Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.028984
Article

Tech Science Press

Optimal Deep Belief Network Enabled Cybersecurity Phishing Email Classification

Ashit Kumar Dutta[1,*], T. Meyyappan[2], Basit Qureshi[3], Majed Alsanea[4], Anas Waleed Abulfaraj[5], Manal M. Al Faraj[1] and Abdul Rahaman Wahab Sait[6]

[1]Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh, 13713, Kingdom of Saudi Arabia
[2]Department of Computer Science, Alagappa University, Karaikudi, 630003, India
[3]Department of Computer Science, Prince Sultan University, Riyadh, 11586, Kingdom of Saudi Arabia
[4]Department of Computing, Arabeast Colleges, Riyadh, 11583, Kingdom of Saudi Arabia
[5]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 23613, Kingdom of Saudi Arabia
[6]Department of Archives and Communication, King Faisal University, Al Ahsa, Hofuf, 31982, Kingdom of Saudi Arabia
*Corresponding Author: Ashit Kumar Dutta. Email: adotta@mcst.edu.sa
Received: 22 February 2022; Accepted: 30 March 2022

Abstract: Recently, developments of Internet and cloud technologies have resulted in a considerable rise in utilization of online media for day to day lives. It results in illegal access to users' private data and compromises it. Phishing is a popular attack which tricked the user into accessing malicious data and gaining the data. Proper identification of phishing emails can be treated as an essential process in the domain of cybersecurity. This article focuses on the design of biogeography based optimization with deep learning for Phishing Email detection and classification (BBODL-PEDC) model. The major intention of the BBODL-PEDC model is to distinguish emails between legitimate and phishing. The BBODL-PEDC model initially performs data pre-processing in three levels namely email cleaning, tokenization, and stop word elimination. Besides, TF-IDF model is applied for the extraction of useful feature vectors. Moreover, optimal deep belief network (DBN) model is used for the email classification and its efficacy can be boosted by the BBO based hyperparameter tuning process. The performance validation of the BBODL-PEDC model can be performed using benchmark dataset and the results are assessed under several dimensions. Extensive comparative studies reported the superior outcomes of the BBODL-PEDC model over the recent approaches.

Keywords: Cybersecurity; phishing email; data classification; deep learning; biogeography based optimization; hyperparameter tuning

1 Introduction

With the rapid development of communication and global networking techniques, lots of our day-to-day life activities like electronic banking, e-commerce, social networks, and so on are transported to cyberspace

[1]. The uncontrolled, open, and anonymous structure of the Internet allows an outstanding environment for cyberattacks that presented severe security susceptibilities for standard computer users, experienced ones, and networks. The procedure of defending cyberspace from attack is called Cyber Security [2]. Cyber Security is recovering, protecting, and preventing each resource that uses the internet from cyberattack [3]. The difficulty in the cybersecurity field rises day-to-day, making controlling, identifying, and analyzing the appropriate risk event important problems. A cyberattack is digital malevolent attempt to intrude, steal, or damage the organizational confidential or personal information [4]. Even though experience of the user and carefulness are significant, it is impossible to entirely prevent users from falling into the phishing scam [5]. A phishing attack is a type of societal production attack widely employed for embezzling user data, including credit card numbers and login testimonials. This happens when an aggressor, hidden as a trusted individual, targets a victim to modify email data, namely content message, or instantaneous message. Fig. 1 illustrates the process of phishing classification model.

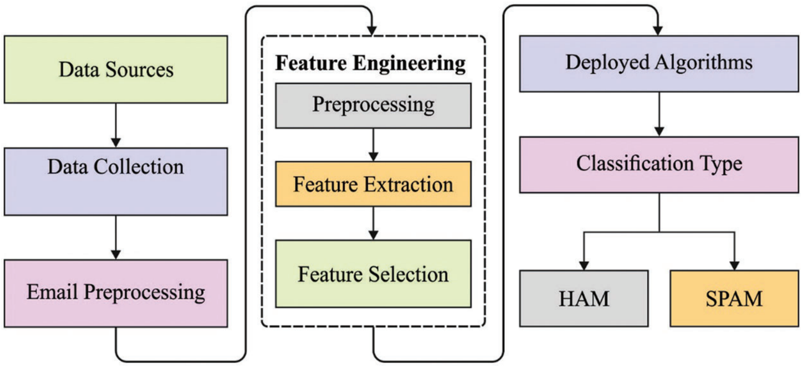

Figure 1: Process in phishing classification model

To obtain personal information, criminal develops illegal replicas of email and real websites, generally from an organization or financial institution handling financial information [6]. This e-mail is rendered by authentic slogans and company logos. The structure and design of hypertext markup language (HTML) allow copying of entire website or an image [7]. As well, it is the major factor for the quick expansion of Internet as a transmission network and allows the misuse of trademarks, brands, and company identifiers that customer relies on as validation mechanism [8]. To trap users, Phisher sends "spooled" emails to largest number of people. Once this e-mail is opened, the customer tends to be distracted from the authentic entity to spoofed websites. There is an important possibility of exploitation of user data. For that reason, phishing in current society is overly critical, very urgent, and challenging [9].

