

## A machine learning model for predicting phishing websites

Grace Odette Boussi<sup>1</sup>, Himanshu Gupta<sup>2</sup>, Syed Akhter Hossain<sup>3</sup>

<sup>1</sup>Department of Information Technology, Amity University, Noida, India

<sup>2</sup>Department of Information Technology, Faculty of Cyber Security, Amity University, Noida, India

<sup>3</sup>Department of Computer Science and Engineering, Head of Department, University of Liberal Arts, Dhaka, Bangladesh

### Article Info

#### Article history:

Received Dec 13, 2023

Revised Mar 7, 2024

Accepted Mar 9, 2024

#### Keywords:

Cybercrime

Cybersecurity

Phishing

Prediction

Random forest algorithm

### ABSTRACT

There are various types of cybercrime, and hackers often target specific ones for different reasons, such as financial gain, recognition, or even revenge. Cybercrimes are not restricted by geographical boundaries and can occur globally. The prevalence of specific types of cybercrime can vary from country to country, influenced by factors such as economic conditions, internet usage levels, and overall development. Phishing is a common cybercrime in the financial sector across different countries, with variations in techniques between developed and developing nations. However, the impact, often leading to financial losses, remains consistent. In our analysis, we utilized a dataset featuring 48 attributes from 5,000 phishing webpages and 5,000 legitimate webpages to predict the phishing status of websites. This approach achieved an impressive 98% accuracy.

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Grace Odette Boussi

Department of Information Technology, Amity University

Noida sector 143, 201301, Uttar Pradesh, India

Email: graceboussi@gmail.com

## 1. INTRODUCTION

In today's world, economic relations, business, and markets are progressively moving towards the digital realm [1]. As a result, the risk of cyberattacks is rapidly escalating due to the innovative methods employed by attackers [2]. The widespread use of mobile technology, in conjunction with the onset of the digital age, poses a socio-technical threat to both government entities and the general public [3]. Over the years, cybercrime has evolved into a sophisticated type of criminal activity, making it challenging for victims to detect [4]. This evolution has led to a significant difference between cybercrime today and its early stages. The increasing prevalence of devices, internet-based services, and expanding user base has contributed to a surge in cybercrimes and their sophistication [5]. Despite the implementation of preventive and security measures to mitigate cybercrime, criminals persist in adapting and innovating new methods to circumvent cyber security [6]. Although cybercrime varies across different countries, certain factors such as phishing and data breaches are observed on a global scale. The current digital environment presents both opportunities and dangers. With the ease of a click, criminals can target unsuspecting individuals from anywhere. Using the anonymity provided by the Internet, attackers utilize methods such as phishing, using fraudulent websites to trick victims into revealing sensitive details like account IDs, usernames, and passwords.

Determining whether a webpage is legitimate or a phishing attempt is a challenging problem, as it exploits the vulnerabilities of computer users [7]. Given its dynamic and intricate nature, effective cybersecurity measures are essential for safeguarding both individuals and organizations. Social engineering techniques are employed by attackers to exploit the carelessness and vulnerabilities of individuals in order to intercept sensitive data [8]. Phishing remains a significant concern in a rapidly changing world [9]. It is a form of social web-engineering attack in the online realm, where criminals illicitly obtain valuable data or

information from unsuspecting or uninformed internet users [10]. These attacks are one of the most prevalent methods employed by attackers, and their consequences can include financial losses, reputational damage, and identity theft [11].

With the evolving nature of phishing emails, more sophisticated approaches are required that leverage all the characteristics of emails to enhance the detection capabilities of machine learning and deep learning classifiers [12]. To counter the rising cyber threats, nations have been formulating resilient cybersecurity initiatives and enacting legislation to counter cybercrime, aiming to shield themselves from digital risks. The private sector has been pivotal in crafting inventive cybersecurity solutions, spanning from antivirus programs to specialized software for fraud prevention.

The effectiveness of phishing emails lies in their ability to exploit human emotions, inducing a sense of urgency that compels recipients to take immediate action. This often leads to financial and data losses. Therefore, solely relying on human detection of phishing attempts is insufficient, and more effective automatic detection mechanisms are necessary [13].

Significance and primary contribution of the proposed model: i) The study conducts an empirical comparison of three machine learning algorithms for phishing detection to understand their individual strengths and weaknesses. It analyzes the performance of logistic regression (LR), support vector machine (SVM), and random forest (RF) models in identifying and classifying phishing websites; and ii) Training data using RF to detect phishing websites: We utilized the RF machine learning model to train their data, enabling the identification of whether a website is a phishing one or not. Subsequently, the model yielded a significantly improved accuracy percentage compared to existing methods. This underscores the efficacy and potential of the RF model in addressing the challenges associated with phishing detection.

Our contributions offer valuable insights into the utilization of machine learning techniques for addressing cybercrime. Particularly in the context of detecting and combating phishing activities. The empirical evaluation and analysis of various machine learning algorithms, along with the utilization of the RF model, signify significant progress in enhancing the accuracy and efficiency of cybercrime detection and mitigation efforts.

