Phishing Attack Detection Using Deep Learning

Sabah M. Alzahrani

Department of Computer Science, College of Computers and Information Technology, Taif University,
Taif P.O. Box 11099, Taif, 21944, Saudi Arabia

Summary
This paper proposes a technique for detecting a significant threat that attempts to get sensitive and confidential information such as usernames, passwords, credit card information, and more to target an individual or organization. By definition, a phishing attack happens when malicious people pose as trusted entities to fraudulently obtain user data. Phishing is classified as a type of social engineering attack. For a phishing attack to happen, a victim must be convinced to open an email or a direct message [1]. The email or direct message will contain a link that the victim will be required to click on. The aim of the attack is usually to install malicious software or to freeze a system. In other instances, the attackers will threaten to reveal sensitive information obtained from the victim. Phishing attacks can have devastating effects on the victim. Sensitive and confidential information can find its way into the hands of malicious people. Another devastating effect of phishing attacks is identity theft [1]. Attackers may impersonate the victim to make unauthorized purchases. Victims also complain of loss of funds when attackers access their credit card information. The proposed method has two major subsystems: (1) Data collection: different websites have been collected as a big data corresponding to normal and phishing dataset, and (2) distributed detection system: different artificial algorithms are used: a neural network algorithm and machine learning. The Amazon cloud was used for running the cluster with different cores of machines. The experiment results of the proposed system achieved very good accuracy and detection rate as well.

Key words:
Artificial Intelligence; Cybersecurity; Detection; Deep Learning; Phishing; Social Engineering; AWS; Spark; Distributed.

1. Introduction and Background

Phishing is a significant online security concern. As per the most recent Google Safe Browsing report, Google search boycotts more than 50,000 malware locales and more than 90,000 phishing destinations month to month. The APWG (Anti–Phishing Working Group) detailed that the number of phishing assaults in 2020 was 65% more than in 2019. Over the most recent 12 years, the quantity of phishing assaults each month has expanded by 5753% [2]. The harm brought about by phishing assaults is, however, broad as it seems to be different. The following are the datasets used for phishing attacks detection and how they differ.

a) URL Features

URL or Lexical highlights are gotten dependent on the properties of the URL of the site. The arrangement of words in the area parcel, part segment, and TLD (Top Level Domain) of the URL and presence of certain unique characters and their positions are huge URL includes that add to discovery of phishing locales. The greater part of the phishing assaults is through email tricks, and thus assailants need a manner by which they can bait clients by making a phishing site that looks precisely like a current considerate site. Phishers regularly use confusion strategies to attract the client to tap on the phishing URL. By muddling the hostname with an IP address or including the objective brand's area name in the URL way with or without spelling mistakes or utilizing longer URLs to install kindhearted glancing tokens in the sub-space or way of the URL, phishers attempt to draw clients into getting to the phishing site, for instance, www.ebay.example.com. For example, the use of uncommon characters, for example, '@,' which cause the program to disregard the string on the left of '@' and treat the string on the right as a real URL, is not many alternate approaches to deceive clients, http://ebay.com/personal_info@www.xyz.com. Phishing pages, by and large, have various sidetracks to divert from the underlying URL to the last site, which the assailant facilitates in any compromised machine [3]. Aggressors likewise taint considerate destinations with a vigorously muddled noxious JavaScript code that implants an iframe with the assailant's malignant space URL and afterward tosses an HTTP 302 redirection to stack the phishing site abuse area.

b) Page content and JavaScript features

Numerous highlights separated from the site page content, such as the presence of login structures, the presence of secret word fields, and the presence of strange prearranging content, can help recognize a phishing site. Most phishing site pages contain structures with input fields to get client installment card subtleties or secret words. Aside from that, few methods like secret components, popups, and prompts to enter touchy data are utilized to bait clients into entering passwords and classified data [4]. JavaScript anomalies, the presence of shellcode in the page, and dubious Active X controls can likewise signify a phishing page – where an assailant abuses any weakness in the page to infuse scripts/code that can download touchy client data from the page to the aggressor's worker. Many exploration works extricate these highlights to identify phishing sites. Extricate the whole website page alongside HTML content and pictures and store it as a profile. At whatever point the client stacks another site, it is checked