There are numerous researchers against phishing according to the faces of domains, like website content, website uniform resource locator (URL), incorporates this two website URL and content, the screenshot of the website and the source code of website [10]. But there is a lack of valuable anti-phishing tools to identify malevolent URLs in an institution for protecting their user. In case of malevolent code being rooted on the website, attackers might install malware and steal user information that possess a severe threat to user privacy and cybersecurity. Malicious URL on the Internet is identified easily by examining them via Machine Learning (ML) approach.

This article focuses on the design of biogeography based optimization with deep learning for Phishing Email detection and classification (BBODL-PEDC) model. The BBODL-PEDC model initially performs data pre-processing in three levels namely email cleaning, tokenization, and stop word elimination. In addition, Term Frequency — Inverse Document Frequency (TF-IDF) model is applied for the extraction of useful feature vectors. Followed by, optimal deep belief network (DBN) model is used for the email classification and its efficacy can be boosted by the BBO based hyperparameter tuning process. The performance validation of the BBODL-PEDC model can be performed using benchmark dataset and the results are assessed under several dimensions.

2 Related Works

Saha et al. [11] introduced the data-driven structure to detect phishing webpage utilizing deep learning (DL) technique. In particular, multilayer perceptron (MLP) that is also mentioned that feed forward neural network (FFNN) was utilized for predicting the phishing webpage. The data set is gathered in Kaggle and comprises data of ten thousand webpages. Opara et al. [12] presented HTMLPhish, a DL based data-driven end-to-end automatic phishing webpage classifier method. Especially, HTMLPhish takes the content of HTML document of a webpage and utilizes convolutional neural network (CNN) for learning the semantic dependence from the textual content of HTML. The CNN learned suitable feature representation in the HTML document embedded with no extensive manual feature engineering.

Ra et al. [13] utilized word embedded and Neural Bag-of-ngrams with DL approaches for detecting phishing emails. Combined word embedded and Neural Bag-of-ngrams enable for extracting syntactic and semantic similarity of emails. DL techniques [14] enable for extracting the abstract and optimum feature representations and fully connected (FC) layer with nonlinear activation function to classifier. According to an enhanced recurrent CNN (RCNN) technique with multilevel vectors and attention process, Fang et al. [15] presented a novel phishing email recognition method called THEMIS that is utilized for modeling email at the word level, email header, email body, and character level concurrently. For evaluating the efficacy of THEMIS, it utilizes an unbalanced data set which is realistic ratios of phishing and legitimate email.

Bagui et al. [16] implemented deep semantic analysis, and ML and DL approach, for capturing inherent features of emails text, and classifying email as phishing/non-phishing. Zamir et al. [17] presented a feature-centric framework (FSEDM) dependent upon current and novel features of emails dataset that is removed after pre-processed. Then, varied supervised learning approaches are executed on the presented feature from conjunction with feature selection (FS) approaches namely gain ratio, information gain, and Relief-F to rank one of the noticeable features and classify the emails to spam/ham (not spam).

3 The Proposed Model

In this article, a new BBODL-PEDC technique has been developed for Phishing Email detection and classification, which effectively distinguished the emails into legitimate and phishing. The BBODL-PEDC model involves a series of subprocesses namely pre-processing, TF-IDF vectorizer, DBN based classification, and BBO based hyperparameter optimization.

3.1 Pre-Processing

Primarily, cleaning of data is performed including the removal of unwanted words as well as characters. Once the data is cleaned, the email data get pre-processed as follows [18].

- Body text extraction
- White space elimination via text parsing
- Convert every character into lowercase and remove non-alphanumeric characters

The BBODL-PEDC model initially performs data pre-processing in three levels namely email cleaning, tokenization, and stop word elimination. Firstly, email cleaning procedure is carried out to remove the unwanted data and non-English characters. Next, tokenization is performed where every email is broken into a set of words, depending upon white spaces. The words obtained are named tokens. Then, stop words which do not carry important data are removed, like conjunction, article, preposition, etc.

3.2 TF-IDF Model

The most commonly utilized measure from the data retrieval is td-idf. These data weight methods are utilized for measuring the probability-weighted count of data in provided documents. During the convention data model, idf is understood as 'the count of data' provided as log of inverse probabilities. By itself, tf-idf has measured that multiples the 2 quantities tf and idf. Thus, term frequency offers evaluation of occurrence probabilities of the term if it can be normalization by the entire frequency from the documents, or document gathering, dependent upon the scope of computation. According to the fundamental equation of data model, the document has been considered that provided disorderly group of terms. Assume $D = \{d_j, \ldots, d_n\}$ be group of documents and $W = \{w_i, \ldots, w_M\}$ be group of various terms limited in D. During this analysis, document D was signified as the corpus of data removed in the tweeter feed but W refers the query term. The parameter N stands for the entire amount of documents but M is the amount of terms. During the adjusting the model, selective of terms w_i in W and selective of documents d_j in D are also regarded.