## 2. LITERATURE REVIEW

Machine learning and modern artificial intelligence (AI) methods have been effectively utilized in various practical applications [14], [15]. Various authors have made noteworthy contributions to the realm of forecasting phishing websites and fortifying against cybercrime. Our work has been shaped by the research conducted by these individuals. The study [16] constructs a knowledge graph called RCTI by combining cybersecurity threat intelligence (CTI) with management security requirements (SR) data. Their innovative  $E(n)$ -equivariant graph neural network (EGNN) model, which is based on GNN, proficiently propagates edge information across the heterogeneous graph. The EGNN model demonstrates top-tier performance in forecasting new insights within the RCTI graph. Additionally, they leverage the EGNN model to forecast new connections within the graph, resulting in a high connectivity rate between CTI and SR entities. The study [17] discusses defense mechanisms against cyber-attacks and presents a threat model for machine learning (ML) security mechanisms in cyber systems. They evaluate the efficacy of machine learning models when subjected to diverse machine learning attacks in cyber-physical systems, offering valuable insights into the effectiveness of distinct security measures. Catal *et al.* [18] presents a comprehensive review of machine learning model life cycles, covering approaches, data sources, feature selection techniques, DL algorithms, evaluation parameters, and validation approaches. They also address the obstacles encountered in the field and put forward potential solutions. Desolda *et al.* [19] focuses on the role of human factors in phishing attacks and presents human factors-based solutions to reduce phishing attacks. Abdillah *et al.* [20] describes common phishing attack vectors, data sources, and identification methods used to mitigate phishing attacks and Das *et al.* [21] delves into the technical and individual attributes of phishing attacks, motivations behind them, and user characteristics. Benavides *et al.* [22] review deep learning algorithms for phishing mitigation, while Arshad *et al.* [23] presents a literature review of phishing and anti-phishing techniques, studies [24] and [25] focus on using natural language processing (NLP) techniques for detecting phishing emails and websites, respectively.

The hybrid features accurately portray emails by merging their content and textual attributes. Additionally, Alani and Tawfik [26] conducts a systematic literature survey comparing various phishing detection approaches, and Nagunwa *et al.* [27] proposes a machine learning-based approach for detecting phishing websites using a novel set of features. To enhance efficiency in anti-phishing techniques, Bahaghighat *et al.* [28] presents an improved predictive model based on machine learning, utilizing six different algorithms and Warraich and Morsi [29] focuses on cyberattacks related to fast-charging stations and introduces a machine learning-based approach for early detection.

For malware detection, Ojewumi *et al.* [30] implements a rule-based approach for phishing detection using machine learning models. They employ three models trained on a dataset comprising fourteen features, demonstrating that the RF model delivers the most favorable performance. Furthermore, the article examines botnets as a noteworthy cybersecurity menace and recommends an ensemble classifier algorithm with a stacking process (ECASP) for identifying and addressing bot attacks through effective feature selection. Kankrale [31] utilized machine learning to identify phishing attacks using a dataset of 1,353 safe website URLs, achieving 90% accuracy with various classifiers. Chiew *et al.* [32] employed machine learning to detect phishing websites, achieving 94.6% accuracy with the RF algorithm.

Yang *et al.* [33] introduced the dynamic category decision algorithm (DCDA) for phishing website detection, integrating deep learning with convolutional neural network-long short-term memory (CNN-LSTM) and XGBoost. Alqahtani [34] developed a method called phishing websites classification using association classification (PWCAC). Ali and Ahmed [35] combined evolutionary and neural network methods for phishing detection. Zamir *et al.* [36] presented a framework for detecting phishing websites using stacking and diverse feature selection methods.

### 3. PROPOSED MODEL

The danger posed by phishing attacks is increasing at a rapid pace, inflicting significant harm by preying on unsuspecting users [37]–[42]. Therefore, our study introduces a novel method to identify and differentiate whether a website is engaging in phishing. Phishing, a pervasive form of cybercrime, involves the use of deceptive websites to deceive users into downloading malware or divulging sensitive personal information to attackers [43].

We utilized a dataset from Kaggle and employed the random forest algorithm to predict phishing threats and Figure 1 illustrates the fundamental operational concept underlying our research endeavor. Prior to model training, we conducted an analysis by splitting the data into three parts to identify uninformative features. Spearman correlation was used to assess variable relationships, enabling the removal of redundant or uninformative features, optimizing the phishing detection process by considering only relevant features.

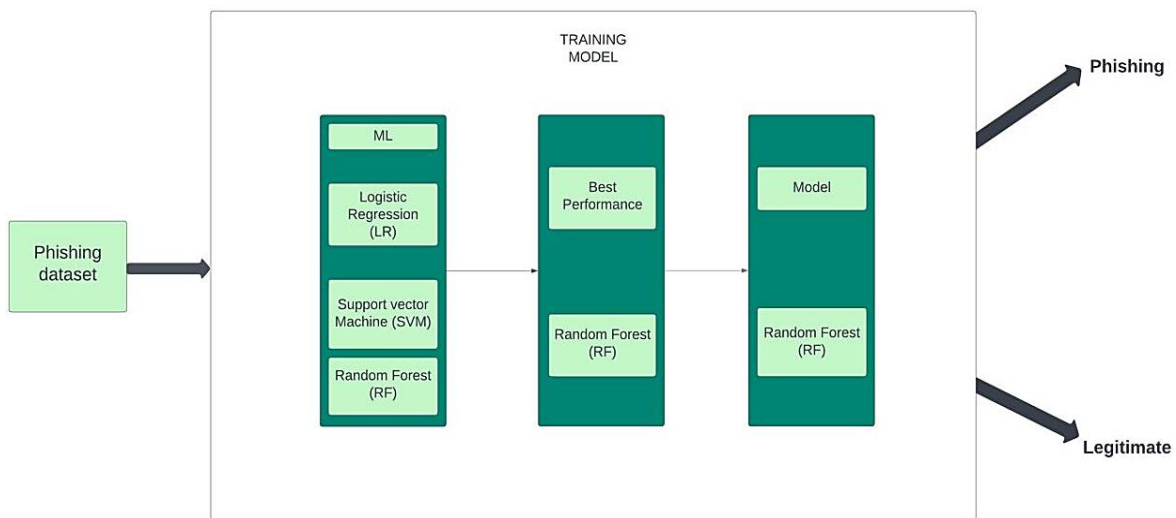


Figure 1. Working principal of proposed model

Our study involved testing websites based on a dataset containing thirty-nine predetermined features to gain in-depth insights into the traits that can impact their classification as phishing sites. Additionally, the study utilized three different ML algorithms to accurately classify the websites and determine the most effective approach for this task. By systematically testing the websites using a diverse dataset and employing multiple machine learning algorithms, the study aims to provide valuable insights into the effectiveness of these approaches in detecting and differentiating phishing websites, ultimately contributing to the advancement of cybersecurity measures. This comprehensive approach underscores the importance of leveraging advanced technology and robust methodologies to address the evolving challenges posed by cybercrime.