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.12.31

against all put-away profiles. If a nearby match is found, there are high possibilities that the stacked site is a phishing site [5]. Profiler framework removes numerous highlights from JavaScript, DOM, and Active X controls of the page alongside URL and Host-based highlights to prepare the classifier to foresee phishing sites. Some examination work checks for the presence of structures and secret word field in structures, source URL match with demand URL and connections on-page as these highlights are asserted as great pointers of phishing sites. This makes the general framework a lightweight activity with less computational overhead when contrasted with frameworks like Profiler and the one planned, which remove numerous pages and JavaScript highlights to recognize phishing [6].

c) Hosting Domain features

Highlights from WHOIS and DNS records of the site explain how and where phishing sites are facilitated. Recorders and vaults give WHOIS administrations to the area names that they support. WHOIS records give data about the site's registrant, enlistment creation, lapse dates, and scarcely any different subtleties [7]. DNS records are planning documents that tell the DNS (space name framework) worker which IP address every area is related with. When somebody visits a site, a solicitation is shipped off the DNS worker and afterward sent to the relating IP address where the site is facilitated. WHOIS highlights like the time of area, the life expectancy of a space, recorder subtleties and barely any others and DNS record highlights, for example, self-ruling framework number, name worker address, and area, time to live worth of DNS record and so on are generally utilized in writing reviewed. Phishers typically target compromised facilitating spaces to dispatch their assaults so that acquiring client data through the phishing site is simple. Phishing destinations are made for a limited ability to focus receive the most extreme in return in a couple of days before the site is distinguished and hindered [8]. Phishing efforts are brief as phishers can't bear to pay for a facilitating space for a significant stretch. Phishers likewise utilize free web facilitating administrations and space tasting to have their sites for a limited ability to focus time.

d) Security Features

SSL endorsement for the site, presence of public key declaration, and if the site treat is strange is not many protections related highlights. While most phishing assaults run over HTTP, a huge number of the sudden spike in demand for destinations for which SSL endorsements have been given as declaration specialists don't investigate who gets their SSL testaments. Bonafide testament proprietors now and again accidentally give offices to phishing because an assailant has undermined their site [9]. In specific cases where phishers have their sites, it could be hard to acquire a phony SSL declaration as certain endorsements require approval by a testament authority.

e) Site Popularity features

Google rank, Alexa traffic rank, and the number of connections highlighting the site are acceptable markers of how mainstream the site is. Alexa traffic rank is determined by Alexa.com and gives three months of amassed information dependent on the number of connections inside the webpage saw my clients and the number of clients saw the site [10]. Phishing site pages are fleeting and consequently either have a low page rank or their page rank in the Alexa information base. Their social standing score, which is the quantity of preferences/shares on Facebook and Twitter, would be less. We utilize this element class alongside URL and facilitate area highlights for phishing site locations. These highlights are simpler to gather and make the phishing discovery framework a lightweight activity.

f) Network Features

Malevolent sites utilize rich web assets that can cause numerous HTTP demands shipped off the web worker, including different redirections, iframes, and external connections to other space names. Consequently, network layer data, for example, the number of bundles sent and set up association and number of ports opened on the web worker, can help identify phishing sites [11].

2. Proposed System

Data has been collected as unstructured data of URLs from Phishtank website, and Alexa website for the normal Data. PhishTank is an online platform dedicated to fighting phishing. On this platform, users can submit, verify, and share phishing data. Use of the PhishTank platform is free to all users. According to [12], the PhishTank platform does not protect against phishing attacks. The PhishTank platform is run and maintained by Cisco Talos Intelligence Group. The data is a total of 10,000 legitimate websites and 10,000 phishing websites.

After the data is collected then it has been stored on Amazon public cloud storage S3. The data then was on the second subsystem. Different deep learning algorithms were applied in a distributed way. According to [13], deep learning is a subset of machine learning that draws inspiration from the human brain. Deep learning is different from traditional machine learning. The concept of deep learning involves algorithms attempting to draw conclusions as a human being would do by analyzing data. Deep learning networks are able to study and analyze data by building computational models [13]. Models built by deep learning networks are made up of different processing layers. The concept of deep learning has found different applications in different fields. Firstly, the concept of deep learning is applied in agriculture as a modern farming technique. Deep learning is used in agriculture to differentiate crops from weeds [13]. Secondly, deep learning is applied in medicine and healthcare settings. Deep learning algorithms are increasingly used in medical imaging and diagnosis [13]. Lastly, the concept of deep learning is also being widely applied in the entertainment industry. For example, Netflix uses deep learning algorithms to recommend films to viewers based on their browsing patterns.

