

4. Domain age

   The domain's lifetime is calculated from the time the domain is registered. The younger the domain age, the more suspicious it will be as a phishing site and will be assigned a value of 1 and vice versa. The attributes of this feature are numeric data types.

5. Is the domain private?

   If a domain from a site is privatized, then it could be that the site is indicated as a phishing site because usually, the domain of a phishing site will be privatized by internet criminals so that other internet users cannot find out information about the site [3]. Therefore, if the domain is private, it will be given a value of 1. But if not, it will be given a value of -1.

6. Site length

   Usually, a site classified as a phishing site has more length than a non-phishing site. Some studies say that site length greatly impacts phishing site detection systems, which is 51% [3]. Therefore, this research also involves the length feature of phishing sites to get accurate results.

7. Contains https

   HTTPS is a protocol used by large sites to maintain the website's security, which is unlikely to be owned by sites detected as phishing sites because phishing sites are usually used to spread viruses, malware, hack, and steal data [3]. If the website has https and the https is accompanied by an SSL (Secure Socket Layer) certificate, it will be given a value of -1. However, if the website has https and the https is not accompanied by an SSL certificate, it will be given a score of 0; otherwise, it will be given a score of 1.

8. Abnormal URL

   Phishing sites usually have URLs that are similar to the original site, but there are some things that are suspicious about abnormal URLs such as misspellings, random characters, or even the addition of some irrelevant words[10].

9. Disable right click

   The right click feature can be used by users to access important features, therefore phishing sites generally do not enable this feature so that users cannot realize or report suspicious activity[10]. These important features include opening links in a new tab or checking link properties.

10. Non-standard port

    There are non-standard HTTP ports that are widely used in phishing sites such as 8080 or 8888, using port numbers that do not comply with this protocol standard can redirect users to unauthorized servers [16].

11. Redirecting '//'

    Phishers sometimes use redirects using the '//' sign in part of the URL, unconsciously users can be directed to a different site[10].

12. Sub domains

    Adding sub-domains to a URL so that it resembles the original site can cause users to believe that they are accessing a legitimate site. For example, phishers may use "paypal-security.com" as a fake subdomain of "paypal.com". The more dots there are in the URL, the stronger the indication that the site is a phishing site.

13. Domain registration length

    Domain registration can be an indication of a phishing site if the domain used on the site is newly registered and is abnormally long or the characters used are random.

14. Favicon

    Favicon is a small icon next to the URL or in the browser tab which is often used by phishers to trick users. the favicon used is a fake favicon that is similar to the site's original favicon.

15. HTTPS domain URL

    Most phishing sites do not use HTTPS which indicates data encryption. Therefore, you should check more carefully whether there is HTTPS or not and whether the HTTPS is real or fake, especially if you want to access logins, credit card details, and so on.

In addition to the 15 features above, there are 15 more features in this study, here are the other features;

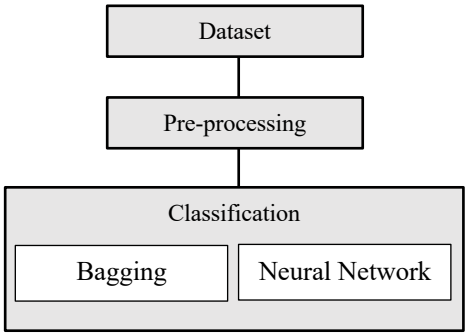
16. Request URL
17. Anchor URL
18. Links in script tags
19. Server form handler
20. Info email
21. Website forwarding
22. Status bar cust
23. Using pop-up window
24. Iframe redirection
25. DNS recording
26. Website traffic
27. Google index

28. Links pointing to page
29. Status reports
30. Class

C. Pre-Processing Data

Data pre-processing is a process/step to make raw data into quality data. The data used amounted to 300 records divided into 200 records (67%) for training data and 100 records (33%) for testing data. The data is stored in .csv format, and each attribute already has a name and not an empty value; besides that, each attribute stores the same data format, so there is no need for data cleaning and data transformation [5].

