Brain Storm Optimization with Long Short Term Memory Enabled Phishing Webpage Classification for Cybersecurity

Mahmoud A. Zaher[1*], Nabil M. Eldakhly[2]

[1] Faculty of Artificial Intelligence, Egyptian Russian University (ERU), Cairo, Egypt

[2] Department of Computer and Information Systems Sadat Academy for Management Sciences (SAMS), Cairo, Egypt

Emails: Mahmoud.zaher@eru.edu.eg; nmeldakhly@yahoo.com

Abstract

Phishing is a familiar kind of cyberattack in the present digital world. Phishing detection with maximum performance accuracy and minimum computational complexity is continuously a topic of much interest. A novel technology was established for improving the phishing detection rate and decreasing computational constraints recently. But, one solution has inadequate for addressing every problem due to attackers from cyberspace. Thus, the initial objective of this work is for analysing the performance of different deep learning (DL) techniques from detection phishing activity. This study introduces a novel Brain Storm Optimization with Long Short Term Memory Enabled Phishing Webpage Classification (BSOLSTM-PWC) for Cybersecurity. The proposed BSOLSTM-PWC technique enables to accomplish cybersecurity by the identification and classification of phishing webpages. To accomplish this, the BSOLSTM-PWC technique initially employs data pre-processing technique to transform the data into actual format. Besides, the BSOLSTM-PWC technique employs LSTM classifier for the identification and categorization of phishing webpages. Moreover, the BSO algorithm is utilized to appropriately adjust the hyperparameters involved in the LSTM model. For reporting the improved outcomes of the BSOLSTM-PWC method, a wide-ranging simulation analysis is made using benchmark dataset. The experimental outcomes reported the enhanced outcomes of the BSOLSTM-PWC method on existing methods.

Keywords: Cybersecurity; Website phishing; Classification; Brain storm optimization; Long short term memory; Deep learning

1. Introduction

The Internet is progressively utilized by people and associations to perform various exercises, for example, individual, exchanges and other business-related errands. Today, more people, associations, and legislatures are progressively taking on and keeping up with online presence [1]. Therefore, this, among different advantages, has added to the development of numerous organizations [2]. Then again, the Internet can be utilized for noxious purposes. One way is to take advantage of the Internet to execute disconnected assaults. For example, fear based oppressors utilize the Internet to plan and arrange their assaults [3]. There are additional assaults that are executed on the Internet. These incorporate spreading malware, counterfeit news and disdain talks, online tricks, personality, and protected innovation burglary, and cyberbullying. These have brought about loss of secret data, enormous monetary misfortunes, reputational harms, to specify yet not many [4]. Another normal assault Internet clients are defenseless to be phishing. With the data accumulated, further cybercrimes, for example, extorting and personality burglaries, can be executed [5].

Phishing URLs are typically created by aggressors to resemble genuine and harmless URLs, to bamboozle clueless Internet clients [6]. The objective is to acquire the trust of such clients to tap the URLs and uncover their touchy data. The aim of the aggressor could likewise be to download malware

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

to casualties' PCs [7]. For this situation, the downloaded pernicious application can take data from the PCs of the people in question, which are then sent to the assailant. The malware could likewise be utilized to download other malevolent documents or make indirect access to the contaminated framework, which would permit the aggressor to remotely control it. Other potential objectives are erasing or altering casualties' data; and scrambling the whole documents, wherein case, the casualty is mentioned to pay a payoff [8].

Numerous strategies have been acquainted with making the IoT climate safer, yet there is right now no powerful technique for recognizing phishing messages. Few examinations have been directed to propose approaches and techniques for recognizing phishing websites for the IoT platform. In the beyond couple of years, deep learning (DL) procedures are shown that a successful arrangement amongst applications across numerous disciplines, including Internet of Things (IoT), IDS, ransomware location, and so forth [9]. Various specialists in digital protection stand out towards DL calculations. Prominently, analysts and security specialists have additionally perceived its importance in the phishing location area [10]. In this manner, different enemy of phishing arrangements has been created to identify phishing dangers ahead of schedule to limit the security gambles and safeguard the end-clients. Among them, website phishing discovery in view of DL calculations has grabbed a lot of eye in late investigations. Security techniques in view of DL systems have become progressively famous to manage to advance phishing assaults. There are various sorts of DL strategies intended to tackle a particular issue or meet a framework's specific necessity; each enjoys its benefits and inconveniences.

Adeyemo et al. [11] presented an ensemble-based Logistic Model Trees (LMT) for detecting web site phishing attacks. LMT is the integration of tree induction and logistic regression approaches as to single model tree. The experiment result shows that the presented method is very efficient for detecting web site phishing attacks. Adebowale et al. [12] introduced an Adaptive Neuro-Fuzzy Inference System (ANFIS) based strong structure with the combined feature of the frames, text, and images for detecting and protecting web-phishing. The projected solution is the initial task that considered the optimally incorporated frame, text, and image feature-based solution for detecting phishing systems. Balogun et al. [13] developed a Rotation Forest-based Logistic Model Trees (RF-LMT) for detecting web site phishing. LMT is a system that integrates tree inference and logistic regression as to single model tree. Three data sets of distinct instance distributions, balanced and imbalanced, are employed for investigating the presented method. From the outcomes, it is noted that LMT implemented well when compared to the chosen baseline classifier. The result reveals that LMT implements comparatively to baseline classifier.