This experiment is running on Amazon EMR with different machines cores and using distributed deep learning. The algorithms are designed with Keras and PySpark through the Elephas platform. Amazon EMR is a cloud-based big data program that allows users to process large amounts of data within a short time. Due to its enhanced processing speed, Amazon EMR offers a cost-effective method of

processing data [14]. Amazon EMR works with the help of open-source tools such as Apache Flink and Apache HBase. Amazon EMR is a useful tool for data analysis for firms of all sizes. Amazon EMR works in tandem with other platforms such as Amazon EC2 and Amazon S3.
Elephas, developed with Keras, enables users to write their deep learning models as standard Spark programs. The main advantage of Elephas is that it is built on the highly scalable Apache Spark platform. The scalable platform allows Elephas to scale out into many small servers. The development of Elephas has made deep learning more accessible to the big data community. Deep learning models such as object detection have greatly benefited from Elephas. Figure 1 shows the diagram of the proposed system.

3. Methodology

Three different deep learning algorithms were used. CNN Convolutional Neural Network, logistic regression and linear regression.

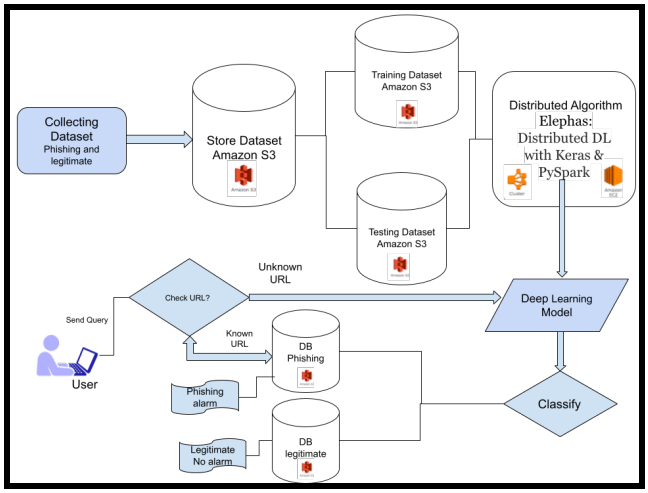

Fig. 1: Proposed Method

3.1. Deep Learning Algorithms

Deep learning uses artificial neural networks to conduct complex computations on massive quantities of data. Learning can take place in a supervised, semi-supervised, or unsupervised environment [15].
In the three layers of the neural networks, data gives information to each node in the form of inputs. The node multiplies the inputs by random weights, computes them, and then adds a bias. Then, nonlinear functions, also known as activation functions, identify which neuron should be activated. Figure 2. Shows a simple diagram of a neural network.

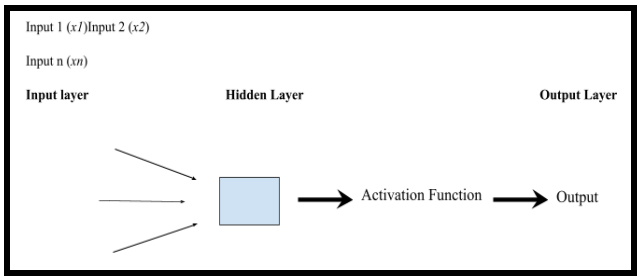

Fig 2. Simple diagram of a neural network.

