WEBSITE PHISING DETECTION APPLICATION USING SUPPORT VECTOR MACHINE (SVM)

Diki Wahyudi[1], Muhammad Niswar[2], A. Ais Prayogi Alimuddin[3]

Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University, Indonesia
dikiwahyudi953@gmail.com, mniswar@gmail.com, ais.prayogi@gmail.com

Abstract -- Phishing is an act to get someone's important information in the form of usernames, passwords, and other sensitive information by providing fake websites that are similar to the original. Phishing (fishing for important information) is a form of criminal act that intends to obtain confidential information from someone, such as usernames, passwords and credit cards, by impersonating a trusted person or business in an official electronic communication, such as electronic mail or instant messages. Along with the development of the use of electronic media, which is followed by the increase in cyber crime, such as this phishing attack. Therefore, to minimize phishing attacks, a system is needed that can detect these attacks. Machine Learning is one method that can be used to create a system that can detect phishing. The data used in this research is 11055 website data, which is divided into two classes, namely "legitimate" and "phishing". This data is then divided using 10-fold cross validation. While the algorithm used is the Support Vector Machine (SVM) algorithm which is compared with the decision tree and k-nearest neighbor algorithms by optimizing the parameters for each algorithm. From the test results in this study, the best system accuracy was 85.71% using SVM kernel polynomial with values of degree 9 and C 2.5.

***Keywords*: feature extraction, machine learning, parameter optimization, phishing support vector machine (SVM).**

I. INTRODUCTION

Along with the development of technology, especially information technology such as the Internet, which has brought people to a different era, namely an era that feels easier, more practical, simpler, and dynamic. With the development of information technology, people can easily and quickly get what information they want, even in getting that information, there are some parties who do it legally and harm others. The internet has become an important part of people's lives. The internet has affected all parts of people's lives, especially social and financial life. For example, social media and websites are used for communication and business. However, along with the development of information technology, it has brought its own concerns to the community, where there are several irresponsible parties who can harm others.

Phishing is an act to get someone's important information in the form of usernames, passwords, and other sensitive information by providing fake websites that are similar to the original. Phishing (fishing for important information) is a form of criminal act that intends to obtain confidential information from someone, such as usernames, passwords and credit cards, by impersonating a trusted person or business in an official electronic communication, such as electronic mail or instant messages [3]. The term phishing in English comes from the word fishing ('fishing'), in this case it means fishing for financial information and user passwords. With so many cases of fraud being reported, additional methods or protections are urgently needed. These efforts include law-making, user training, and technical measures. Phishing is usually difficult to detect, especially for ordinary people who are not engaged in the technical field. This matter further exacerbated by the increasing use of smartphones which usually do not show the URL of a website as a whole.

Talking about phishing, it will also be associated with social media whose use has increased every year. And if you pay attention to social media, it is the most important medium for a hacker to launch an attack. This is because the use of social media is so large and it is easy to get someone's information through their social media accounts. Phishing attacks like the origin of the word "fishing" which is fishing by giving bait to the victim and waiting for the victim to take the bait.

From several studies [1][4][5][10], in this study the author tries to build a phishing detection application using machine learning methods [6] and the Support Vector Machine (SVM) classification algorithm [9] by extracting features from a URL-based web page using the Python programming language.

II. METHODE

A. Research Stages

The proposed system in this study consists of two main systems, namely the Evaluation System and the Implementation System, in which the Implementation System contains an important process, namely feature extraction, while in the Evaluation System there will be a process of optimizing the parameters of the classification algorithm that will be used. The algorithm that will be used is the Support Vector Machine (SVM) with Decision Tree and K-Nearest Neighbors as a comparison of system performance. The stages that will be carried out in this research can be seen in Figure 1.