Lakshmi et al. [14] proposed an advanced method to recognize phishing web site through hyperlinks presented from the source-code of the HTML page from the respective web site. The suggested technique employs a feature vector with thirty variables for detecting malignant websites. This feature is utilized in training the supervised DNN with Adam optimizer to differentiate fraudulent web sites from reliable web sites. Jain and Gupta [15] assessed the efficiency of the presented phishing detection technique on different classification algorithms with the phishing and non-phishing web sites. The presented technique is a completely client-side solution and doesn’t need other services from the 3rd parties.

This study introduces a novel Brain Storm Optimization with Long Short Term Memory Enabled Phishing Webpage Classification (BSOLSTM-PWC) for Cybersecurity. The proposed BSOLSTM-PWC technique initially employs data pre-processing technique to transform the data into actual format. Besides, the BSOLSTM-PWC technique employs LSTM classifier for the identification and categorization of phishing webpages. Moreover, the BSO algorithm is utilized to appropriately adjust the hyperparameters involved in the LSTM model. For reporting the enhanced outcomes of the BSOLSTM-PWC approach, a wide-ranging simulation analysis is made using benchmark dataset.

2. The Proposed Model

In this study, a novel BSOLSTM-PWC method was established to accomplish cybersecurity by the identification and classification of phishing webpages. The BSOLSTM-PWC technique incorporates data pre-processing technique to transform the data into actual format. Besides, the BSOLSTM-PWC technique employs LSTM classifier for the identification and categorization of phishing webpages. Moreover, the BSO algorithm is utilized to appropriately adjust the hyperparameters involved in the LSTM model.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

2.1 LSTM based Phishing Classification

At the time of phishing detection and classification, the LSTM classifier for the identification and categorization of phishing webpages. A main disadvantage of the conventional RNN approach is that if the time step enhances, the network attains failed for deriving the context from the time step of earlier state so many fear later as phenomenon is named long-term dependency. According to the deep layer of networks and recurrent efficiency of usual RNN, exploding and vanishing gradient issue is also encountered quite frequently. Besides, in order to report this problem, the LSTM approaches are recognized by utilizing memory cells with numerous gates from hidden layers [16, 17]. The block of hidden layer with LSTM cell unit and 3 drives of gate controller as:

- The forget gate f_t select that measured of long-term state c_t is misplaced;
- An input gate i_t mechanism that measures of $\tilde{c}_t$ is more to long-term procedure c_t;
- The output gate g_t determines that quantity of c_t is read and output to h_t and o_t.

The subsequent procedures demonstrate the long- and short-terms procedures of cells and output of every layer from time step:

$$i_t = \sigma(W_x^T i \cdot x_t + W_h^T i \cdot h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma\left(W_{xf}^T \cdot x_t + W_{hf}^T \cdot h_{t-1} + b_f\right) \quad (2)$$

$$o_t = \sigma(W_x^T o \cdot x_t + W_h^T o \cdot h_{t-1} + b_i) \quad (3)$$

$$g_t = tanh(W_x^T g \cdot x_t + W_h^T g \cdot h_{t-1} + b_i) \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t. \quad (5)$$

$$o_t, h_t = g_t \otimes tanh(c_t). \quad (6)$$

whereas $W_x f, W_x i, W_x o, W_x g$ represents the weighted matrices to corresponding connected input vector, $W_h f, W_h i, W_h o, W_h g$ refers the weighted matrices of short-term procedure of earlier time step, and $b_f, b_i, b_o,$ and b_g are bias. Fig. 1 illustrates the framework of LSTM technique.

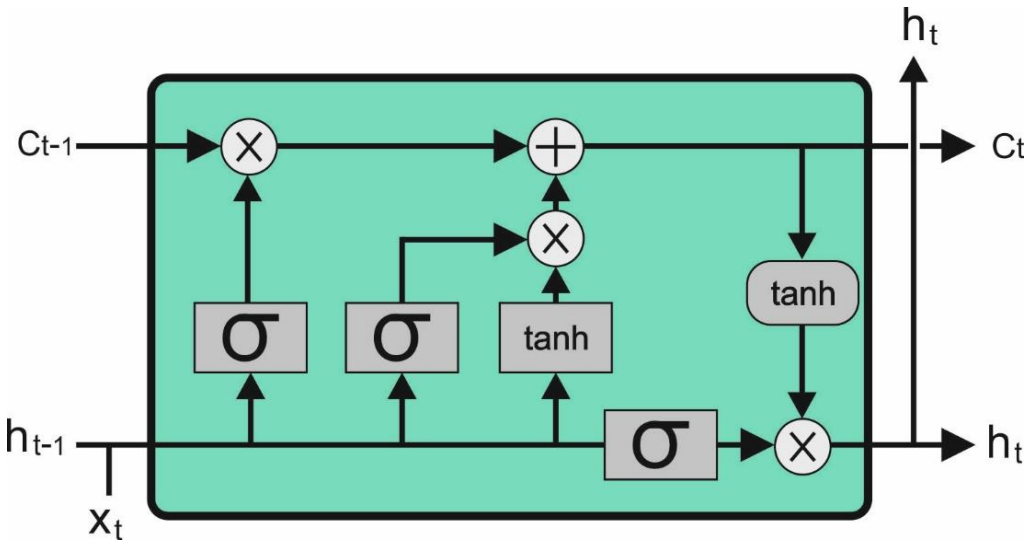

Figure 1: LSTM structure

2.2 BSO based Hyperparameter Optimization

For enhancing the classifier results of the LSTM technique, the BSO method was employed to appropriately adjust the hyperparameters contained in the LSTM technique. The BSO approach is stimulated by the concept of brainstorming, and it is a widely employed mechanism to increase creativeness from the organization that has obtained wider acceptance [18, 19]. In BSO method, first, N models are initiated arbitrarily from the solution space, and next all the concepts are evaluated

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

according to their FF. Next, m point of cluster center has been elected arbitrarily and introduced as N concepts, while m is less than N. The remaining process of BSO is described in the following.