3.3 DBN Based Email Classification

At this stage, the DBN model is utilized for the classification of emails into phishing and legitimate ones. DBN is a type of probabilistic generative method that establishes the joint distributions amongst input and label information via the learning procedure [19]. Rationally developing the architecture of the DBN models like the amount of layers of the restricted Boltzmann machine (RBM), could efficiently enhance the classifier performance. Determine rational DBN operating parameter includes, including the amount of positive unsupervised learning, the quantity of hidden layers, and the learning rate could significantly enhance the performance of the classifier outcomes. Considering the training efficiency and classification effect of the models, DBN models using a network requirement of 124-250-250-2 is created.

Through comparing the classifier efficacy and setting up a control experiment of the models, it can be defined that the RBM layer fixed by the DBN architecture are 2 layers. RBM is a generative neural network (NN) system. A single RBM is a 2-layer NN comprised of hidden and visible layers. The neuron in all the layers isn't linked, and there is no self-feedback phenomenon from the layer. The neuron in visible and hidden layers are FC in two directions. The energy function among the hidden and visible layers is formulated by:

$$E(v,\ h,\ \omega,\ b_1,\ b_2) = -\sum_i \sum_j \omega_{ij} v_i h_j - \sum_i b_{1i} v_i - \sum_j b_{2j} h_j \tag{1}$$

whereas ω_{ij} indicates the weight connects i and j visible and hidden layers. b_1 and b_2 indicates the biases of visible and hidden layer neurons, correspondingly. Amongst them, the joint likelihood distribution among neurons was estimated by:

$$P(v,\ h,\ \omega,\ b_1,\ b_2) = \frac{1}{Q} e^{-E(v,h,\omega,b_1,b_2)}, \tag{2}$$

$$Q = \sum_v \sum_h e^{-E(v,h,\omega,b_1,b_2)}. \tag{3}$$

assume the input value of DBN architecture is X and resultant value of hidden layer is H, later the weight as well as bias updating equation connect the hidden and output layer neurons as

$$\omega_{ij} = \omega_{ij} + \varepsilon H_j(1 - H_j)X(i)\sum_k \omega_{jk}\delta_k \tag{4}$$

In which δ_k shows the variance among the true type of input values and the actual output value of DBN. ε represent the learning rate of DBN. The classification method of DBN architecture comprises reverse supervised "fine-tuning" learning and forward unsupervised "layer-by-layer initialization" learning. The initial phase of training is named as pretraining method. Fig. 2 demonstrates the framework of DBN.

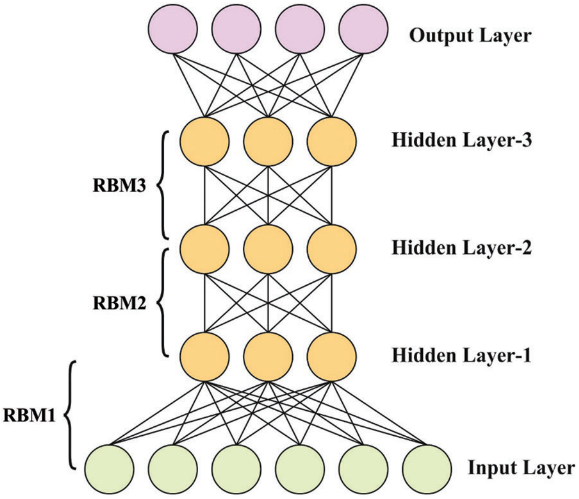

Figure 2: DBN structure

The DBN framework implements forward training via a layer-wise initialization learning model. Through stacking the RBM layer, transfer and map the characteristics data of the input information sequentially. The suggested technique has a Softmax classification on top of RBM. The Softmax classification receives the output data of the top RBM as input data. The Softmax classifiers output the results of forward learning method with the comparison of likelihood distribution. The Softmax classification is created by a multinomial distribution. It is realized that the LR classification confronts generalized induction of various classifiers and is utilized for multiclass classifier problems. The aim is for translating the output data of RBM to a likelihood distribution. The arithmetical depiction of Softmax classification is given below:

$$(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}} \tag{5}$$

whereas y denotes the output vector of RBM. The next phase of training is named the finetuning method. By using the initial phase of pre-training, the RBM layer ensures that the weights of layer reach the optimum perform of feature data of layer and makes the mapping of input data of whole DBN reaches the optimum.

3.4 BBO Based Hyperparameter Optimization

At the final stage, the BBO algorithm [20] is employed for the optimal hyperparameter tuning of the DBN model. Biogeography is the analysis of mutation, migration, speciation, and extinction of species. Biogeography is often supposed that process is compelled equilibrium from the amount of species from the islands. But, the equilibrium in a method is also observed as minimal-energy configuration, thus it can be realized that biogeography was regarded as an optimized procedure. BBO algorithm is a novel evolution technique established to the global optimized. It can be simulated as the immigration as well as emigration of species amongst islands (or habitats) from the searching to further well-suited islands. All the solutions are named as "habitat" (or "island") with habitat suitability index (HSI) and demonstrated as n-dimensional real vectors. A primary individual of habitat vectors is created at arbitrary.