D. Classification Model



Classification Model Chart

This study's phishing site detection classification model refers to research [3]. In the algorithm testing stage, this research will using confusion matrix. Confusion matrix is a method used to measure the performance of the classification method used. In its application, confusion matrix consists of a table with 4 combinations of values, namely A (Accuracy), P (Precision), R (Recall), and F (F-Measure) A (Accuracy) to evaluate the performance of the classification model created[17]. The following is the formula for the 4 combination values in the confusion matrix:

1. Accuracy

   $Accuracy = (TP + TN) / (TP + FP + FN + TN)$

2. P (Precision)

   $Precision = (TP) / (TP + FP)$

3. R (Recall)

   $Recall = (TP) / (TP + FN)$

4. F (F-Measure)

   $F\ (F\text{-}Measure) = (2 * Recall * Precision) / (Recall + Precision)$

Classification modelling at the trial stage in this research involves two algorithms that have the highest accuracy in detecting phishing sites from previous studies. The two algorithms are Bagging Algorithm and Neural Network Algorithm.

1. Bagging Algorithm

   Bootstrap Aggregating, often abbreviated as Bagging, is an algorithm that will produce results in the form of a training set L (learning) by utilizing sub-datasets (bootstrap) in its calculations [18]. The Bagging algorithm is widely used in regression-related research because it can produce a decision based on existing datasets and process it into a single prediction [3].

2. Neural Networks

   The neural network is an algorithm that works by imitating the function of the human brain combined with simple computational elements (neurons) in a system. This algorithm is often used to solve problems related to classification. Based on [3], the neural network algorithm has a specific type of algorithm, one of which is the multilayer perceptron (MLP). This multilayer perceptron algorithm will work by creating a hidden layer to be studied to adjust its weights and biases to be modelled into patterns in the data. Therefore, this research will also involve one of the specific algorithms of the neural network, namely the multilayer perceptron.

IV. RESULT

1. Results using the Neural Network method.

Testing in this study was carried out using the Neural Network method or more precisely the classic type of Neural Network. Multilayer Perceptron (MLP) is a type of Classic Neural Network used in this research where this method consists of one or more neural layers.

With the MLP method, one of the steps taken is data splitting. In this research, data splitting is assisted by the sklearn.model.selection library, and using the train_test_split function. The dataset used is split with test_size = 0.25 and random state = 42, where test size is a measure of how much test data you want to split. while random state is a parameter used to control when randomizing data and dividing it into new sub datasets. So with this random state, it will produce the same train_test_split when running the code[19].

Neural Network Method				
	Precision	Recall	F1-score	Accuracy
Training Data	0.976	0.957	0.964	0.963
Test Data	0.960	0.935	0.962	0.936

Table of Results with Neural Network Method

2. Results using the Bagging method.

Using the help of DecisionTreeClassifier with random_state = 0 to model the data properly. The use of DecisionTreeClassifier in the bagging method is done because decission tree can handle complex datasets. Then aided by DecisionTreeClassifier, the model is adjusted with the help of BaggingClassifier with base_estimator = tree which is the result of DecisionTreeClassifier model, with n_estimators = 100 and random_state = 0;

Bagging Method				
	Precision	Recall	F1-Score	Accuracy
Training Data	0.995	0.986	0.990	0.989
Test Data	0.986	0.922	0.990	0.944

Table of Results with Bagging Method

V. CONCLUSION

Phishing sites are generally made as similar as possible by phishers to the original site. However, if examined carefully, there are many characteristics that can indicate a fake site such as the features described in the methodology section of this research. Therefore, internet users can be more careful if they find sites that have suspicious indications which can minimize phishing cases in Indonesia.

By using 30 features, the results of this research can be concluded that:

a. The use of classification with Bagging and Neural Network methods can be applied in detecting phishing sites.

b. The test results show that the Bagging method has better performance than the Neural Network method in detecting phishing sites with higher value in Precision = 0.995, Recall = 0.986, F1-Score = 0.990, and Accuracy = 0.989 on training data and Precision = 0.986, Recall = 0.922, F1-Score = 0.990, and Accuracy = 0.944 on testing data.

Training Data Result Comparison Graph

Test Data Result Comparison Graph

VI. REFERENCE

[1] Sri Indah Wijayanti, "Cyber Crime Meningkat Tajam di Masa Pandemi," *ui.ac.id*, Jul. 16, 2021. https://fisip.ui.ac.id/bhakti-cybercrime-menjadi-jenis-kejahatan-yang-mengalami-peningkatan-cukup-tinggi/ (accessed Mar. 31, 2023).

