# Phishing Website Detection using Neural Network and PCA based on Feature Selection

**Deyanara Tuapattinaya, Antoni Wibowo**

***Abstract*: Phishing is a criminal activity that tries to steal user account password or other confidential information by tricking user into believing they are on the actual website. In order to phishing, they must get user to go from an email to a website. User can also land on phishing site by mistyping a URL (web address). However, the numbers of phishing attacks have been growing and need the protection technique. Neural network and Principal Component Analysis (PCA) can be combined to detect phishing website. This study uses back-propagation algorithm based on neural network method and PCA based on feature selection to reduce large attributes into small attributes. Neural network without using PCA will be compared with neural network using PCA. The result shows that neural network using PCA has better accuracy in 55.67% and neural network without using PCA only reaches 54.43% accuracy. However neural network without using PCA has faster computing time than neural network using PCA. This study can be used as a phishing protection technique.***

***Keywords* : back-propagation, neural network, phishing website, principal component analysis (PCA)***

## I. INTRODUCTION

Phishing is a fraudulent attack by tricking the user to land on illegitimate site to steal their personal information [1]. In order for social engineering to successfully "phish" user personal data, they must get users to go from an email to a website [2]. Usually phishing email come from an outstanding organization and asks for user personal data such as username, password and credit card detail. Most phishing email are done by telling users to follow a link that send user to a website where their personal data is requested [3]. Legitimate organization would never request data information from user via email. Phishing activity began to be monitored since 2014 and the highest result achieved in 2016. The phishing attacks reached 1.220.523 in 2016 [4].

Therefore, phishing attacks is needs to be prevented, it requires research on detection phishing website to get a good classification in order to produce good performance. In this problem, a feature extraction is needed to reduce and compress the data in order to eliminate noise and to produce a good classification.

This study proposed two models to be compared. The first model is neural network without using PCA, and the second model is neural network using PCA. As of the large number of features in this study, thus PCA is needed to extract the features. The first and the second model will be compared based on accuracy and computational time. This study uses back-propagation algorithm based on neural network method and PCA based on feature selection to reduce large attributes into small attributes. From this comparison it is expected that this study can provide an overview of the most efficient and accurate method for predicting phishing website based on accuracy and time.

## II. RELATED WORK

Various studies have been carried out in detecting phishing website. The result of the study literature can be seen in the table below, which contains author, output, methods and performance measure.

**Table- I: Table of study literature**

| Author | Output | Method | Performance Measure |
|---|---|---|---|
| [5] | Neural network reduces the error & gives better classification | Neural Network | Comparison against data mining classification (error rate) |
| [6] | Decision Tree outperforms another algorithm | Neural Network, Decision Tree, KNN, NB, SVM, K-Means | Comparison against machine learning techniques |
| [7] | C4.5 outperforms another algorithm | C4.5, Ripper, Prism, CBA | Error Rate |
| [8] | Wrapper feature selection gives better performance | Neural network (BPNN), RBFN, NB, SVM, C4.5, RF, KNN | Accuracy with feature selection and without feature selection |
| [9] | Improved Accuracy | Fuzzy & Neural network | RMSE & Accuracy |
| [10] | MLP outperforms all algorithm | MLP, Random forest, Decision tree, J48, REP tree, Decision stump, Hoeffding tree | Accuracy |

Based on Table I, neural network is the most widely used method for detecting phishing website. Neural network has been proven to reduce errors and give better classification. Therefore, based on previous research, this study uses a neural network to produce great performance.

**Revised Manuscript Received on February 01, 2020.**
***** Correspondence Author
**Deyanara Tuapattinaya***, Computer Science Department, Bina Nusantara University, Jakarta, Indonesia. Email: deyanara@binus.ac.id
**Antoni Wibowo**, Computer Science Department, Bina Nusantara University, Jakarta, Indonesia. Email: anwibowo@binus.edu

## III. THEORY AND METHOD

### A. Artificial Neural Network (ANN)

Neural network is a model that is made based on neurons that work in the human brain. Each neuron in human brain is interrelated and provides information. Each neuron accepts an input and performs a dot operation with weight to be summed. The result of this model will be used as a criterion of the activation function to produce output neuron. This research uses backpropagation which is a part of neural network. Back-propagation is the most frequently used algorithm on neural network methods [11]. Back-propagation consists of input layer, hidden layer and output layer and works to minimize errors in output.