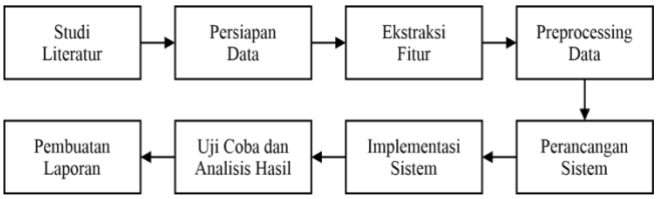

Fig 1. Research Stages

Based on the diagram in Figure 1, the stages in general, research can be described as follows as:

1. Literature study is the initial stage in this research. This stage is carried out to collect research related to the methods used in both feature extraction and classification. At this stage, the collection of the theoretical basis that underlies the research to be carried out is also carried out.
2. Data preparation is a step taken to obtain a dataset that will be used in classifying websites. The dataset used in this study was obtained from Kaggle Datasets.
3. Feature extraction is done to extract the features contained on a website based on the dataset obtained from Kaggle Datasets. The results of this feature extraction will then be used to detect phishing websites.
4. Preprocessing data [8] is done in the form of analyzing the data to be used and selecting data based on the results of feature extraction that has been done previously.
5. The system design in this study was carried out to determine the algorithm to be used in the classification of phishing websites. And in this study, the algorithm used in classifying is the Support Vector Machine (SVM) algorithm with Decision Tree and K-Nearest Neighbors as a comparison of system performance. At this stage, a flowchart related to the workflow of the system is also carried out.
6. Implementation of the system is carried out according to the flowchart that has been made in the previous stage. In this study, the system was built using the python programming language.
7. System testing is carried out to find out how accurate the system is. The trial will be carried out with several different scenarios.
8. The final stage in this research is to write a research report in the form of a thesis as publication material.

B. Research Time and Place

This research was conducted for 9 months starting from the approval of this research proposal in March 2019 until the process of reporting the results of this research in November 2019. This research was conducted at *Ubiquitous Computing and Networking Laboratorium* (UBICON) Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University.

C. Feature Extraction

Feature extraction [2] in this study was conducted to extract the features contained on a website by entering a URL into the system and then the system will access the URL and perform feature extraction on the website. Because to do the extraction, you have to access the website that will be predicted, then this system must be connected to the internet to do this. So this system will automatically be an online system.

The methods used in extracting this feature include checking the URL, which in this case does not need to directly access the website, but simply performs operations on the URL entered. Then there is another method, namely checking the source page, which means the system must be able to access the website online. Other methods include checking the Whois data, DNS Record, Alexa database, PhishTank website and StopBadware.

D. System Design and Implementation

The system made consists of two parts, namely the implementation system and the evaluation system. The implementation system is a system that can be

used directly by the user by entering input in the form of a URL. While the evaluation system is a system created to analyze the performance of the implementation system.

1. Evaluation System

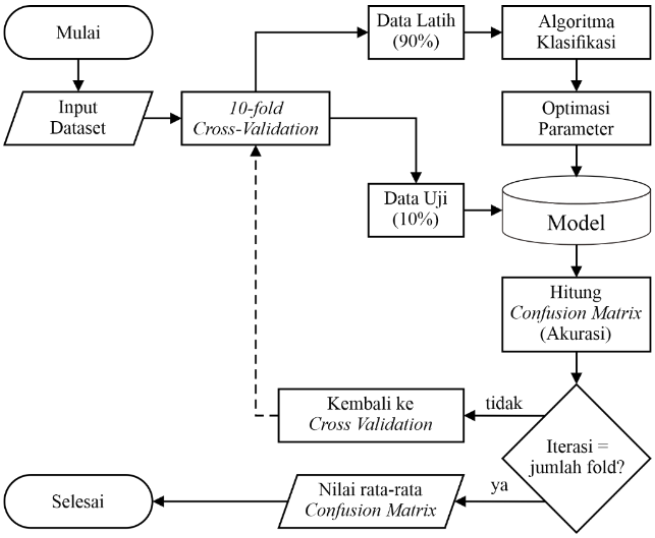

Fig 2. Evaluation System Flowchart

This evaluation system is a system designed as a production system performance evaluation system. Evaluation of this system is done by evaluating the dataset used in the implementation system. The flowchart of the evaluation system can be seen in Figure 2. This evaluation system is a system designed as a production system performance evaluation system. Evaluation of this system is done by evaluating the dataset used in the implementation system. The flowchart of the evaluation system can be seen in Figure 2.

2. Implementation System

This production system is a system that designed to be used and implemented in performing phishing detection. The thing that distinguishes this production system from the evaluation system is that the production system is designed to be used by users with input in the form of a URL. In addition to these differences there is a feature extraction process that is very important and determines system performance. For more details, the system flowchart can be seen in Figure 3.