We evaluated the performance of three classifiers: LR, SVM, and RF. The LR, is a supervised machine learning method used to predict discrete output classes (binary in this very case) [44]. It relies on various hypothesis functions to forecast binary-value outputs. This paper specifically considers the sigmoid function as a hypothesis function, which is expressed as (1).

$$h_w(x^{(i)}) = \frac{1}{1+e^{-\sum_{j=0}^n w_j x^{(i)}_j}} \quad (1)$$

SVM is a machine learning algorithm used for addressing classification and regression problems [45]. It relies on a hyperplane classifier that separates and maximizes the margin between distinct classes. For a given dataset  $D$ , denoted as  $\{(x_1, y_2), (x_2, y_2), \dots, (x_n, y_n)\}$ , where  $(x_1, y_2)$  represents the labelled data mapping for training, SVM aims to find the optimal decision boundary to separate these classes using a hyperplane, denoted as  $h(x)$ . This decision boundary is designed to effectively distinguish between the two classes, typically represented as +1 for 'phishing websites and -1 for 'legitimate websites.

$$h(x) = \text{sign}(W * X + b) = \text{sign}(\sum_{i=1}^N a_i y_i(x_i, x) + b) \quad (2)$$

Using SVM, we fit the model to the data by minimizing the following function with slack variables:

$$\frac{1}{2} \|W\|^2 + C \sum_{i=1}^N \xi_i \quad (3)$$

Subject to:

$$Y_i (W \cdot X_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall_i \quad (4)$$

For a RF model consisting of  $T$  decision trees, the prediction can be represented, given an input feature vector  $x$  with the predicted class  $\hat{y}$ , as (5):

$$\hat{y} = \text{mode}(f_t(X)) \quad (5)$$

where  $f_t(X)$  is the prediction of the  $t^{\text{th}}$  decision tree. For regression tasks, the predicted value  $\hat{y}$  is computed as (6):

$$\hat{y} = \frac{1}{T} \sum_{i=1}^T f_t(X) \quad (6)$$

Therefore, we can say that the novel aspects of our work lie in the comprehensive and strategic approach to feature analysis, selection, and model training, as well as the demonstration of superior performance in phishing detection compared to existing research. These aspects collectively contribute to the advancement of phishing website detection and classification.

#### 4. INITIALIZING MACHINE LEARNING MODELS FOR THE TRAINING AND SELECTING

This study provides valuable insights into using machine learning techniques to address cybercrime, specifically in detecting and combating phishing activities. While previous studies have examined the significance of machine learning in detecting phishing attacks, they have not explicitly discussed how it influences the evaluation and analysis of different machine learning algorithms to improve the accuracy and efficiency of cybercrime detection and mitigation. We found that a deep understanding and analyst of data correlates with the performance of the algorithm. The proposed method in this study tended to have an inordinately higher proportion in the comprehensive and strategic approach to feature analysis, selection, and model training, as well as the demonstration of superior performance in phishing detection compared to existing research as these aspects collectively contribute to the advancement of phishing website detection and classification.

##### 4.1. Dataset

The initial analysis involved breaking down the data into three parts for a thorough examination of features relevant to detecting suspicious websites. Figures 2, 3, and 4 depict the three segments of our analysis. Figure 2 illustrates the first segmentation, encompassing 0-15 features, while Figure 3 illustrates the second segmentation, covering 15-30 features. Finally, Figure 4 represents the third segmentation, comprising 30-50 features. Spearman Correlation was employed to evaluate the connections between

variables and detect redundant or uninformative features. This method was pivotal in refining the phishing detection process to prioritize only pertinent and meaningful features. By pinpointing these correlated features, we can improve the accuracy and efficacy of our algorithm in discerning between authentic and deceptive websites. Figure 3 illustrates the second segmentation, covering 15-30 features. Figure 4 represents the third segmentation, comprising 30-50 features.