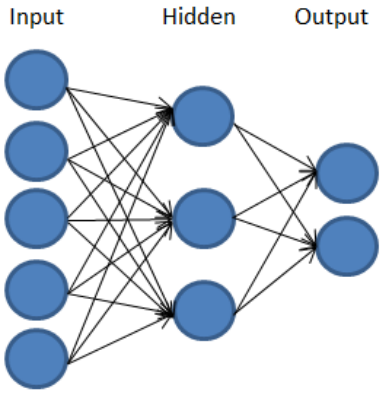


**Fig. 1.Neural network. (*back-propagation*) structure**

### B. Principal Component Analysis

Principal Component Analysis (PCA) is an algorithm used to reduce the dimensionality of data by converting a large number of attributes into a small attributes [12]. PCA is one of the algorithms based on feature selection, also known as feature extraction. Feature selection works by selecting the right attributes and removing non-essential attributes to get good analysis results [13].

## IV. PROPOSED METHOD

This study proposed two models to be compared. The first model is neural network without using PCA, and the second model is neural network using PCA. As of the large number of features in this study, thus PCA is needed to extract the features. The first and the second model will be compared based on accuracy and computational time. This study uses back-propagation algorithm based on neural network method and PCA based on feature selection to reduce large attributes into small attributes. To measure this accuracy, this paper uses dataset from UCI Machine Learning Repository. The dataset has 11055 training set and 31 attributes. Total phishing website is 4898 and total legitimate website is 6157. This dataset has been proven effective in predicting phishing website.

In this paper, all experiments were carried out with python software. Python is a programming language used to developing numeric applications and complex scientific. Python is designed with features to facilitate data analysis and visualization. This paper compares neural network model without using PCA and neural network using PCA. The dataset is divided into 3 parts, 80% training dataset, 10% validation dataset, and 10% testing dataset. The first model is neural network without using PCA. It has 31 attributes as can be seen in Table II. The second model is neural network using PCA, the attributes is selected into 18 out of 31. As you can see in Table III, total number of inputs is 18.

Property from the first model can be seen in Table II, which contains 31 input layers, 1 hidden layer, 0.3 learning rate, 100 iterations, 80% training dataset, 10% validation dataset, and 10% testing dataset.

**Table- II: Parameters of neural network without using PCA**

| Parameters | Total |
|---|---|
| Number of inputs | 31 |
| Hidden layer | 2 |
| Learning rate | 0.3 |
| Iteration | 100 |
| Training | 80% |
| Validation | 10% |
| Testing | 10% |

Property from the second model can be seen in Table III, which contains 18 input layers, 1 hidden layer, 0.3 learning rate, 100 iterations, 80% training dataset, 10% validation dataset, and 10% testing dataset. In this model, the attributes have been extracted from 31 to 18. Based on previous research in [11], 18 features are proven to produce good accuracy in detecting phishing. Therefore, the second model which is neural network with PCA uses 18 features.

**Table- III: Parameters of neural network using PCA**

| Parameters | Total |
|---|---|
| Number of inputs | 18 |
| Hidden layer | 2 |
| Learning rate | 0.3 |
| Iteration | 100 |
| Training | 80% |
| Validation | 10% |
| Testing | 10% |

## V. RESULT AND DISCUSSION

The result can be seen in Table IV, from the first model, the best accuracy is achieved at 54.43%, and the lowest accuracy is 42.04%. This model requires ±15 seconds computing time per one run. From the second model, the best accuracy is achieved at 55.67% and the lowest accuracy is 52.88%. This model requires ±4 hours computing time per one run. Neural network using PCA can give better performance than neural network without using PCA. Neural network without using PCA have faster computing time in ±15 seconds and neural network using PCA needed ±4 hours in computing time.

**Table- IV: Result of testing neural network without using PCA and neural network using PCA**

| Number of runs | Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | ANN | | | ANN-PCA | | |
| | Training | Validation | Testing | Training | Validation | Testing |
| 1 | 44.27% | 46.24% | 42.67% | 54.24% | 54.22% | 54.26% |
| 2 | 46.89% | 46.79% | 49% | 52.86% | 52.58% | 52.88% |
| 3 | 44.57% | 44.43% | 42.04% | 55.02% | 55.01% | 54.88% |
| 4 | 54% | 54.3% | 54.43% | 55.7% | 55.69% | 55.67% |
| Average | 47.43% | 47.94% | 47.04% | 54.46% | 54.38% | 54.42% |
| Min | 44.27% | 44.43% | 42.04% | 52.86% | 52.58% | 52.88% |
| Max | 54% | 54.3% | 54.43% | 55.70% | 55.69% | 55.67% |

From the two models that have been studied, neural network using PCA which is consisting 1 hidden layer give better performance, it can be seen from the value of accuracy 55.67%. Compared to neural network without using PCA which is consisting 1 hidden layer has an accuracy value 54.43%. Based on this experiment, the best classification obtained using neural network using PCA. This shows that the opportunity for phishing website and legitimate website have 44.33% classification, it means that from 11055 there are 4897 incorrectly classified data.