A. Convoluted Neural Networks (CNNs)

Convolutional Neural Networks are multi-layer networks primarily utilized for image processing and object detection. Yann LeCun created the first CNN, known as LeNet, in 1988. LeNet was used for character recognition, such as ZIP codes and numbers. CNN's are frequently used to detect abnormalities, identify satellite pictures, analyze medical imaging, forecast time series, and find anomalies [15]. CNN has a convolution layer with many filters to conduct a rectified Linear Unit (ReLU) layer for element operations. The result is a rectified feature map, which is fed into a pooling layer. Pooling is a downsampling process that lowers the feature map's size. The pooling layer flattens the resultant two-dimensional arrays from the pooled feature map to create a single, long, continuous, linear vector. And a Fully connected layer, which classifies and labels the output when the flattened matrix from the pooling layer is presented as an input.

B. Logistic Regression

Logistic regression is a prominent Machine Learning method that is part of the Supervised Learning approach. It is employed in the prediction of the categorical dependent variable from a group of independent factors. The outcome must be an absolute or discrete value. It can be Yes or No, 0 or 1, true or false, but instead of presenting the precise values like 0 and 1, it offers the probability values that fall between 0 and 1. Logistic regression assumes that the dependent variable should be categorical and that the independent variable is not multi-collinear [16].
Except for how they are utilized, Logistic Regression and Linear Regression are pretty similar. Logistic regression is used to solve classification difficulties, whereas linear regression is used to solve regression problems. An S-shaped logistic function that predicts two maximum values in logistic regression (0 or 1) is provided instead of fitting a regression line. The logistic regression hypothesis tends to restrict the cost function between 0 and 1. As a result, linear functions fail to describe it since they might have a value

larger than one or less than 0, which is not conceivable according to the expectation of the logistic regression hypothesis.

$$0 \leq h\Theta(x) \leq 1.$$

The sigmoid function maps predictions to probabilities. The function maps any real value into another value between 0 and 1.

$$f(x) = 1 / 1+e^{-(x)}$$

The sigmoid function is a statistical function that is used to convert expected values into probabilities. It converts any real value between 0 and 1 into another value.

C. Long-Short Term Memory Network (LSTMs)

LSTMs are a kind of Recurrent Neural Network (RNN) capable of learning and remembering long-term dependencies. The default habit is to recall prior knowledge over extended periods. LSTMs keep track of information throughout time. Because they recognize previous inputs, they are valuable in time-series prediction. LSTMs feature a chain-like structure with four interacting layers that communicate distinctly. LSTMs ignore unnecessary portions of the initial state, then update the cell-state values selectively, and ultimately output specific parts of the cell state.

The dataset has been collected and stored on Amazon S3 for the next detection stage. After the data is collected the distributed algorithms are developed on Amazon cloud AWS EMR with different machines cores. The table of the experiment training time is shown on Table III. It's clear that the training time with a single system takes much more time than a distributed framework.

Hardware	Number of Nodes	NO. of cores	Training Time
MAC OSx, processor 2,5 GHz Intel Core i7- RAM 16 GB -1600 MHz DDR3	Locally	2	-CNN: 120 minutes -logistic_regression : 98 minutes -linear regression : 90 minutes
1 m4.4xlarge EMR	1 node	8	-CNN: 90 minutes -logistic_regression : 84 minutes -linear regression : 70 minutes
3 m4.4xlarge EMR, 3 nodes	3 nodes	24	-CNN: 45 minutes -logistic_regression : 42 minutes -linear regression : 37 minutes
6 m4.4xlarge EMR	6 nodes	48	-CNN: 19 minutes -logistic_regression : 14 minutes -linear regression : 10 minutes
8 m4.4xlarge EMR	8 nodes	64	-CNN: 6 minutes -logistic regression : 4 minutes -linear regression : 2 minutes

TABLE 1. The Training Time On EMR Amazon Cloud.

The experimental results are displayed using the following measurement metric:

(i) True positive (TP): a phishing website is correctly classified as phishing.
True Positive (TP) (Recall) = TP / (TP+FN)
(ii) True negative (TN): a legitimate website is correctly classified as legitimate.
True Negative (TN) = TN/ (TN+FP)
(iii) False positive (FP): a legitimate website is misclassified as phishing.
False Positive (FP) = FP / (FP+TN)
(iv) False negative (FN): a phishing website is misclassified as legitimate.
False Negative (FN) = FN / (FN+TP)