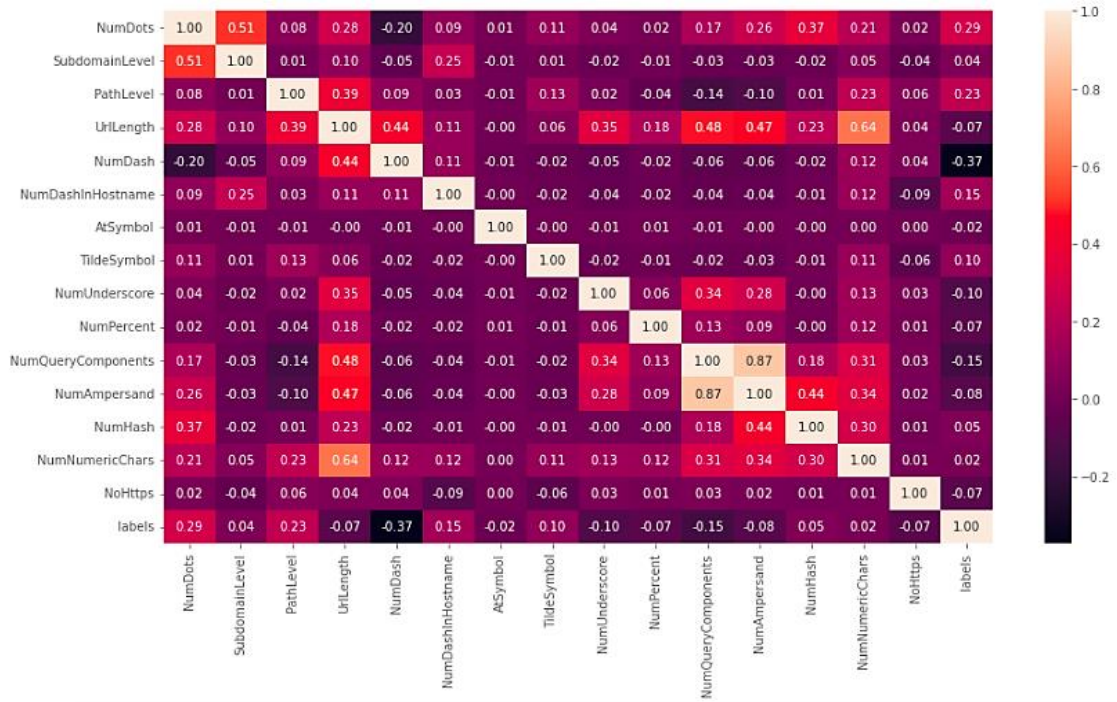


Figure 2. 1<sup>st</sup> segmentation

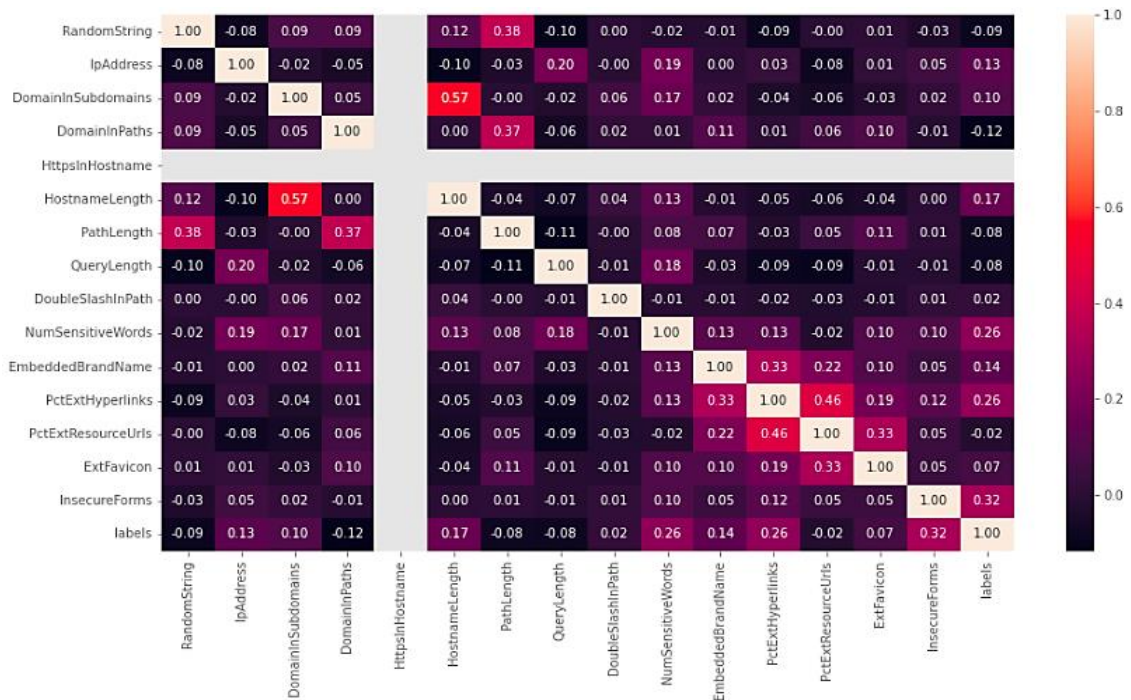


Figure 3. 2<sup>nd</sup> segmentation

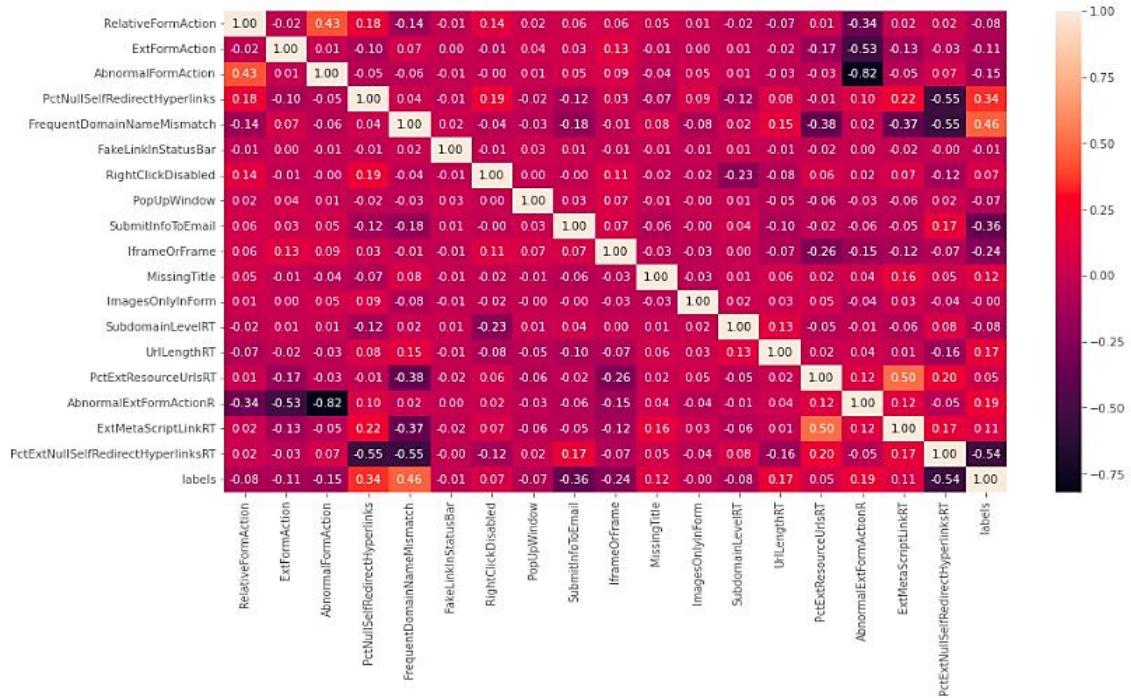


Figure 4. 3<sup>rd</sup> segmentation

#### 4.2. Visualize logistic regression performance

Our analysis of the logistic regression performance using the wrapper method involved visualizing the results. From the visualization, we observed that the regression model achieved the best performance when utilizing 39 features. Based on this insight, we decided to select these 39 features for our final model training. In order to determine the optimum number of features, we examined the plot and identified a region where all performance metrics exhibited favorable results. This region, as depicted in Figure 5, indicated that using 39 features would lead to improved performance across all metrics, resulting in a well-balanced and accurate model.