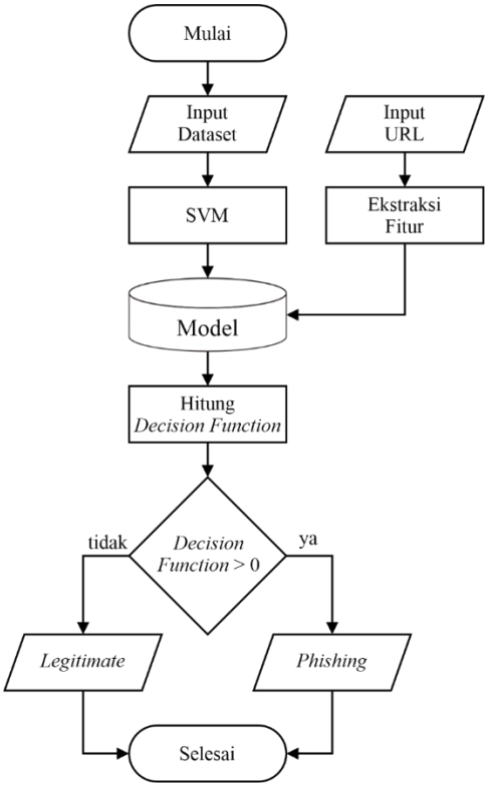

Fig 3. Flowchart of Implementation System

E. System Performance Analysis

The work analysis of the phishing website prediction system is carried out by calculating the accuracy value based on the accuracy of the predictions made by the system. The calculation of the accuracy value is carried out using the help of a confusion matrix as shown in Figure 4. below.

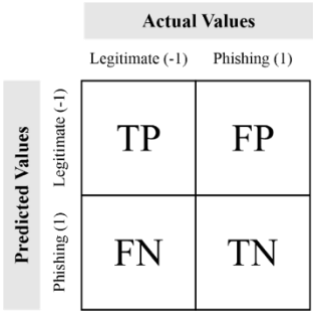

Fig 4. *Confusion matrix*

The confusion matrix consists of four parts, that is:

1. True Positive: the number of correct predictions from positive data.
2. False Positive: the number of false predictions from positive data.
3. False Negative: the number of false predictions from negative data.

4. True Negative: the number of correct predictions from negative data.

To calculate the value of system accuracy, it can be done with the following equation (1).

$$Accuration = \left(\frac{TP + TN}{TP + FP + FN + TN}\right) \times 100\% \quad (1)$$

III. Result And Discussion

In this section, the results of the performance of a phishing detection system using the Support Vector Machine (SVM) algorithm are presented. In this research, a confusion matrix is used to assist in calculating system accuracy. The confusion matrix table for each experiment with the parameter optimization method can be seen in the appendix. The amount of data used is 11055 data, which is divided by the 10-fold cross-validation method.

In this research, experiments are carried out on three algorithms with parameter optimization for each algorithm. The three algorithms are:

a. Support Vector Machine (SVM)

1) Linear Kernels

a) The value of parameter C is 1.0, 1.5, 2.5, and 5.0.

2) Polynomial Kernels

a) The value of parameter C is 1.0, 1.5, 2.5, and 5.0.

b) The value of the degree parameter is 1, 3, 5, 7, and 9.

3) RBF Kernels

a) The value of parameter C is 1.0, 1.5, 2.5, and 5.0.

b) The gamma parameter values are 0.1, 0.5, 1.0, 1.5, and 2.0.

b. Decision Tree

1) Parameter values for *criterion* are *gini* and *entropy*.

2) *max_depth* parameter values are 1, 5, 10, 15, and 20.

c. K-Nearest Neighbors

1) Parameter values for *weights* are *uniform* dan *distance*.

2) The *n_neighbors* parameter values are 1, 5, 10, 15, and 20.