The habitat with maximum HSI was regraded that optimum solution, but the habitat with minimum HSI was regraded that poor solution. The minimum HSI is taken in several novel optimum features procedure the maximum HSI, and this minimum HSI solution has a comparatively higher possibility which developed maximum HSI solution. In BBO, habitat H refers the vector of n (suitable index vector (SIV)) initialize arbitrarily and then executes migration and mutation function for achieving the optimum solutions. A novel candidate solutions are created in the total habitat from population utilizing the migration as well as mutation functions. In BBO, the migration function is to modify present habitat and alter present solution. Migration is a probabilistic function which adjusts habitat X_i. The probability X_i altered has proportional to their immigration rate λ_i, and the source of altered probability in X_j has proportional to rate of emigration μ_j.

The mutation is also a probabilistic function which arbitrarily changes habitat SIV dependent upon the habitat a priori probability of existences. The extremely higher HSI solution and extremely lower HSI solution were correspondingly improbable. Medium HSI solution is comparatively probable. The mutation rate m has formulated as:

$$m = m_{\max}\left(\frac{1 - P_s}{P_{\max}}\right), \tag{6}$$

where $m_{\max}$ implies the adjustable parameters. Moreover, the mutation function deals with improving the population diversity Mutation. Fig. 3 depicts the process flow of BBO technique.

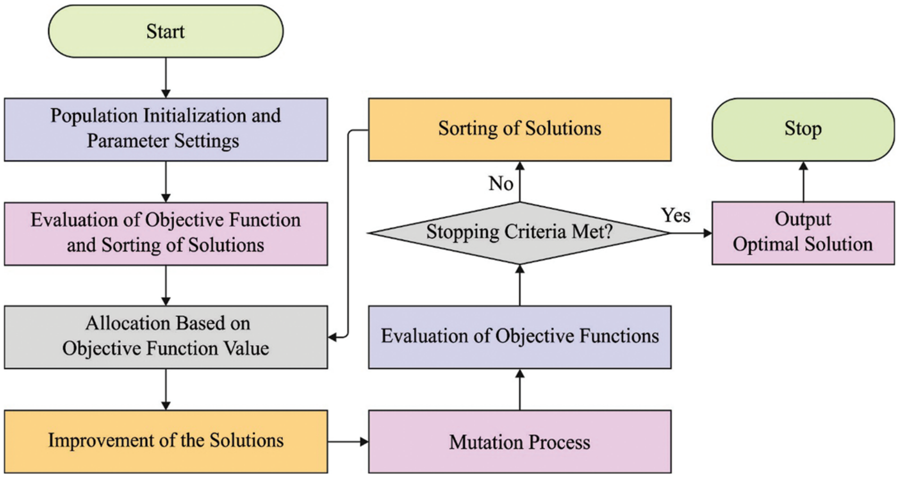
Start
Population Initialization and Parameter Settings
Evaluation of Objective Function and Sorting of Solutions
Allocation Based on Objective Function Value
Improvement of the Solutions
Mutation Process
Evaluation of Objective Functions
Stopping Criteria Met?
No
Sorting of Solutions
Yes
Output Optimal Solution
Stop

Figure 3: Process flow of BBO

4 Experimental Validation

The experimental result analysis of the proposed model is validated using a benchmark CLAIR dataset [21], which contains 3685 phishing and 4894 legitimate Emails.

Fig. 4 highlights the confusion matrices created by the BBODL-PEDC model on distinct runs. The figure reported that the BBODL-PEDC model has effectually categorized the instances into appropriate classes. For instance, with run-1, the BBODL-PEDC model has identified 3660 instances into legitimate classes and 4861 instances into Phishing classes. In line with, with run-1, the BBODL-PEDC model has identified 3660 instances into legitimate class and 4861 instances into Phishing classes. Along with that, with run-3, the BBODL-PEDC model has recognized 3648 instances into legitimate class and 4865 instances into Phishing class. At last, with run-10, the BBODL-PEDC model has identified 3648 instances into legitimate class and 4858 instances into Phishing class.