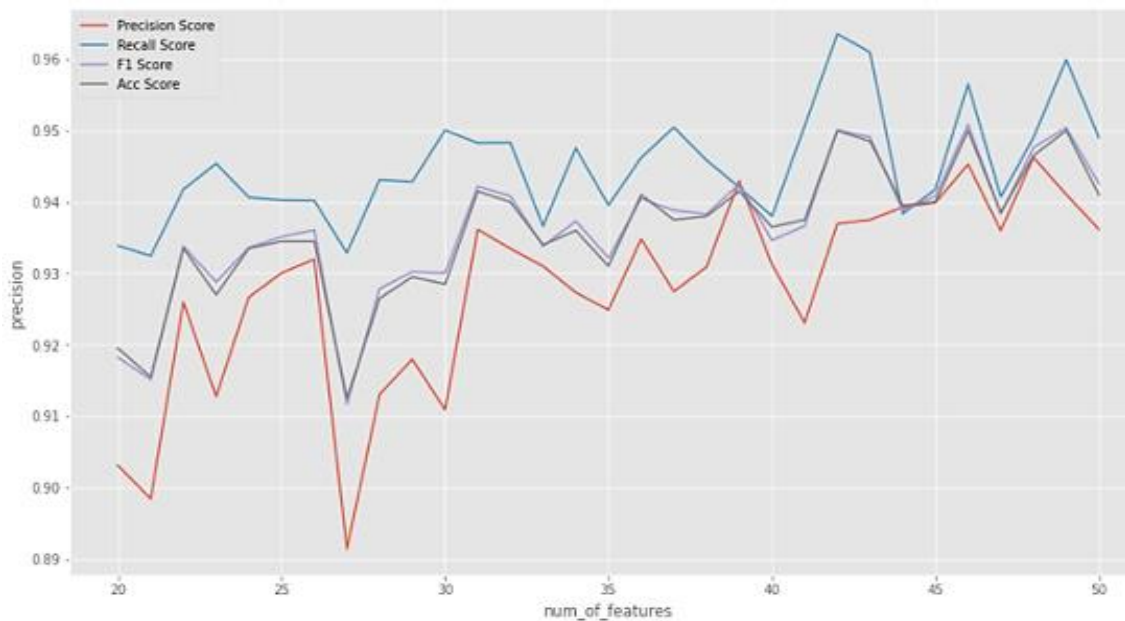


Figure 5. Logistic regression

Figure 5 illustrates the performance of the logistic regression model, serving as a foundational basis for the final model. The analysis of the results in this figure informs decisions about the model's effectiveness and suitability for the specific task. The insights gained from Figure 5 are crucial in shaping and optimizing the final model for improved performance.

In Figure 6, we employed the SVM algorithm along with the wrapper method for feature selection. Similar to our prior analysis with logistic regression, our goal was to choose the most suitable features using this method and assess the performance of SVM. However, it was noted that SVM did not surpass linear regression in terms of performance.

Following the training of the random forest model, we proceeded to predict on the test dataset. Remarkably, the model demonstrated an accuracy of 98% during this phase. This notable accuracy indicates that the random forest model has adeptly discerned patterns and relationships within the data, resulting in precise predictions on previously unseen test data. Our study shows that RF outperforms LR and SVM in phishing detection and Figure 7 illustrates the model's performance. Future research could investigate integrating LR with RF to create an ensemble model, combining the strengths of both algorithms for an overall enhanced predictive performance.