Clustering Individual is a process of sorting equivalent objects, and, each generation, is re-clustered into m clusters according to the process (or individual) feature. K means has popular method applied from the cluster; now it can be employed from the cluster method. The cluster center disrupting method arbitrarily chooses a cluster center and replaced it with a newly developed concept with a possibility of $p\_replace$ that is named as a replacing function. The $p\_replace$ value has been utilized to control the possibility of replacing a cluster center with arbitrarily created solution.

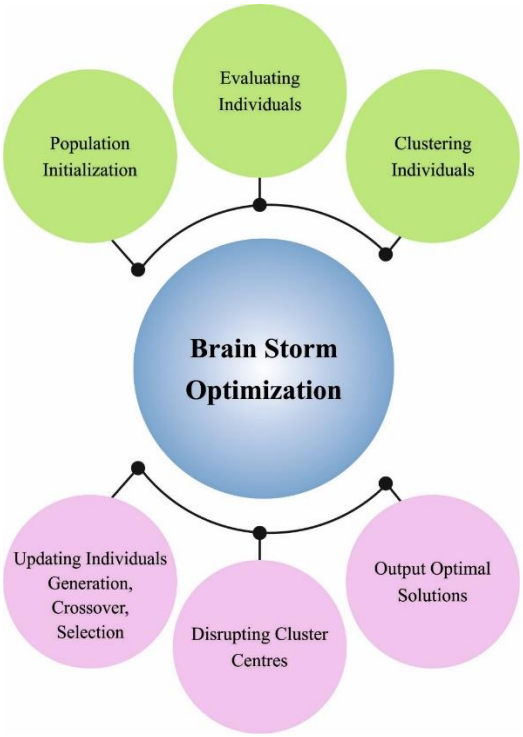

Figure 2: Process involved in BSO technique

For maintaining diversity of the population, a concept (individual) is generated according to 1 or 2 in 1 cluster or 2. Next, the selection of 1 cluster or 2, the method of cluster center or arbitrary process can be elected by a possibility of p_one_center and p_two_center.

$$X_{select} = \begin{cases} X_i & one\ cluster \\ rand * X_{i1} + (1 - rand) * X_{i2} & two\ clusters' \end{cases} \quad (7)$$

while rand represents an arbitrary number within [0, 1]. Afterward choosing one or two concept, the selected idea(s) has been upgraded by the following equation. Fig. 2 depicts the process involved in BSO technique.

$$X_{\text{new}} = X_{\text{select}} + \xi * normrnd\ (0,1), \quad (8)$$

Now, $normrnd$ signifies the Gaussian random number with mean zero and variance one and ξ characterizes a changing factor that slows down the convergence rate as follows

$$\xi = rand * \log sig\left(\frac{(0.5 * \max\_iterat - current\_iterat)}{k}\right), \quad (9)$$

Here rand signifies an arbitrary value within [0, 1]. The max and existing iteration represent the present amount of iterations and maximal amount of iterations. $logsig$ shows the logarithmic sigmoid transfer function, and the technique is effective for global searching capability at the beginning of development and enhances local searching ability when the process is approaching at the end. k represents an existing parameter to change the slope of logsig function.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

3. Experimental Validation

In this section, the performance validation of the BSOLSTM-PWC model is performed using three datasets. The first UCI dataset includes 11055 instances with 4898 phishing and 6157 legitimate instances. The second Huddersfield_1 dataset comprises 2456 instances with 1094 phishing and 1362 legitimate instances. The third Huddersfield_2 dataset includes 2670 instances with 1485 phishing and 1185 legitimate instances as shown in Table 1.

Table 1 Datasets details

Datasets	No. of Instances	Phishing	Legitimate
UCI	11055	4898	6157
Huddersfield_1	2456	1094	1362
Huddersfield_2	2670	1485	1185

Fig. 3 illustrates a set of three confusion matrices offered by the BSOLSTM-PWC model. On the test UCI dataset, the BSOLSTM-PWC model has identified 4805 phishing and 6096 legitimate instances. In addition, on the test Huddersfield_1 dataset, the BSOLSTM-PWC method has identified 1066 phishing and 1353 legitimate instances. Also, on the test Huddersfield_2 dataset, the BSOLSTM-PWC approach has identified 1299 phishing and 1146 legitimate instances.

Table 2 provides detailed phishing classification outcomes of the BSOLSTM-PWC model on three distinct datasets.