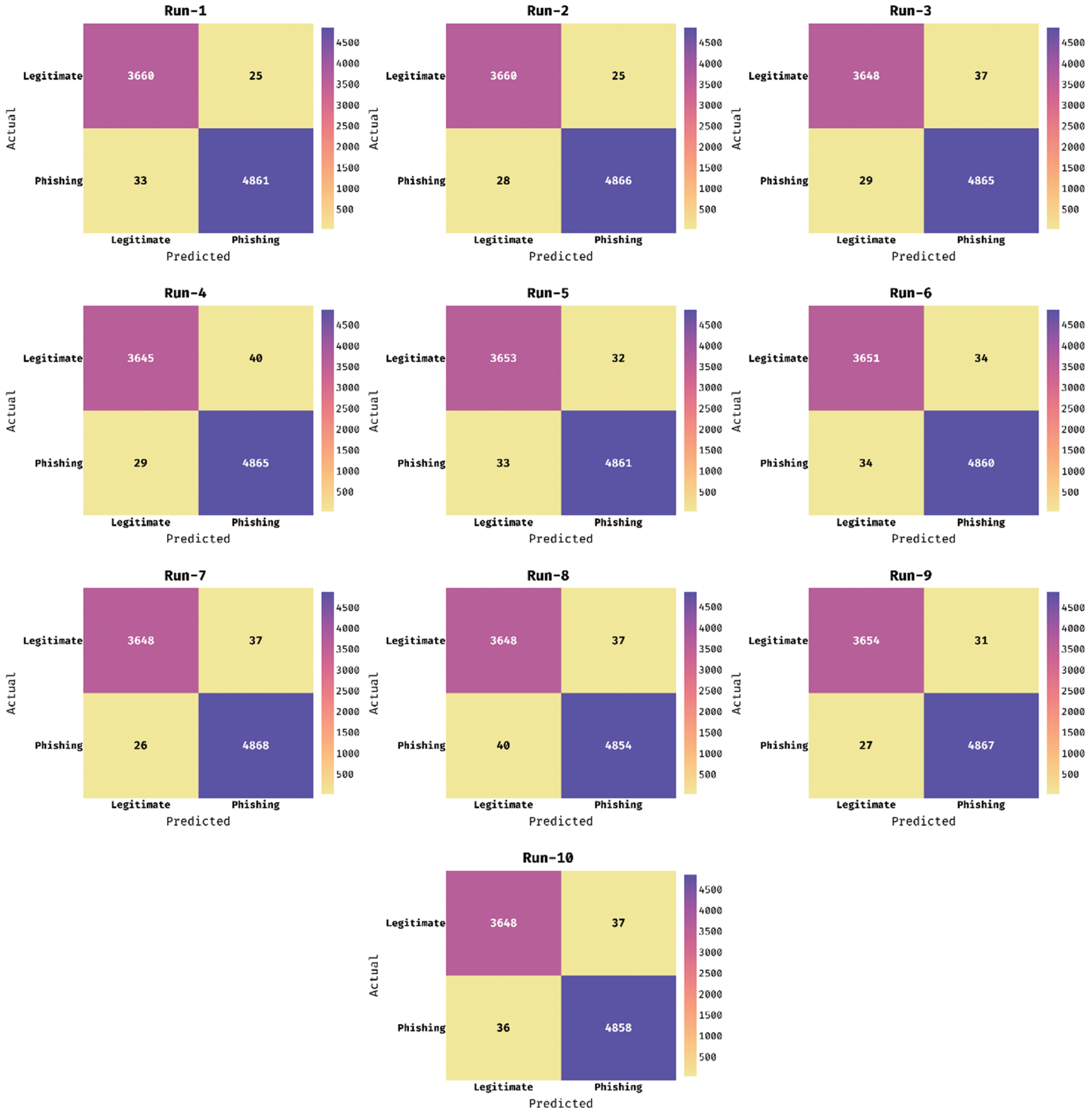

Figure 4: Confusion matrix of BBODL-PEDC technique under different runs

Tab. 1 and Fig. 5 reports the overall classifier outcomes of the BBODL-PEDC model under diverse runs. The table values highlighted that the BBODL-PEDC model has resulted in effectual outcomes under all runs.

For instance, under run-1, the BBODL-PEDC model has attained $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.11%, 99.32%, 99.32%, and 99.21% respectively. In addition, on run-2, the BBODL-PEDC model has obtained $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.24%, 99.32%, 99.38%, and 99.28% respectively. Along with that, on run-3, the BBODL-PEDC model has offered $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.21%, 99.00%, 99.23%, and 99.10% respectively. Followed by, on run-4, the BBODL-PEDC model has reached $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.21%, 98.91%, 99.20%, and 99.06% respectively. In line with, on run-5, the BBODL-PEDC model has exhibited $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.10%, 99.13%, 99.24%, and 99.12% respectively. Finally, on run-10, the BBODL-PEDC model has accomplished $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.02%, 99.00%, 99.15%, and 99.01% respectively.