## VI. CONCLUSION

The results show that neural network using PCA can give better performance than neural network without using PCA. It can be seen on Table IV, neural network using PCA produce the best accuracy in 55.67%, while neural network without using PCA only reaches 52.85% accuracy. However, neural network without using PCA have faster computing time in ±15 seconds and neural network using PCA needed ±4 hours in computing time. The end result is ANN-PCA has better accuracy but slow in computing time, and ANN without PCA has a lower accuracy than ANN-PCA but faster in computing time.

## REFERENCES

1. P. Syiemlieh, G. M. Khongsit, U. M. Sharma, and B. Sharma, "Phishing-An Analysis on the Types , Causes , Preventive Measuresand Case Studies in the Current Situation," pp. 1–8.
2. R. J. Sassi et al., "Artificial Neural Network for Websites Classification with Phishing Characteristics," Soc. Netw., vol. 07, no. 02, pp. 97–109, 2018.
3. R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Comput. Appl., vol. 25, no. 2, pp. 443–458, 2014.
4. P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," Wirel. Commun. Mob. Comput., vol. 2018, 2018.
5. A. Bergholz, J. Chang, G. Paass, F. Reichartz, and S. Strobel, "Improved Phishing Detection using Model-Based Features.," Ceas, no. September, 2008.
6. N. Zhang and Y. Yuan, "Phishing detection using neural network," Dep. Comput. Sci. Dep. Stat. Stanford Univ. Web, vol. 29, no. Department of Computer Science, Department of Statistics, Stanford University, 2013.
7. L. McCluskey, F. Thabtah, and R. M. Mohammad, "Intelligent rule-based phishing websites classification," IET Inf. Secur., vol. 8, no. 3, pp. 153–160, 2014.
8. W. Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, pp. 72–78, 2017.
9. ] L. Anh and T. Nguyen, "Phishing Identification Using a Novel Non-Rule Neuro-Fuzzy Model," vol. 14, no. 4, 2016.
10. A. Hodžić and J. Kevrić, "Comparison of Machine Learning Techniques," ICESoS 2016 - Proc. B., pp. 249–256, 2016.
11. A. F. Thabtah, T. L. McCluskey, and R. M. Mohammad, "Predicting Phishing Websites using Neural Network trained with Back-Propagation," Proc. 2013 World Congr. Comput. Sci. Comput. Eng. Appl. Comput. WORLDCOMP 2013 ., no. January, pp. 682–686, 2013.
12. M. Zareapoor and S. K. R, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection," Int. J. Inf. Eng. Electron. Bus., vol. 7, no. 2, pp. 60–65, 2015.
13. P. Singh, N. Jain, and A. Maini, "Investigating the effect of feature selection and dimensionality reduction on phishing website classification problem," Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015, no. June, pp. 388–393, 2016.

## AUTHORS PROFILE

**Deyanara Tuapattinaya** was born in Ambon, Indonesia on January 14, 1998. She has received her first degree of Computer Science from Bina Nusantara University on 2019. She is currently pursuing a master's degree in computer science from Bina Nusantara University, 2019.

**Antoni Wibowo** has received his first degree of Applied Mathematics in 1995 and a master's degree of Computer Science in 2000. In 2003, He was awarded a Japanese Government Scholarship (Monbukagakusho) to attend Master and PhD programs at Systems and Information Engineering in University of Tsukuba-Japan. He completed the second master's degree in 2006 and PhD degree in 2009, respectively. His PhD research focused on machine learning, operations research, multivariate statistical analysis and mathematical programming, especially in developing nonlinear robust regressions using statistical learning theory. He has worked from 1997 to 2010 as a researcher in the Agency for the Assessment and Application of Technology – Indonesia. From April 2010 – September 2014, he worked as a senior lecturer in the Department of Computer Science - Faculty of Computing, and a researcher in the Operation Business Intelligence (OBI) Research Group, Universiti Teknologi Malaysia (UTM) – Malaysia. From October 2014 – October 2016, he was an Associate Professor at Department of Decision Sciences, School of Quantitative Sciences in Universiti Utara Malaysia (UUM). Dr. Eng. Wibowo is currently working at Binus Graduate Program (master's in computer science) in Bina Nusantara University-Indonesia as a Specialist Lecturer and continues his research activities in machine learning, optimization, operations research, multivariate data analysis, data mining, computational intelligence and artificial intelligence