Fig. 4 reports a comprehensive phishing detection and classification performance of the BSOLSTM-PWC model on the UCI dataset. The figure indicated that the BSOLSTM-PWC model has classified samples under phishing class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 98.61%, 98.75%, 98.10%, 98.42%, and 98.56% respectively. In addition, the BSOLSTM-PWC model has categorized samples under legitimate class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 98.61%, 98.50%, 99.01%, 98.75%, and 98.56% respectively.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

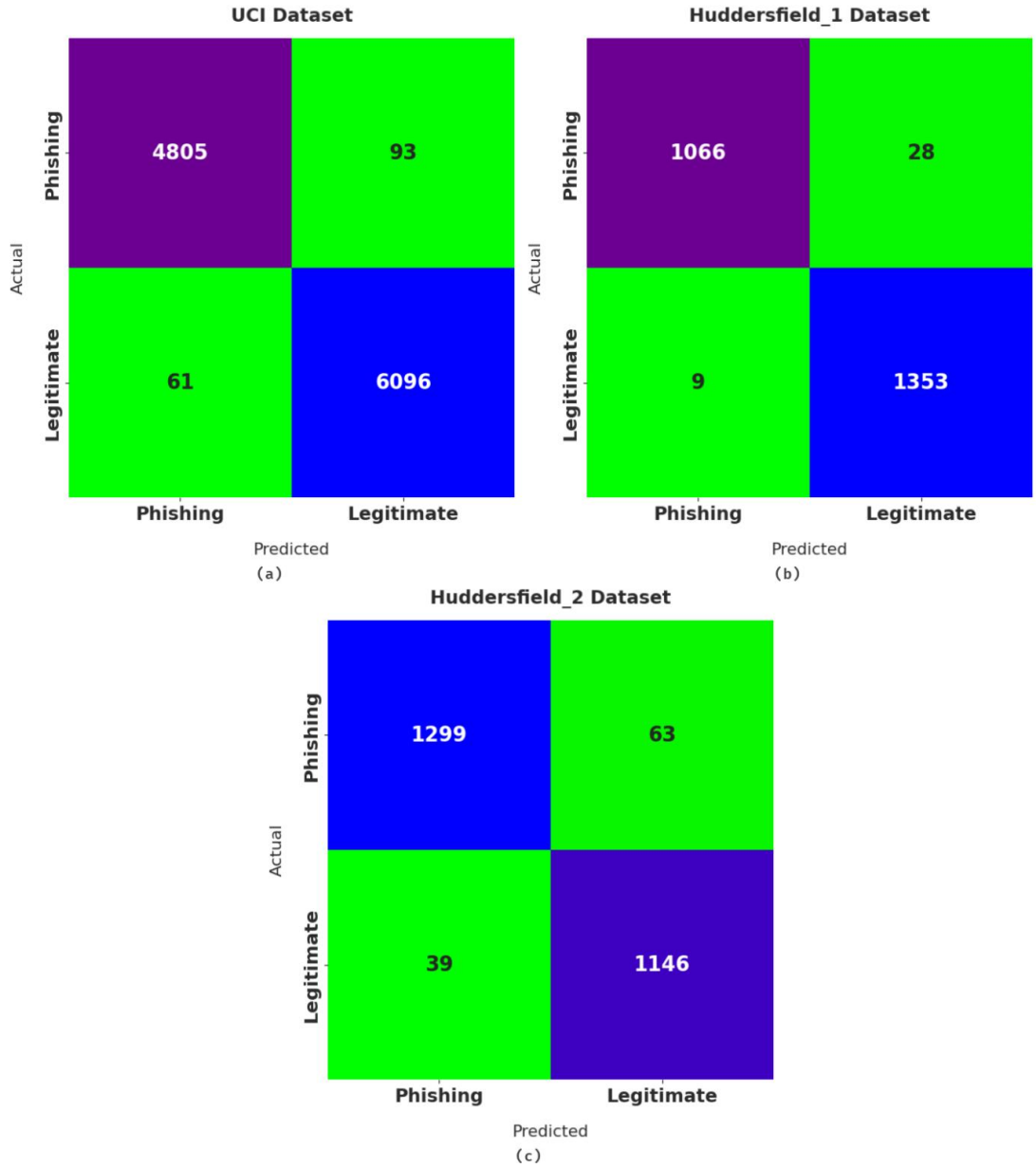

Figure 3: Confusion matrix of BSOLSTM-PWC technique under three datasets

Fig. 5 demonstrates a comprehensive phishing detection and classification performance of the BSOLSTM-PWC model on the Huddersfield_1 dataset. The figure referred that the BSOLSTM-PWC algorithm has classified samples under phishing class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 98.49%, 99.16%, 97.44%, 98.29%, and 98.39% correspondingly. At the same time, the BSOLSTM-PWC approach has categorized samples under legitimate class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 98.49%, 97.97%, 99.34%, 98.65%, and 98.39% correspondingly.

Fig. 6 showcases a comprehensive phishing detection and classification performance of the BSOLSTM-PWC algorithm on the Huddersfield_2 dataset. The figure exposed that the BSOLSTM-PWC technique has classified samples under phishing class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 96%, 97.09%, 95.37%, 96.22%, and 96.04% correspondingly. Additionally, the BSOLSTM-PWC model has categorized samples under legitimate class with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 96%, 94.79%, 96.71%, 95.74%, and 96.04% correspondingly.