The results of the system performance for each experiment are shown as follows.

a. Support Vector Machine (SVM)

1) Linear Kernels

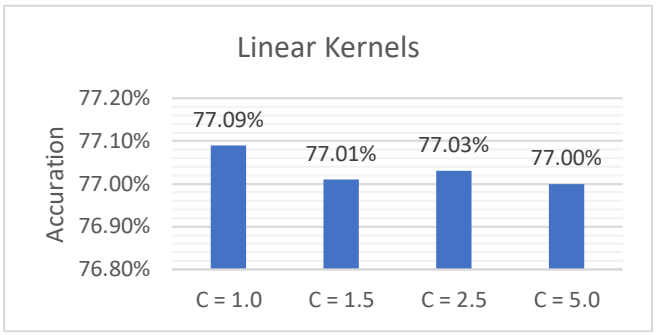

Fig 5. Comparison Graph of Accuracy based on *C* value

2) Polynomial Kernels

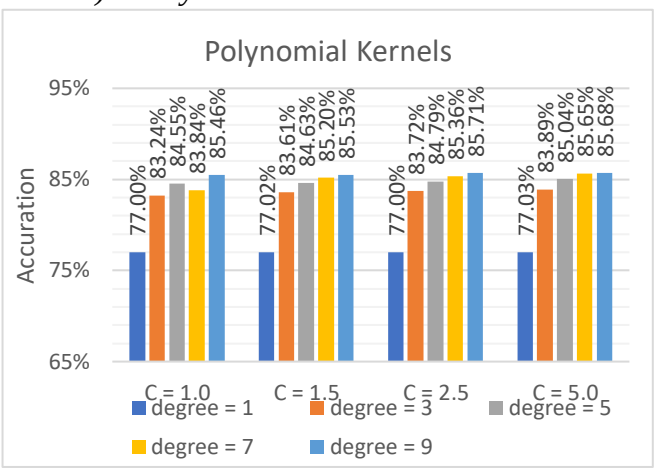

Fig 6. Comparison Graph of Accuracy based on Degree and C values

3) RBF Kernels

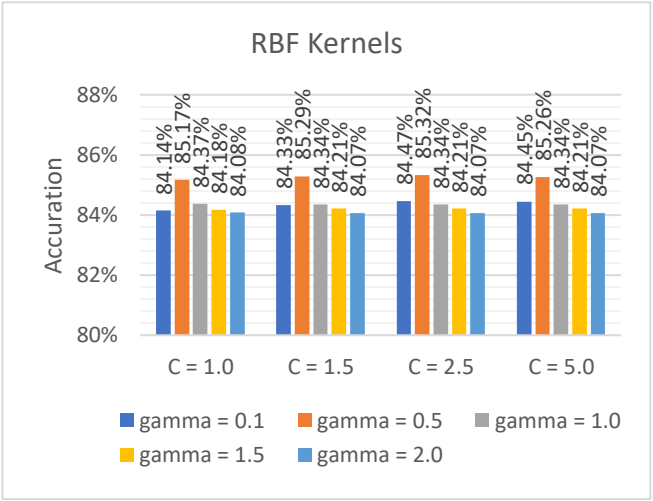

Fig 7. Comparison Graph of Accuracy based on Gamma and C values

b. *Decision Tree*

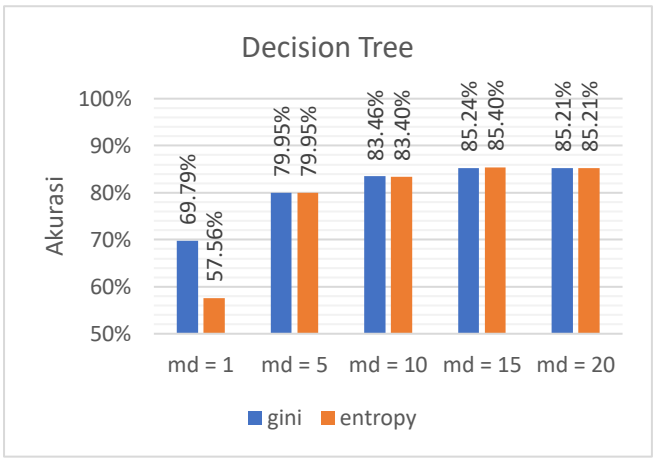

Fig 8. Accuracy Comparison Graph based on the value of *max_depth* and *criterion*

c. *K-Nearest Neighbors*

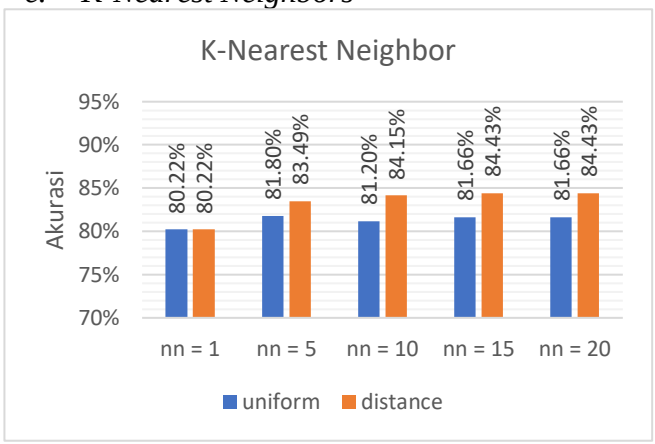

Fig 9. Comparison Graph of Accuracy based on the Value of *n_neighbors* and *weights*