Table 1: Result analysis of BBODL-PEDC technique interms of various measures under 10 runs

No. of runs	Precision	Recall	Accuracy	F-Score
Run-1	99.11	99.32	99.32	99.21
Run-2	99.24	99.32	99.38	99.28
Run-3	99.21	99.00	99.23	99.10
Run-4	99.21	98.91	99.20	99.06
Run-5	99.10	99.13	99.24	99.12
Run-6	99.08	99.08	99.21	99.08
Run-7	99.29	99.00	99.27	99.14
Run-8	98.92	99.00	99.10	98.96
Run-9	99.27	99.16	99.32	99.21
Run-10	99.02	99.00	99.15	99.01
Average	99.15	99.09	99.24	99.12

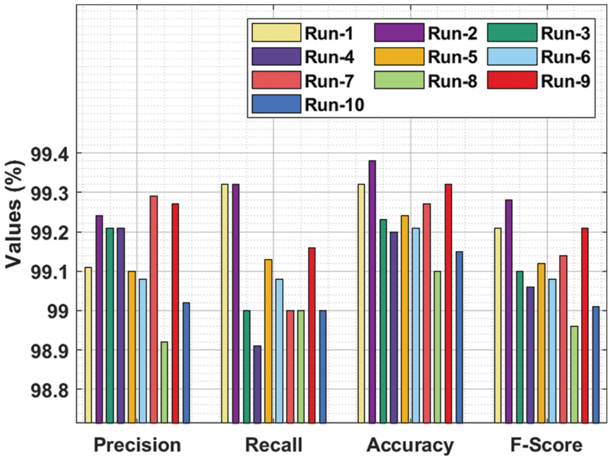

Figure 5: Result analysis of BBODL-PEDC technique under distinct runs

Fig. 6 demonstrates the average classification results of the BBODL-PEDC model on the test dataset applied. The figure reported that the BBODL-PEDC model has accomplished effectual performance with the average $prec_n$, $reca_l$, $accu_y$, and F_{score} of 99.15%, 99.09%, 99.24%, and 99.12% respectively.

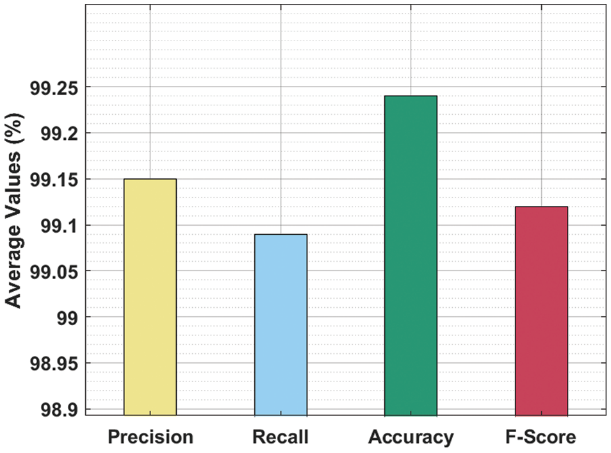

Figure 6: Average analysis of BBODL-PEDC technique

Finally, an extensive comparison study of the BBODL-PEDC model with recent approaches is made in Tab. 2.

Table 2: Comparative analysis of BBODL-PEDC technique with recent approaches

Methods	Precision	Recall	Accuracy	F-Measure
RCNN	96.53	96.74	96.94	97.12
DL	97.00	95.00	99.00	96.00
H-LSTM	97.45	95.98	96.74	96.71
MLA-NLP	95.00	91.00	94.89	95.36
GCN	98.50	98.30	98.20	98.55
BBODL-PEDC	99.15	99.09	99.24	99.12

Fig. 7 exhibits a comparative $prec_n$ examination of the BBODL-PEDC model with existing ones. The figure portrayed that RCNN and machine learning accelerator-natural language processing (MLA-NLP) models have obtained lower performance with rec_n of 96.53% and 95% respectively. In addition, the DL and hierarchical long short term memory (H-LSTM) models have attained moderately reduced $prec_n$ values of 97% and 97.45% respectively. Along with that, the graph convolutional network (GCN) model has resulted in competitive $prec_n$ of 98.50%. However, the BBODL-PEDC model has outperformed the other methods with $prec_n$ of 99.15%.

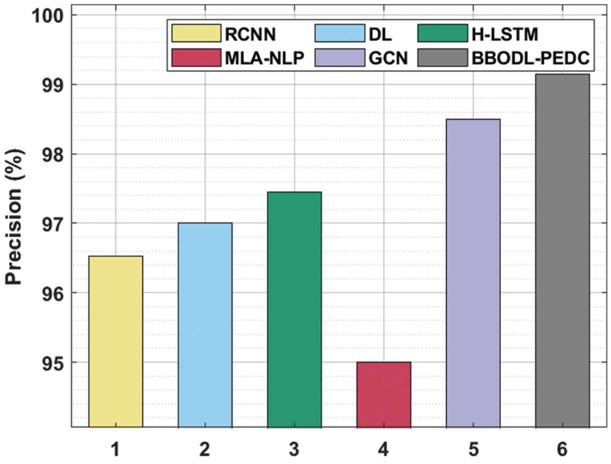

Figure 7: *$Prec_n$* analysis of BBODL-PEDC technique with recent algorithms

Fig. 8 displays a relative $reca_l$ inspection of the BBODL-PEDC model with existing ones. The figure portrayed that the DL, H-LSTM and MLA-NLP models have gotten lower performance with $reca_l$ of 95%, 95.98%, and 91% respectively. Furthermore, the RCNN model has reached reasonably reduced $reca_l$ value of 96.74%. Also, the GCN model has resulted in competitive $reca_l$ of 98.30%. However, the BBODL-PEDC model has outpaced the other methods with $reca_l$ of 99.09%.