Table 2: Result analysis of BSOLSTM-PWC technique with distinct measures and datasets

Class Labels	Accuracy	Precision	Recall	F-Score	ROC Score

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

UCI Dataset					
Phishing	98.61	98.75	98.10	98.42	98.56
Legitimate	98.61	98.50	99.01	98.75	98.56
Average	**98.61**	**98.62**	**98.56**	**98.59**	**98.56**
Huddersfield_1 Dataset					
Phishing	98.49	99.16	97.44	98.29	98.39
Legitimate	98.49	97.97	99.34	98.65	98.39
Average	**98.49**	**98.57**	**98.39**	**98.47**	**98.39**
Huddersfield_2 Dataset					
Phishing	96.00	97.09	95.37	96.22	96.04
Legitimate	96.00	94.79	96.71	95.74	96.04
Average	**96.00**	**95.94**	**96.04**	**95.98**	**96.04**

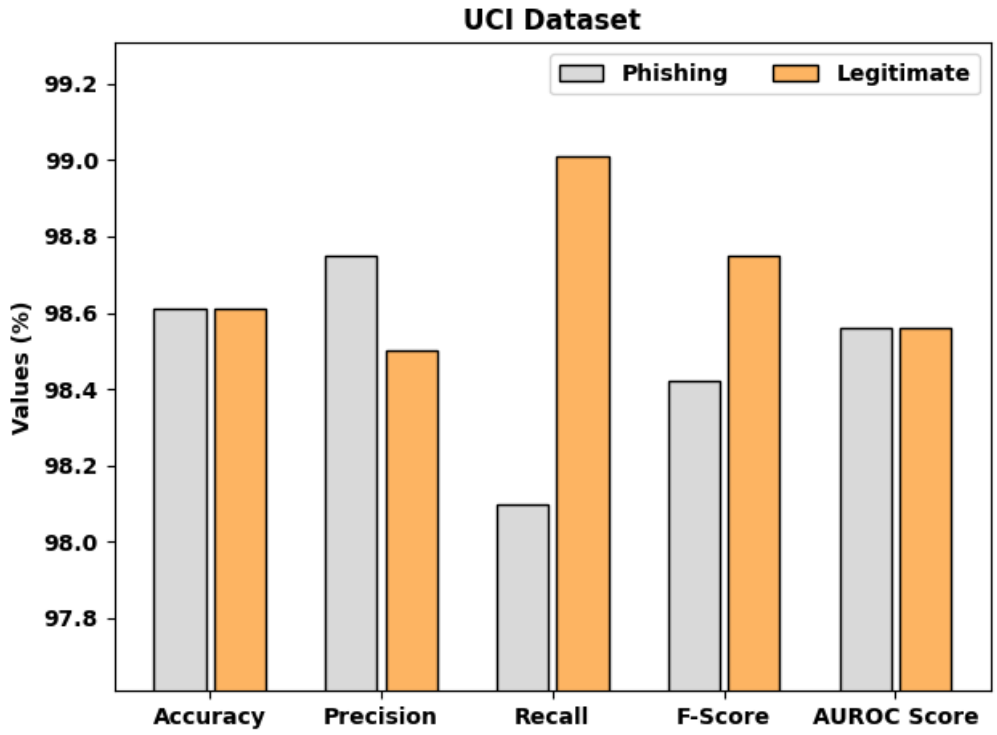

Figure 4: Result analysis of BSOLSTM-PWC technique under UCI dataset

Table 3 demonstrates a comprehensive comparison study of the BSOLSTM-PWC model with recent models interms of different measures. [20] The comparative results indicated that the BSOLSTM-PWC model has accomplished maximum performance with $accu_y$, $prec_n$, $reca_l$, F_{score}, and ROC_{score} of 98.61%, 98.62%, 98.56%, 98.59%, and 98.56% respectively.

Fig. 7 indicates the comparative $prec_n$ and $reca_l$ investigation of the BSOLSTM-PWC model. The figure implied that the RT model has gained worse outcome with minimal values of $prec_n$ and $reca_l$. Followed by, the MLP model has resulted in slightly improved values of $prec_n$ and $reca_l$. Next to that, the bagging, LMT, Kstar, and RF models have accomplished moderately closer values of $prec_n$ and $reca_l$. But the BSOLSTM-PWC model has resulted in superior $prec_n$ and $reca_l$ values of 98.61% and 98.62% respectively.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