Based on the table that has been presented, it can be seen that the best accuracy of 85.71% is obtained in the SVM kernel polynomial algorithm with a value of *degree* 9 and a value of *C* 2.5.

1. *Implementation*

In this section, we present the results of implementing a phishing website detection system using the support vector machine (SVM) algorithm. In its implementation, the system has been able to receive input from the user in the form of a URL that will be detected. After the URL is entered by the user, the system will perform a feature extraction process against the URL. The results of the feature extraction will be saved in the form of a comma separated value (CSV). An example of the results of feature extraction from the system can be seen in Table 1.

Fig 10. Implementation Results of Phishing Website Detection Program

TABLE 1

Feature Extraction Results

Having IP Address	URL Length	Shortining Service	Having At Symbol	Double Slash Redirecting	Prefix Suffix	Having Sub Domain	Domain Registration Length	HTTPS Token	Submitting to Email	Right Click	Iframe	Age of Domain	DNS Record	Web Traffic	Statistical Report
-1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1

From the results of feature extraction [7], classification will be carried out using the SVM model using training data from the Kaggle dataset. And with this model, the system will be able to classify the URLs that were previously entered, as in the example in Figure 10.

1. *Evaluation*
   a. *Support Vector Machine (SVM)*
      - The effect of the value of C on the results of algorithm performance

Parameter C is the amount of penalty given to the classification error. In the linear kernel, the greatest accuracy is obtained at the value of C 1.0. By looking at the data in Figure 4.1. it can be said that for a linear kernel with a dataset in this study it is not good to use a large C value. Almost the same as the linear kernel, the highest accuracy results obtained from the polynomial kernel and RBF is at the value of C 2.5 with an accuracy value of 85.71% and 85.32%, respectively.

From the results of the experiment with the optimization of the C parameter on the three kernels above, it can be concluded that the phishing website dataset used in this study is not good at using a C value that is too large.

- Effect of degree value on algorithm performance results

The degree parameter is a parameter that can only be optimized in the polynomial kernel. From the test results, the best accuracy is obtained at the value of degree 9 with a value of C 2.5 as well as being the best accuracy among other algorithms. And by paying attention to the accuracy in each degree value, it is seen that there is a significant increase in accuracy. The one with the lowest degree value obtained 77.00% can increase to 85.00%. So from these results it can be concluded that by optimizing the degree parameter at a certain value can increase the accuracy of the system.

- Effect of gamma value on algorithm performance results

The gamma parameter is an optimization parameter in the SVM kernel RBF algorithm using the values, 0.1, 0.5, 1.0, 1.5, 2.0. From the test results, the best accuracy is obtained at a gamma value of 0.5 with a C value of 2.5. By paying attention to the accuracy of each gamma and C values, it can be concluded that the optimal gamma parameter is at a value of 0.5.

b. Decision Tree

In an experiment using a decision tree, two criteria are used in splitting the data, namely Gini and Entropy. And the max_depth values are 1, 5, 10, 15, and 20.

Based on Figure 8. By using Gini, the highest accuracy is obtained at max_depth 15, which is 85.24%. while for entropy the highest accuracy is obtained at max_depth 15, which is 85.40%. From these results, it can be concluded that for the dataset in this study, it is more suitable to use the entropy criterion. And not always with a large max_depth will get better accuracy depending on the number of features used.

c. K-Nearest Neighbors

In the experiments conducted on the KNN algorithm, the values of the weights are used, namely, unform and distance. While n_neighbors are 1, 5, 10, 15, and 20. Based on Figure 4.5., using the uniform value the highest accuracy is obtained at the value - n_neighbors 5, while using the distance value the highest accuracy is obtained at the value of n_neighbors 20. These results indicate that with uniform it is not always the greater the value of n_neighbors, the greater the accuracy obtained. Unlike the case with distance, the greater the value of n_neighbors, the better the accuracy obtained. In uniform, the weight given to each n_neighbors is the same. So, every -n_neighbors has the same effect on the point. With every n_neighbors having the same effect, if - n_neighbors is chosen too large it can cause the classification to be less good, this is because there are many n_neighbors that affect the point. In contrast to the greater distance n_neighbors the better the accuracy. This is because the weight given to each n_neighbors is not the same but is determined based on the distance from the point. So even if the value of n_neighbors is large, only the one closest to the point will have a big effect on the point.