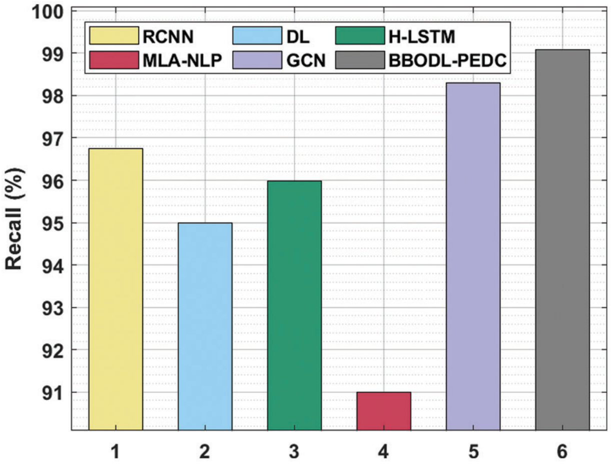

Figure 8: *$Reca_l$* analysis of BBODL-PEDC technique with recent algorithms

Fig. 9 exhibits a comparative $accu_y$ examination of the BBODL-PEDC model with existing ones. The figure portrayed that the RCNN and MLA-NLP models have obtained lower performance with $accu_y$ of 96.94% and 94.89% respectively. In addition, the DL and H-LSTM models have attained moderately reduced $accu_y$ values of 99% and 96.74% respectively. Moreover, the GCN model has resulted in competitive $accu_y$ of 98.20%. However, the BBODL-PEDC model has outpaced the other methods with $accu_y$ of 99.24%.

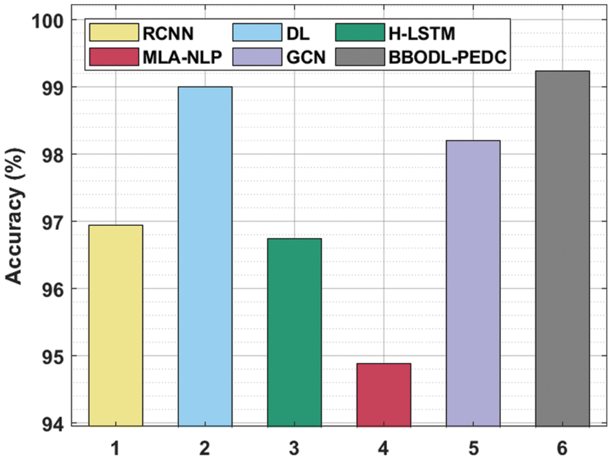

Figure 9: Acc_y analysis of BBODL-PEDC technique with recent algorithms

Fig. 10 reveals a comparative F_{score} inspection of the BBODL-PEDC model with existing ones. The figure portrayed that the DL and MLA-NLP models have found poor results with F_{score} of 96.00% and 95.36% respectively. Additionally, the RCNN and H-LSTM models have accomplished moderately reduced F_{score} values of 97.12% and 96.71% respectively. Besides, the GCN model has reached near optimal F_{score} of 98.55%. However, the BBODL-PEDC model has outperformed the other methods with F_{score} of 99.12%.

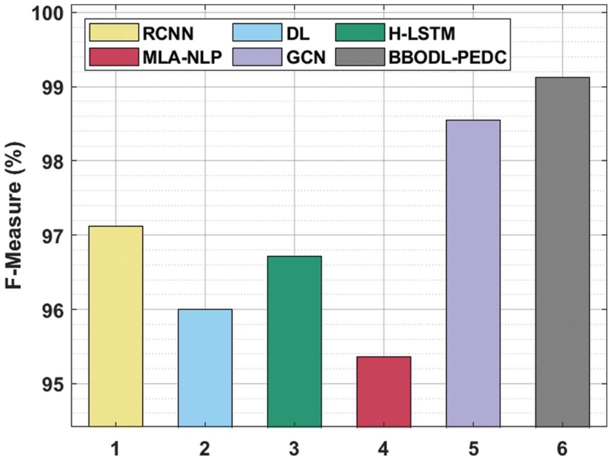

Figure 10: F_{score} analysis of BBODL-PEDC technique with recent algorithms

After examining the above mentioned tables and figures, it is evident that the BBODL-PEDC model has shown effective results over the other methods.

5 Conclusion

In this article, a new BBODL-PEDC technique has been developed for Phishing Email detection and classification, which effectively distinguished the emails into legitimate and phishing. The BBODL-PEDC model involves a series of subprocesses namely pre-processing, TF-IDF vectorizer, DBN based classification, and BBO based hyperparameter optimization. The efficacy of the DBN model can be boosted by the BBO based hyperparameter tuning process. The performance validation of the BBODL-PEDC model can be performed using benchmark dataset and the results are assessed under several dimensions. The extensive comparative studies reported the superior outcomes of the BBODL-PEDC model over the recent approaches. In future, advanced DL models with hybrid metaheuristic optimization algorithms can be designed for phishing email detection.