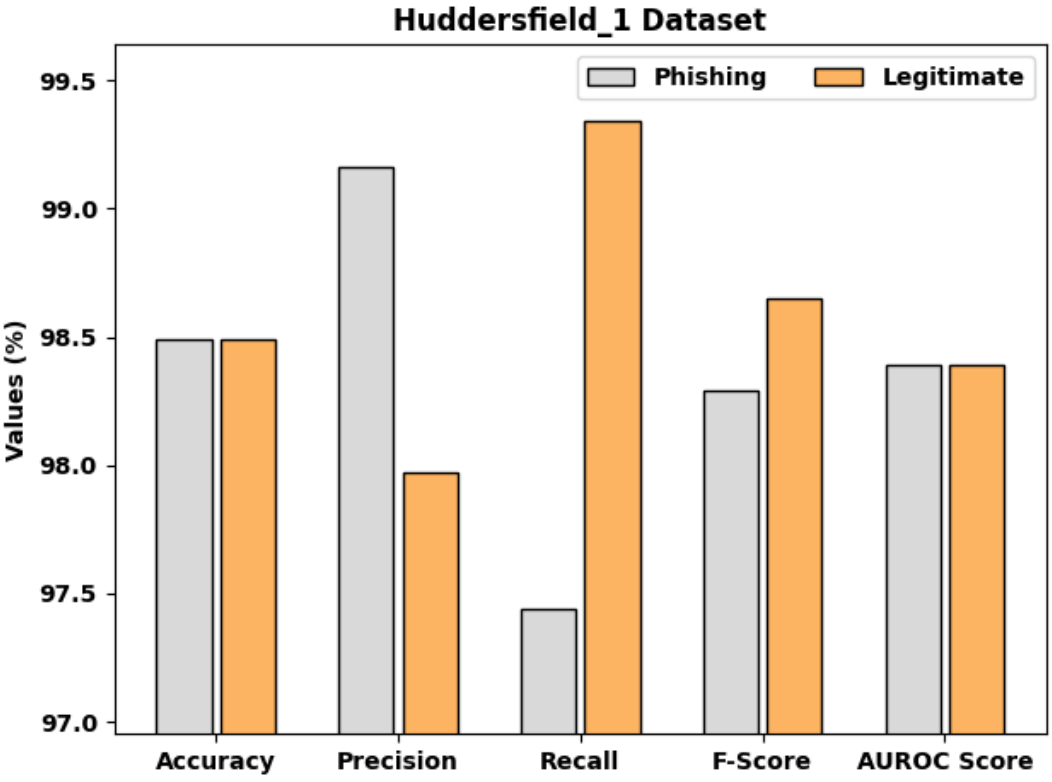

Figure 5: Result analysis of BSOLSTM-PWC technique under Huddersfield_1dataset

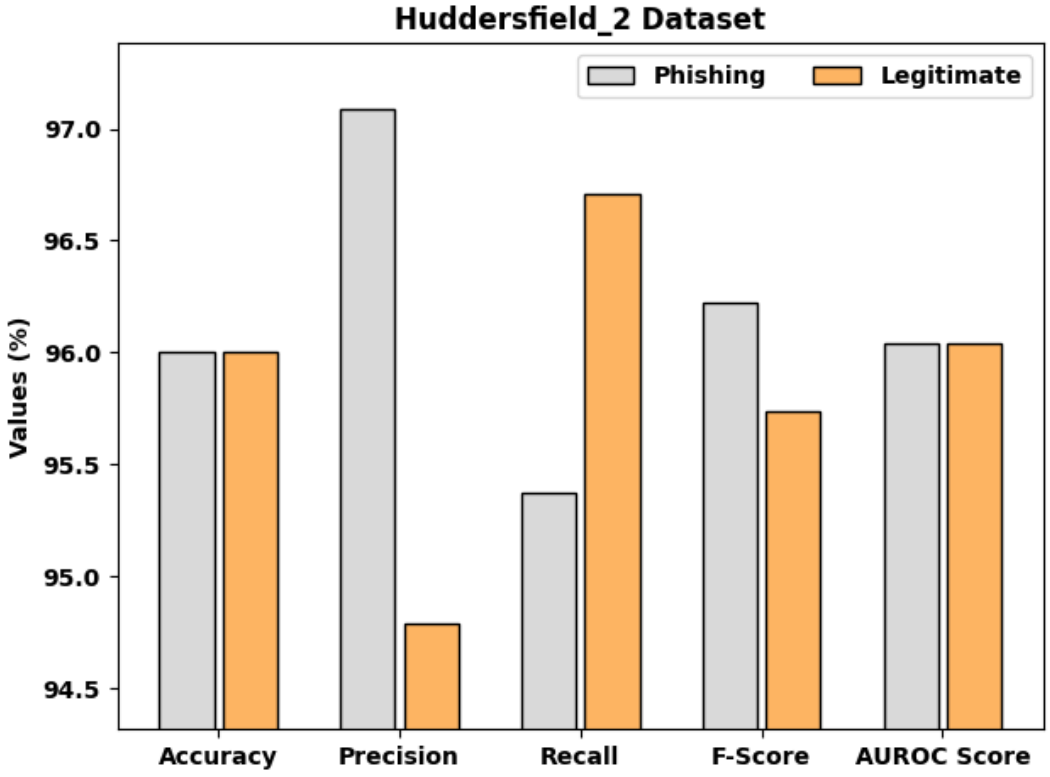

Figure 6: Result analysis of BSOLSTM-PWC technique under Huddersfield_2dataset

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

Table 3: Comparative analysis of BSOLSTM-PWC technique with recent algorithms

Methods	Accuracy	Precision	Recall	F-Measure	ROC Score
Bagging Algorithm	98.41	97.54	98.39	96.75	98.02
LMT Algorithm	96.43	97.21	97.21	97.54	97.57
MLP Algorithm	96.58	96.53	97.08	96.40	98.35
Kstar Algorithm	97.19	97.30	96.68	98.47	98.44
RandomForest Algorithm	98.27	97.98	97.62	98.22	98.13
RandomTree Algorithm	97.49	95.75	96.28	96.68	96.62
BSOLSTM-PWC	98.61	98.62	98.56	98.59	98.56