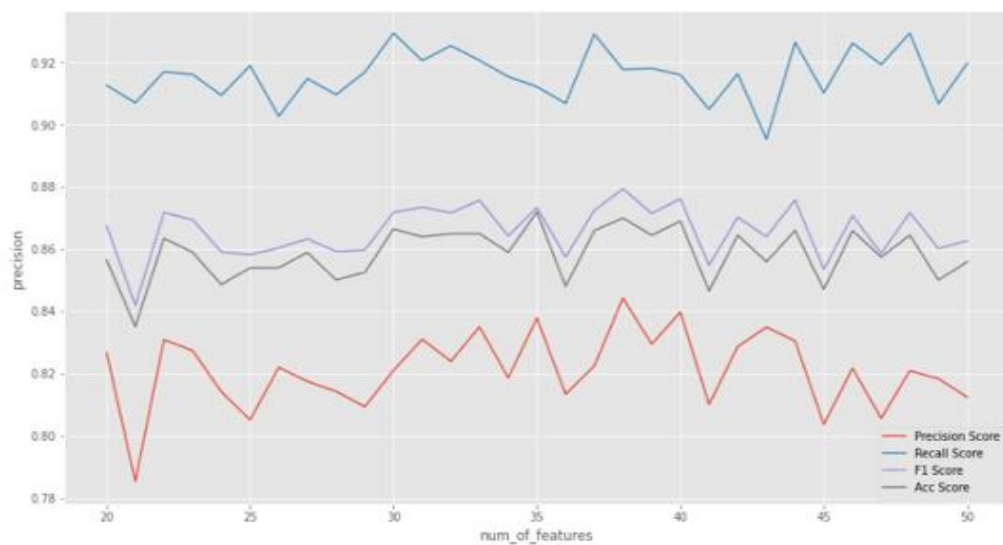


Figure 6. Support vector machine

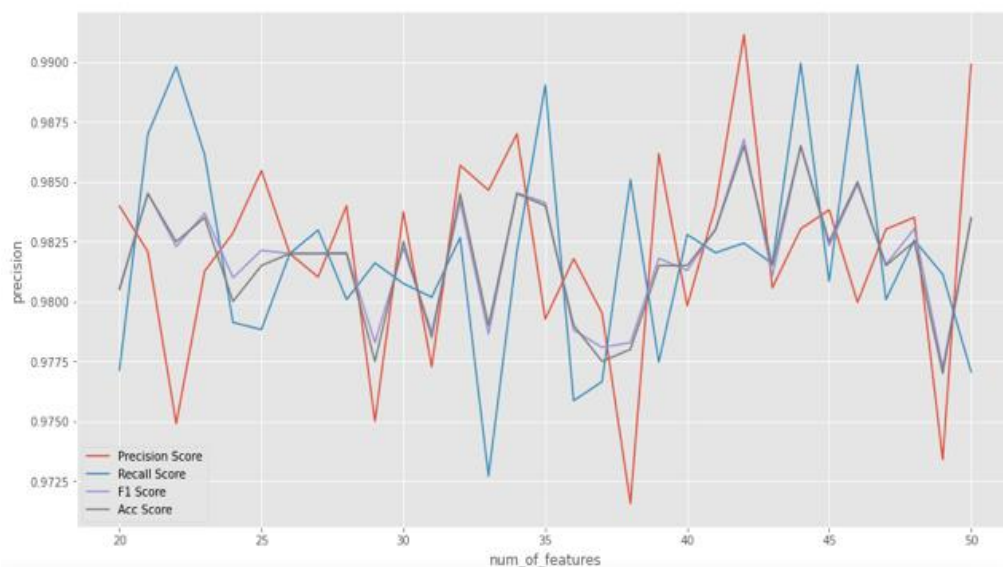


Figure 7. Random forest model

## 5. RESULT AND DISCUSSION

The model's performance was assessed using four primary metrics: accuracy, precision, recall, and F1-Score. Accuracy measures overall correctness, precision evaluates accurate positive predictions, recall measures the model's ability to identify positive instances, and F1-score combines precision and recall for a balanced assessment. These metrics offer a comprehensive evaluation of the model's ability to classify websites as legitimate or phishing. The confusion matrix was used for further performance assessment, capturing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{F1-Score} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (10)$$

The model has demonstrated impressive performance, achieving a high accuracy of 98%, as well as strong precision and recall scores. This indicates the model's robust capability to effectively differentiate between legitimate and potentially malicious websites, showcasing its effectiveness in making accurate predictions. Figure 8 showcases the outcomes of our model, presenting metrics such as precision, recall, F1-score, and accuracy.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	999
1	0.98	0.98	0.98	1001
accuracy			0.98	2000
macro avg	0.98	0.98	0.98	2000
weighted avg	0.98	0.98	0.98	2000

Figure 8. Result

### 5.1. Comparison

The table 1 provides a comparison between our study and existing research in the phishing domain, emphasizing precision, recall, F1-score, and accuracy. The results clearly indicate that our work surpasses previous research in all four metrics. We specifically compare our work with several existing studies, including those referenced as [26], [46]–[49]. With an accuracy rate of 98%, our model excels in predicting phishing websites, revealing the superiority of our approach over prior studies in the field. This contribution significantly enhances cybersecurity efforts as the increasing reliance on the internet has led to a rise in cybercrime, which presents a significant challenge for researchers and law enforcement [50].

Table 1. Comparison

Reference	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)
[46]	96.38 %	90.06 %	93.12 %	93.91 %
[47]	94.85 %	97.86 %	96.33 %	95.94 %
[48]	98.28 %	94.56 %	96.38 %	96.76 %
[26]	95.65 %	96.70 %	96.17 %	97.56 %
[49]	98.47 %	96.58 %	97.51 %	97.54 %
Our proposed model	98.20 %	98.69 %	98.45 %	98.45 %

## 6. CONCLUSION

Our study presents a machine learning model designed to predict phishing websites effectively. Recent observations suggest that the technology has provided both organizations and cybercriminals with advanced tools, leading to a shift from physical to cybercrimes. The evolving nature of technology has made



it difficult for people to differentiate between legitimate and fraudulent links, increasing the risk of falling victim to these attacks. In this context, machine learning plays a crucial role, as it can help predict phishing websites with confidence. By utilizing machine learning algorithms, it becomes possible to analyze various features and patterns, empowering users to identify and avoid potentially dangerous phishing links, thereby reducing the risk of data breaches and financial losses. Our research contributes valuable insights into the utilization of machine learning techniques for phishing detection in the financial sector. The empirical comparison of algorithms, emphasis on feature analysis, and the demonstrated superiority of the RF model collectively advance the field of cybersecurity. The study serves as a foundation for future research, and its findings can inform the development of more effective and accurate phishing detection systems. Future research may explore ensemble models combining LR and RF for even more robust predictive performance.