[2] Rahma Yati, "Survei APJII: Penetrasi Internet di Indonesia Capai 78,19 Persen pada 2023," *teknologi.bisnis.com*, Mar. 08, 2023. https://teknologi.bisnis.com/read/20230308/101/1635191/survei-apjii-penetrasi-internet-di-indonesia-capai-7819-persen-pada-2023 (accessed Apr. 08, 2023).

[3] F. E. Purwiantono and A. Tjahyanto, "Model Klasifikasi Untuk Deteksi Situs Phising Di Indonesia," *Surabaya: Institut Teknologi Sepuluh Nopember*, 2017.

[4] Praditya Fauzi Rahman, "Ada 34.622 Kasus Phising di Indonesia Selama 5 Tahun Terakhir.," *detikJatim*, Dec. 27, 2022. https://www.detik.com/jatim/berita/d-6483650/ada-34622-kasus-phising-di-indonesia-selama-5-tahun-terakhir (accessed Apr. 30, 2023).

[5] A. S. Sunge, "Komparasi Machine Learning Memprediksi Phising Dalam Keamanan Website," *Prosiding Sains dan Teknologi*, vol. 1, no. 1, pp. 135–140, 2022.

[6] P. Subarkah and A. N. Ikhsan, "Identifikasi Website Phishing Menggunakan Algoritma Classification And Regression Trees (CART)," *Jurnal Ilmiah Informatika*, vol. 6, no. 2, pp. 127–136, Dec. 2021, doi: 10.35316/jimi.v6i2.1342.

[7] A. Rayan, "Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique," *Comput Intell Neurosci*, vol. 2022, p. 2500772, 2022, doi: 10.1155/2022/2500772.

[8] W. Agwil, H. Fransiska, and N. Hidayati, "Analisis Ketetapan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging CART," *FIBONACCI: Jurnal Pendidikan Matematika dan Matematika*, vol. 6, no. 2, p. 155, Dec. 2020, doi: 10.24853/fbc.6.2.155-166.

[9] T. Setiyorini and R. S. Wahono, "Penerapan Metode Bagging untuk Mengurangi Data Noise pada Neural Network untuk Estimasi Kuat Tekan Beton," *Journal of Intelligent Systems*, vol. 1, no. 1, pp. 37–42, 2015.

[10] S. Widodo, "Klasifikasi Situs Phishing dengan Menggunakan Neural Network dan K-Nearest Neighbor," *INFORMATION MANAGEMENT FOR EDUCATORS AND PROFESSIONALS: Journal of Information Management*, vol. 1, no. 2, pp. 145–154, 2017.

[11] S. Sunaryono, "Penelitian Komparasi Algoritma Klasifikasi dalam Menentukan Website Palsu," *Teknikom*, vol. 1, no. 1, pp. 1–11, 2017.

[12] E. P. Rohmawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural

Network," *Jurnal Ilmiah MATRIK*, vol. 20, no. 1, pp. 21–30, 2018.

[13] E. Saraswati, Y. Umaidah, and A. Voutama, "Penerapan Algoritma Artificial Neural Network untuk Klasifikasi Opini Publik Terhadap Covid-19," *Generation Journal*, vol. 5, no. 2, pp. 109–118, 2021.

[14] A. Hannousse, "Web page phishing detection." Mendeley, 2021. doi: 10.17632/C2GW7FY2J4.3.

[15] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites," *Information & Management*, vol. 51, no. 7, pp. 845–853, 2014, doi: https://doi.org/10.1016/j.im.2014.08.003.

[16] Bradley Mitchell, "Port Number Basics and Services Overview," *Lifeware*, Sep. 17, 2021. https://www.lifewire.com/port-numbers-on-computer-networks-817939 (accessed Jun. 05, 2023).

[17] S. KOM. , M. KOM. DR. MARIA SUSAN ANGGREANY, "Confusion Matrix," *socs.binus.sc.id*. DR. MARIA SUSAN ANGGREANY, S.KOM., M.KOM. (accessed Jun. 04, 2023).

[18] A. Awasthi and N. Goel, "Phishing website prediction using base and ensemble classifier techniques with cross-validation," *Cybersecurity*, vol. 5, no. 1, p. 22, 2022, doi: 10.1186/s42400-022-00126-9.

[19] Michael Galarnyk, "Understanding Train Test Split," *builtin.com*, Jul. 28, 2022. https://builtin.com/data-science/train-test-split (accessed Jun. 04, 2023).