2. *Implementasi*

In the implementation system, there are several things that distinguish it from the evaluation system, namely in the implementation system there has been contacted with the user who will enter the URL for detection of phishing websites. And in the evaluation system, there is a process that plays an important role, namely feature extraction. With feature extraction, the system will be able to extract the required features according to the dataset used from the Kaggle dataset. The results of this feature extraction will later be used as test data as well as data that will be classified by the system using the Machine Learning Support Vector Machine (SVM) algorithm.

IV. CONCLUSION

From the results of the analysis that has been carried out in testing the phishing website detection system using machine learning methods, it can be concluded that:

1. The phishing website detection system is made using the *Machine Learning Support Machine Learning (SVM)* algorithm as an algorithm to detect phishing websites. Before the URL that is input to the system is processed using the Support Vector Machine (SVM) algorithm, a feature extraction process will first be carried out on the URL that has been entered. The method used in performing feature extraction is different for each feature to be extracted. The methods used include URL Obfuscation extraction, Whois data extraction, DNS Record extraction, data extraction from Alexa database, and data extraction on PhishTank and StopBadware websites.
2. The results of the test using 10-fild cross-validation with machine learning algorithms *Support Vector Machine (SVM)*, *Decision Tree*, and *K-Nearest Neighbor* as well as the parameter optimization process for each algorithm, the best accuracy of each algorithm is SVM kernel linear 77.09% with parameter values of C 1.0, kernel polynomial 85.71% with parameter values of degree 9 and C 2.5, kernel RBF 85.32% with parameter values of gamma 0.5 and C 2.5, decision tree 85.40% with parameter values of criterion entropy and max_depth 15, and KNN 84 ,43% with parameter values of weights distance and n_neighbors 15 and 20. By comparing the best accuracy of each algorithm, it can be concluded that the best accuracy is obtained in the SVM kernel polynomial algorithm with parameter values of degree 9 and C 2.5, namely 85.71%.

V. ACKNOWLEDGMENT

We thank to all people that contribute to this research. Special to Hasanuddin University.

VI. REFERENCE

[1] Natan, Oskar dkk. 2019. “Grid SVM: Aplikasi Machine Learning dalam Pengolahan Data Akuakultur”. Jurnal Rekayasa Elektrika Volume 15 (hlm 7-17). Surabaya: Politeknik Elektronika Negeri Surabaya.

[2] Karima, Inna Sabily. 2014. Optimasi Parameter Pada Support Vector Machine untuk Klasifikasi Fragmen Metagenome Menggunakan Algoritme Genetika. Bogor: Institut Pertanian Bogor.

[3] Halim Z. 2017. “Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM)”. Information System for Educators and Professionals. 2 (1): 71 – 82

[4] Diani, Rima dkk. 2017. “Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker”. IndonesiaJournal on Computing Volume 2 (hlm. 109-118).

[5] Preethi, V dan Velmayil, G. 2016. “Automated Phishing Website Detection Using URL Features and Machine Learning Technique”. International Journal of Engineering and Technique Volume 5.

[6] Putra, Jan Wira Gotama. 2019. Pengenalan Konsep Pembelajaran Mesin dan Deep Learning.Tokyo: Tokyo Institute of Technology.

[7] Mohammad, Rami M dkk. “Phishing Websites Features”.

[8] Purwiantono, Febry Eka, dan Aris, Tjahyanto. 2017. “Model Klasifikasi untuk Deteksi Situs Phising di Indonesia”. Surabaya: Institut Teknologi Sepuluh Nopember.

[9] Nugroho, Anto Satriyo. 2007. “Pengantar Support Vector Machine”. Jurnal Teknologi.

[10] Erton, M Wiby. 2018. “Support Vector Machine”. Jombang: Universitas KH. A. Wahab Hasbullah.