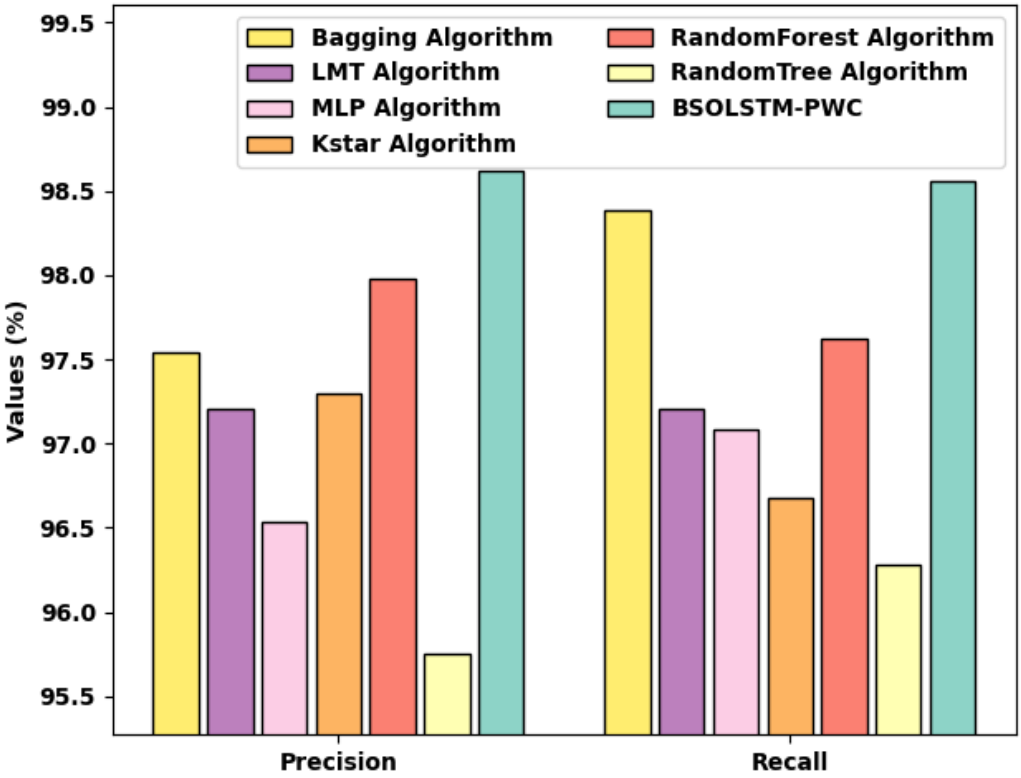

Figure7: $Prec_n$ and $reca_l$ analysis of BSOLSTM-PWC technique with recent algorithms

Fig. 8 specifies the comparative $F_{measure}$ and ROC score analysis of the BSOLSTM-PWC method. The figure implied that the RT model has gained worse outcome with minimal values of $F_{measure}$ and ROC score. Likewise, the MLP model has resulted to somewhat enhanced values of $F_{measure}$ and ROC score. Then, the bagging, LMT, Kstar, and RF approaches have accomplished moderately closer values of $F_{measure}$ and ROC score. However the BSOLSTM-PWC approach has resulted to superior $F_{measure}$ and ROC score values of 98.59% and 98.56% correspondingly.

After observing the above mentioned results and discussion, it is apparent that the BSOLSTM-PWC model has resulted in maximum classification outcome over the other methods.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

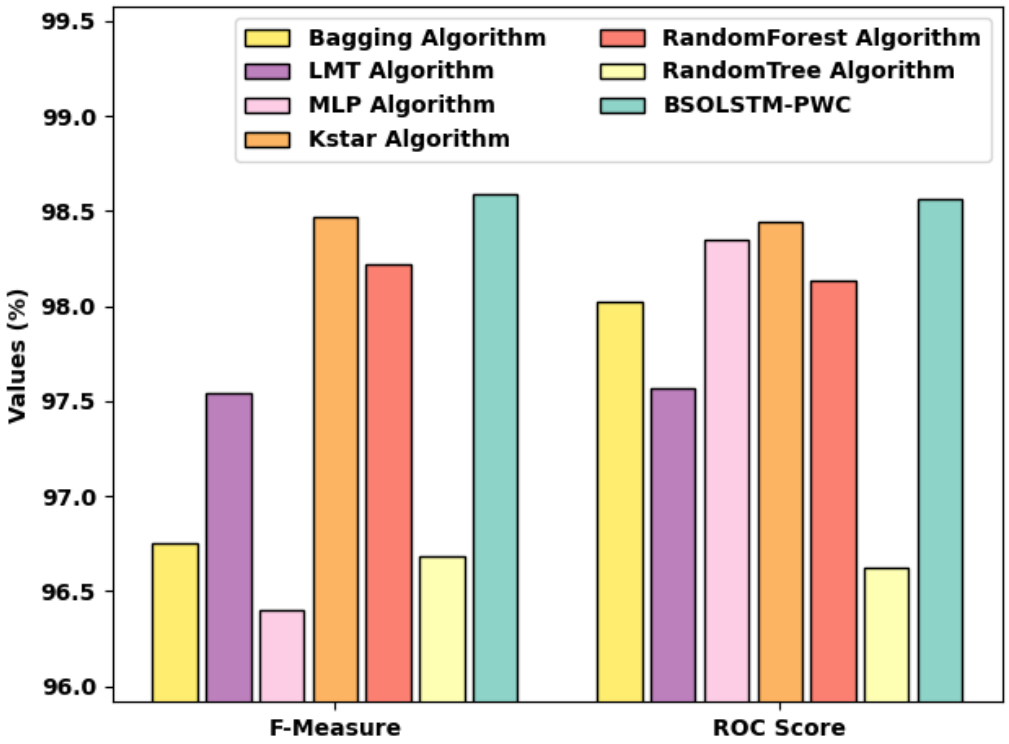

Figure 8: $F_{measure}$ and ROC score analysis of BSOLSTM-PWC technique with recent algorithms