## REFERENCES

- [1] V. Ivanyuk, "Forecasting of digital financial crimes in Russia based on machine learning methods," *Journal of Computer Virology and Hacking Techniques*, May 2023, doi: 10.1007/s11416-023-00480-3.
- [2] M. Aljabri *et al.*, "Detecting malicious URLs using machine learning techniques: review and research directions," *IEEE Access*, vol. 10, pp. 121395–121417, 2022, doi: 10.1109/ACCESS.2022.3222307.
- [3] M. Arshey and K. S. Angel Viji, "Thwarting cyber crime and phishing attacks with machine learning: a study," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2021, pp. 353–357, doi: 10.1109/ICACCS51430.2021.9441925.
- [4] A. Djenna, E. Barka, A. Benchikh, and K. Khadir, "Unmasking cybercrime with artificial-intelligence-driven cybersecurity analytics," *Sensors*, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146302.
- [5] V. Babanina, I. Tkachenko, O. Matiushenko, and M. Krutevych, "Cybercrime: history of formation, current state and ways of counteraction," *Revista Amazonia Investiga*, vol. 10, no. 38, pp. 113–122, Apr. 2021, doi: 10.34069/AI/2021.38.02.10.
- [6] S. Islam, M. M. Haque, and A. N. M. Rezaul Karim, "A rule-based machine learning model for financial fraud detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 759–771, Feb. 2024, doi: 10.11591/ijece.v14i1.pp759-771.
- [7] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, Mar. 2019, doi: 10.1016/j.eswa.2018.09.029.
- [8] C. M. R. da Silva, B. J. T. Fernandes, E. L. Feitosa, and V. C. Garcia, "Piracema.io: a rules-based tree model for phishing prediction," *Expert Systems with Applications*, vol. 191, Apr. 2022, doi: 10.1016/j.eswa.2021.116239.
- [9] S. H. Ahammad *et al.*, "Phishing URL detection using machine learning methods," *Advances in Engineering Software*, vol. 173, Nov. 2022, doi: 10.1016/j.advengsoft.2022.103288.
- [10] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, "AI Meta-learners and extra-trees algorithm for the detection of phishing websites," *IEEE Access*, vol. 8, pp. 142532–142542, 2020, doi: 10.1109/ACCESS.2020.3013699.
- [11] B. Naqvi, K. Perova, A. Farooq, I. Makhdoom, S. Oyedeji, and J. Porras, "Mitigation strategies against the phishing attacks: a systematic literature review," *Computers & Security*, vol. 132, Sep. 2023, doi: 10.1016/j.cose.2023.103387.
- [12] P. Bountakas and C. Xenakis, "HELPHED: hybrid ensemble learning phishing email detection," *Journal of Network and Computer Applications*, vol. 210, Jan. 2023, doi: 10.1016/j.jnca.2022.103545.
- [13] A. Alhogaib and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," *Computers & Security*, vol. 110, Nov. 2021, doi: 10.1016/j.cose.2021.102414.
- [14] S. AlZu'bi, D. Aqel, and M. Lafi, "An intelligent system for blood donation process optimization - smart techniques for minimizing blood wastages," *Cluster Computing*, vol. 25, no. 5, pp. 3617–3627, Oct. 2022, doi: 10.1007/s10586-022-03594-3.
- [15] D. Aqel, S. Al-Zubi, A. Mughaid, and Y. Jararweh, "Extreme learning machine for plant diseases classification: a sustainable approach for smart agriculture," *Cluster Computing*, vol. 25, no. 3, pp. 2007–2020, Jun. 2022, doi: 10.1007/s10586-021-03397-y.
- [16] Y. Zhang *et al.*, "Edge propagation for link prediction in requirement-cyber threat intelligence knowledge graph," *Information Sciences*, vol. 653, Jan. 2024, doi: 10.1016/j.ins.2023.119770.
- [17] J. Singh, M. Wazid, A. K. Das, V. Chamola, and M. Guizani, "Machine learning security attacks and defense approaches for emerging cyber physical applications: a comprehensive survey," *Computer Communications*, vol. 192, pp. 316–331, Aug. 2022, doi: 10.1016/j.comcom.2022.06.012.
- [18] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: a systematic literature review," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1457–1500, Jun. 2022, doi: 10.1007/s10115-022-01672-x.
- [19] G. Desolda, L. S. Ferro, A. Marrella, T. Catarci, and M. F. Costabile, "Human factors in phishing attacks: a systematic literature review," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–35, Nov. 2022, doi: 10.1145/3469886.
- [20] R. Abdillah, Z. Shukur, M. Mohd, and T. M. Z. Murah, "Phishing classification techniques: a systematic literature review," *IEEE Access*, vol. 10, pp. 41574–41591, 2022, doi: 10.1109/ACCESS.2022.3166474.
- [21] S. Das, A. Kim, Z. Tingle, and C. Nippert-Eng, "All about phishing: exploring user research through a systematic literature review," *arXiv preprint arXiv:1908.05897*, 2019.
- [22] E. Benavides, W. Fuertes, S. Sanchez, and M. Sanchez, "Classification of phishing attack solutions by employing deep learning techniques: a systematic literature review," in *Developments and Advances in Defense and Security: Proceedings of MICRADS 2019*, 2020, pp. 51–64.
- [23] A. Arshad, A. U. Rehman, S. Javaid, T. M. Ali, J. A. Sheikh, and M. Azeem, "A systematic literature review on phishing and anti-phishing techniques," *arXiv preprint arXiv:2104.01255*, 2021.
- [24] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/ACCESS.2022.3183083.
- [25] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.004.
- [26] M. M. Alani and H. Tawfik, "PhishNot: a cloud-based machine-learning approach to phishing URL detection," *Computer Networks*, vol. 218, Dec. 2022, doi: 10.1016/j.comnet.2022.109407.