Acknowledgement: The authors would like to acknowledge the support provided by AlMaarefa University while conducting this research work. The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

Funding Statement: This research was supported by the Researchers Supporting Program (TUMA-Project-2021–27) Almaarefa University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. Qabajeh, F. Thabtah and F. Chiclana, “A recent review of conventional *vs.* automated cybersecurity anti-phishing techniques,” *Computer Science Review*, vol. 29, pp. 44–55, 2018.

[2] G. Sonowal, “Phishing email detection based on binary search feature selection,” *SN Computer Science*, vol. 1, no. 4, pp. 191, 2020.

[3] M. Butavicius, K. Parsons, M. Lillie, A. McCormac, M. Pattinson *et al.*, “When believing in technology leads to poor cyber security: Development of a trust in technical controls scale,” *Computers & Security*, vol. 98, pp. 102020, 2020.

[4] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He, X. Chen, “TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2021. https://doi.org/10.1109/TITS.2021.3130403.

[5] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, “RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring,” *Applied Intelligence*, pp. 1–16, 2021. https://doi.org/10.1007/s10489-021-02893-3.

[6] N. B. Harikrishnan, R. Vinayakumar and K. P. Soman, “A machine learning approach towards phishing email detection,” in *Proc. of the Anti-Phishing Pilot at ACM Int. Workshop on Security and Privacy Analytics (IWSPA AP)*, Coimbatore, vol. 2013, pp. 455–468, 2018.

[7] G. D. L. T. Parra, P. Rad, K. K. R. Choo and N. Beebe, “Detecting internet of things attacks using distributed deep learning,” *Journal of Network and Computer Applications*, vol. 163, pp. 102662, 2020.

[8] F. Tchakounte, J. C. T. Ngnintedem, I. Damakoa, F. Ahmadou and F. A. K. Fotso, “Crawl-shing: A focused crawler for fetching phishing contents based on graph isomorphism,” *Journal of King Saud University - Computer and Information Sciences*, pp. S1319157821003037, 2021, https://doi.org/10.1016/j.jksuci.2021.11.003.

[9] A. S. Bozkir and M. Aydos, “LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition,” *Computers & Security*, vol. 95, pp. 101855, 2020.

[10] G. Varshney, M. Misra and P. K. Atrey, “A phish detector using lightweight search features,” *Computers & Security*, vol. 62, pp. 213–228, 2016.

[11] I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana *et al.*, “Phishing attacks detection using deep learning approach,” in *2020 Third Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp. 1180–1185, 2020.

[12] C. Opara, B. Wei and Y. Chen, “HTMLPhish: Enabling phishing web page detection by applying deep learning techniques on HTML analysis,” in *2020 Int. Joint Conf. on Neural Networks (IJCNN)*, Glasgow, United Kingdom, pp. 1–8, 2020.

[13] V. Ra, B. G. HBa, A. K. Ma, S. KPa, P. Poornachandran *et al.*, “DeepAnti-PhishNet: Applying deep neural networks for phishing email detection,” in *Proc. of the 1st AntiPhishing Shared Pilot at 4th ACM Int. Workshop on Security and Privacy Analytics (IWSPA 2018)*, Tempe, Arizona, USA, pp. 1–11, 2018.

[14] S. Jeyaraj and T. Raghuveera, “A deep learning based end-to-end system (F-gen) for automated email FAQ generation,” *Expert Systems with Applications*, vol. 187, pp. 115896, 2022.

[15] Y. Fang, C. Zhang, C. Huang, L. Liu and Y. Yang, “Phishing email detection using improved rcnn model with multilevel vectors and attention mechanism,” *IEEE Access*, vol. 7, pp. 56329–56340, 2019.

[16] S. Bagui, D. Nandi, S. Bagui and R. J. White, “Classifying phishing email using machine learning and deep learning,” in *2019 Int. Conf. on Cyber Security and Protection of Digital Services (Cyber Security)*, Oxford, United Kingdom, pp. 1–2, 2019.

[17] A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal and A. U. Akram, “A Feature-centric spam email detection model using diverse supervised machine learning algorithms,” *the Electronic Library*, vol. 38, no. 3, pp. 633–657, 2020.

[18] A. Alhogail and A. Alsabih, “Applying machine learning and natural language processing to detect phishing email,” *Computers & Security*, vol. 110, pp. 102414, 2021.

[19] K. Zhang, C. Hu and H. Yu, “Remote sensing image land classification based on deep learning,*” Scientific Programming*, vol. 2021, pp. 1–12, 2021.

[20] H. Duan, W. Zhao, G. Wang and X. Feng, “Test-sheet composition using analytic hierarchy process and hybrid metaheuristic algorithm TS/BBO,” *Mathematical Problems in Engineering*, vol. 2012, pp. 1–22, 2012.

[21] Dataset:D. Radev, “CLAIR collection of fraud email, ACL data and code repository, adcr2008t001,” 2008. http://aclweb.org/aclwiki.