4. Conclusion

In this study, a novel BSOLSTM-PWC approach was established to accomplish cybersecurity by the identification and classification of phishing webpages. The BSOLSTM-PWC technique incorporates data pre-processing technique to transform the data into actual format. Besides, the BSOLSTM-PWC technique employs LSTM classifier for the identification and categorization of phishing webpages. Moreover, the BSO algorithm is utilized to appropriately adjust the hyperparameters involved in the LSTM model. For reporting the enhanced outcomes of the BSOLSTM-PWC technique, a wide-ranging simulation analysis is made using benchmark dataset. The experimental outcomes reported the enhanced outcomes of the BSOLSTM-PWC approach on existing methods. In future, hybrid DL models are employed to enhance the classifier outcomes of the BSOLSTM-PWC technique.

References

[1] Gandotra, E. and Gupta, D., 2021. An efficient approach for phishing detection using machine learning. In Multimedia Security (pp. 239-253). Springer, Singapore.

[2] Shahrivari, V., Darabi, M.M. and Izadi, M., 2020. Phishing Detection Using Machine Learning Techniques. arXiv preprint arXiv:2009.11116.

[3] Aljofey, A., Jiang, Q., Qu, Q., Huang, M. and Niyigena, J.P., 2020. An effective phishing detection model based on character level convolutional neural network from URL. Electronics, 9(9), p.1514.

[4] Al-Ahmadi, S., 2020. PDMLP: phishing detection using multilayer perceptron. International Journal of Network Security & Its Applications (IJNSA) Vol, 12.

[5] El Aassal, A., Baki, S., Das, A. and Verma, R.M., 2020. An in-depth benchmarking and evaluation of phishing detection research for security needs. IEEE Access, 8, pp.22170-22192.

[6] Jain, A.K. and Gupta, B.B., 2018. PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[7] Wang, W., Zhang, F., Luo, X. and Zhang, S., 2019. Pdrcnn: precise phishing detection with recurrent convolutional neural networks. Security and Communication Networks, 2019.

[8] Vishva, E.S. and Aju, D., 2022. Phisher Fighter: Website Phishing Detection System Based on URL and Term Frequency-Inverse Document Frequency Values. Journal of Cyber Security and Mobility, pp.83-104.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022

[9] Balogun, A.O., Mojeed, H.A., Adewole, K.S., Akintola, A.G., Salihu, S.A., Bajeh, A.O. and Jimoh, R.G., 2021, October. Optimized Decision Forest for Website Phishing Detection. In Proceedings of the Computational Methods in Systems and Software (pp. 568-582). Springer, Cham.

[10] Tang, L. and Mahmoud, Q.H., 2021. A survey of machine learning-based solutions for phishing website detection. Machine Learning and Knowledge Extraction, 3(3), pp.672-694.

[11] Adeyemo, V.E., Balogun, A.O., Mojeed, H.A., Akande, N.O. and Adewole, K.S., 2020, December. Ensemble-based logistic model trees for website phishing detection. In International Conference on Advances in Cyber Security (pp. 627-641). Springer, Singapore.

[12] Adebowale, M.A., Lwin, K.T., Sanchez, E. and Hossain, M.A., 2019. Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications, 115, pp.300-313.

[13] Balogun, A.O., Akande, N.O., Usman-Hamza, F.E., Adeyemo, V.E., Mabayoje, M.A. and Ameen, A.O., 2021, September. Rotation Forest-Based Logistic Model Tree for Website Phishing Detection. In International Conference on Computational Science and Its Applications (pp. 154-169). Springer, Cham.

[14] Lakshmi, L., Reddy, M.P., Santhaiah, C. and Reddy, U.J., 2021. Smart phishing detection in web pages using supervised deep learning classification and optimization technique adam. Wireless Personal Communications, 118(4), pp.3549-3564.

[15] Jain, A.K. and Gupta, B.B., 2019. A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 10(5), pp.2015-2028.

[16] Tasdelen, A. and Sen, B., 2021. A hybrid CNN-LSTM model for pre-miRNA classification. Scientific reports, 11(1), pp.1-9.

[17] Srinivasu, P.N., SivaSai, J.G., Ijaz, M.F., Bhoi, A.K., Kim, W. and Kang, J.J., 2021. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. Sensors, 21(8), p.2852.

[18] Shi, Y., 2011, June. Brain storm optimization algorithm. In International conference in swarm intelligence (pp. 303-309). Springer, Berlin, Heidelberg.

[19] Zhan, Z.H., Zhang, J., Shi, Y.H. and Liu, H.L., 2012, June. A modified brain storm optimization. In 2012 IEEE Congress on Evolutionary Computation (pp. 1-8). IEEE.

[20] Osho, O., Oluyomi, A., Misra, S., Ahuja, R., Damasevicius, R. and Maskeliunas, R., 2019, November. Comparative evaluation of techniques for detection of phishing URLs. In International Conference on Applied Informatics (pp. 385-394). Springer, Cham.

Doi : https://doi.org/10.54216/JCIM.090202
Received: February 05, 2022 Accepted: March 08, 2022