- [27] T. Nagunwa, P. Kearney, and S. Fouad, "A machine learning approach for detecting fast flux phishing hostnames," *Journal of Information Security and Applications*, vol. 65, Mar. 2022, doi: 10.1016/j.jisa.2022.103125.
- [28] M. Bahaghighat, M. Ghasemi, and F. Ozen, "A high-accuracy phishing website detection method based on machine learning," *Journal of Information Security and Applications*, vol. 77, Sep. 2023, doi: 10.1016/j.jisa.2023.103553.
- [29] Z. S. Warraich and W. G. Morsi, "Early detection of cyber-physical attacks on fast charging stations using machine learning considering vehicle-to-grid operation in microgrids," *Sustainable Energy, Grids and Networks*, vol. 34, Jun. 2023, doi: 10.1016/j.segan.2023.101027.
- [30] T. O. Ojewumi, G. O. Ogunleye, B. O. Oguntunde, O. Folorunsho, S. G. Fashoto, and N. Ogbu, "Performance evaluation of machine learning tools for detection of phishing attacks on web pages," *Scientific African*, vol. 16, Jul. 2022, doi: 10.1016/j.sciaf.2022.e01165.
- [31] P. R. Kankrale, "Phishing website detection using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. VI, pp. 3216–3220, Jun. 2021, doi: 10.22214/ijraset.2021.35671.
- [32] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, May 2019, doi: 10.1016/j.ins.2019.01.064.
- [33] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209, 2019, doi: 10.1109/ACCESS.2019.2892066.
- [34] M. Alqahtani, "Phishing websites classification using association classification (PWCAC)," in *2019 International Conference on Computer and Information Sciences (ICICIS)*, Apr. 2019, pp. 1–6, doi: 10.1109/ICICISci.2019.8716444.
- [35] W. Ali and A. A. Ahmed, "Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting," *IET Information Security*, vol. 13, no. 6, pp. 659–669, Nov. 2019, doi: 10.1049/iet-ifs.2019.0006.
- [36] A. Zamir *et al.*, "Phishing web site detection using diverse machine learning algorithms," *The Electronic Library*, vol. 38, no. 1, pp. 65–80, Mar. 2020, doi: 10.1108/EL-05-2019-0118.
- [37] B. A. V and A. K. Koundinya, "Detection of phishing websites using machine learning techniques," *International Journal of Computer Science and Information Security*, vol. 18, no. 7, pp. 1–5, 2020.
- [38] A. A. Akinyelu, "Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques," *Journal of Computer Security*, pp. 1–57, Aug. 2021, doi: 10.3233/JCS-210022.
- [39] S. Anupam and A. K. Kar, "Phishing website detection using support vector machines and nature-inspired optimization algorithms," *Telecommunication Systems*, vol. 76, no. 1, pp. 17–32, Jan. 2021, doi: 10.1007/s11235-020-00739-w.
- [40] A. Almomani *et al.*, "Phishing website detection with semantic features based on machine learning classifiers," *International Journal on Semantic Web and Information Systems*, vol. 18, no. 1, pp. 1–24, Feb. 2022, doi: 10.4018/IJSWIS.297032.
- [41] F. Song, Y. Lei, S. Chen, L. Fan, and Y. Liu, "Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers," *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 5210–5240, Sep. 2021, doi: 10.1002/int.22510.
- [42] A. Awasthi and N. Goel, "Phishing website prediction: a machine learning approach," in *Advances in Intelligent Systems and Computing*, 2021, pp. 143–152.
- [43] Y. Wei and Y. Sekiya, "Sufficiency of ensemble machine learning methods for phishing websites detection," *IEEE Access*, vol. 10, pp. 124103–124113, 2022, doi: 10.1109/ACCESS.2022.3224781.
- [44] F. Salahdine and N. Kaabouch, "Security threats, detection, and countermeasures for physical layer in cognitive radio networks: a survey," *Physical Communication*, vol. 39, Apr. 2020, doi: 10.1016/j.phycom.2020.101001.
- [45] K. Veena, K. Meena, R. Kuppasamy, Y. Teekaraman, R. V. Angadi, and A. R. Thelkar, "Cybercrime: identification and prediction using machine learning techniques," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Aug. 2022, doi: 10.1155/2022/8237421.
- [46] H. Le, Q. Pham, D. Sahoo, and S. C. Hoi, "URLNet: learning a URL representation with deep learning for malicious URL detection," *arXiv preprint arXiv:1802.03162*, 2018.
- [47] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," *Electronics*, vol. 9, no. 9, Sep. 2020, doi: 10.3390/electronics9091514.
- [48] A. Aljofey *et al.*, "An effective detection approach for phishing websites using URL and HTML features," *Scientific Reports*, vol. 12, no. 1, May 2022, doi: 10.1038/s41598-022-10841-5.
- [49] M. M. Elsheh and K. Swayeb, "Phishing website detection using a hybrid approach based on support vector machine and ant colony optimization," in *2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, May 2023, pp. 402–406, doi: 10.1109/MI-STA57575.2023.10169464.
- [50] J. Curtis and G. Oxburgh, "Understanding cybercrime in 'real world' policing and law enforcement," *The Police Journal: Theory, Practice and Principles*, vol. 96, no. 4, pp. 573–592, Dec. 2023, doi: 10.1177/0032258X221107584.

## BIOGRAPHIES OF AUTHORS






**Grace Odette Boussi**     obtained her bachelor of computer application in Haryana, India, in 2016. She then pursued a master of science in networking technology and management at Amity University Noida from 2016 to 2018, where she received the silver medal for her academic achievements. Since 2019, Grace has been pursuing her Ph.D. in Cyber Security at Amity University Noida, India. She can be contacted at email: graceboussi@gmail.com.



**Himanshu Gupta**    is a respected senior faculty member at Amity University in Uttar Pradesh, India. He completed his education at Aligarh Muslim University and has an extensive academic and professional background in information technology. He has published numerous research papers and articles in the field, with his first patent in network security being published in the International Journal of Patents by the Government of India in December 2010. Additionally, he is a member of various prestigious international technical and research organizations and has delivered online lectures to students from 16 African countries. He can be contacted at email: [hgupta@amity.edu](mailto:hgupta@amity.edu).



**Syed Akhter Hossain**    is an esteemed computer scientist, educator, columnist, and technology consultant from Bangladesh. He is currently serving as a professor and the head of the Computer Science and Engineering Department at the University of Liberal Arts Bangladesh. He can be contacted at email: [aktarhossain@daffodilvarsity.edu.bd](mailto:aktarhossain@daffodilvarsity.edu.bd).