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

Metric	CNN	logistic regression	linear regression
Accuracy	98.00%	99.97%	96.62%
Recall / TP / Detection Rate	80.68%	98.05%	75.10%
FP	1.70%	0.0001%	3.40%
Precision / TP	28.60%	99.74%	22.15%
FN	19.31%	1.90%	24.89%
TN	98.26%	99.99%	97.70%

TABLE 2. Evaluation of The Performance for Proposed System

4. Conclusion

Nowadays, social engineering attacks are one of the most significant cyber attacks. Phishing attack is a kind of the social engineering attack. In this paper, distributed framework was developed using deep learning algorithms. It has been trained three classification models to distinguish phishing websites and normal ones.
In conclusion, the proposed system could provide significant benefits to current traditional techniques against phishing attacks. The proposed system has two subsystems. Firstly: collection data corresponds to normal and phishing websites. Secondly: distributed framework to apply different neural networks with different machines cores. The result of the proposed system achieved high accuracy and detection rate. This application can be extended with real time detection.

References

[1] Atkins, B., & Huang, W. (2013). A study of social engineering in online frauds. *Open Journal of Social Sciences*, *1*(03), 23.

[2] Parra, G. D. L. T., Rad, P., Choo, K. K. R., & Beebe, N. (2020). Detecting Internet of Things attacks using distributed deep learning. *Journal of Network and Computer Applications*, *163*, 102662.

[3] Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, July). Detection of Phishing Websites by Using Machine Learning-Based URL Analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

[4] Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., & Woźniak, M. (2020). Accurate and fast URL phishing detector: a convolutional neural network approach. *Computer Networks*, *178*, 107275.

[5] Sánchez-Paniagua, M., Fidalgo, E., González-Castro, V., & Alegre, E. (2020, September). Impact of current phishing strategies in machine learning models for phishing detection. In *Conference on Complex, Intelligent, and Software Intensive Systems* (pp. 87-96). Springer, Cham.

[6] Shie, E. W. S. (2020). Critical analysis of current research aimed at improving detection of phishing attacks. *Selected computing research papers*, 45.

[7] Wang, Zhilong, et al. "To detect stack buffer overflow with polymorphic canaries." *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2018.

[8] Kalaharsha, P., & Mehtre, B. M. (2021). Detecting Phishing Sites--An Overview. *arXiv preprint arXiv:2103.12739*.

[9] Azeez, N., Misra, S., Margaret, I. A., & Fernandez-Sanz, L. (2021). Adopting Automated Whitelist Approach for Detecting Phishing Attacks. *Computers & Security*, 102328.

[10] Khan, S. A., Khan, W., & Hussain, A. (2020, October). Phishing attacks and websites classification using machine learning and multiple datasets (A comparative analysis). In *International Conference on Intelligent Computing* (pp. 301-313). Springer, Cham.

[11] Alsariera, Y. A., Elijah, A. V., & Balogun, A. O. (2020). Phishing website detection: forest by penalizing attributes algorithm and its enhanced variations. *Arabian Journal for Science and Engineering*, *45*(12), 10459-10470.

[12] Bell, S., & Komisarczuk, P. (2020, February). An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank. In *Proceedings of the Australasian Computer Science Week Multiconference* (pp. 1-11). https://doi.org/10.1145/3373017.3373020

[13] Ahmad, J., Farman, H., & Jan, Z. (2018). Deep learning methods and applications. *Deep Learning: Convergence to Big Data Analytics*, 31–42. https://doi.org/10.1007/978-981-13-3459-7_3

[14] Pradhananga, Y., Karande, S., & Karande, C. (2015, February). CBA: cloud-based bigdata analytics. In *2015 International Conference on Computing Communication Control and Automation* (pp. 47-51). IEEE. https://doi.org/10.1109/iccubea.2015.18

[15] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-6). IEEE.

[16] Connelly, L. (2020). Logistic regression. *Medsurg Nursing, 29*(5), 353-354.

Sabah Alzahrani received the B.Sc. degree in Computer Science from Taif University, Saudi Arabia, in 2007. the M.Sc. degree and Ph.D degree. in computer and information systems engineering, from Tennessee State University, United States in 2015 and 2018 respectively. He is currently an Assistant Professor with department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia. Her research interests include the Internet of Things, Cyber Security, Computer Networking, Cloud, and Big Data.