

Enhancing phishing URL detection through comprehensive feature selection: a comparative analysis across diverse datasets

Preeti, Priti Sharma

Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak Haryana, India

Article Info

Article history:

Received Mar 14, 2024

Revised Jul 30, 2024

Accepted Aug 5, 2024

Keywords:

Decision tree

Feature selection technique

Phishing attack

Random forest

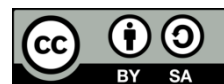
URL classification

XGBoost and MLP

ABSTRACT

Malicious attacks have developed a prominent risk to the safety of online users, with attackers employing increasingly sophisticated systems to deceive unsuspecting victims. This research focuses on the critical aspect of feature selection in optimizing phishing uniform resource locator (URL) detection system. Feature selection boosts machine learning (ML) and deep learning (DL) by picking vital attributes efficiently. This research paper provides a comprehensive examination of feature selection techniques using five diverse datasets. Various methods, including random forest (RF) select from model, SelectKBest with chi-square statistic, principal component analysis (PCA) and recursive feature elimination (RFE), were employed. The experiments, with a particular emphasis on PCA and fourth dataset, revealed that all four models RF, decision trees (DTs), XGBoost, and multilayer perceptron) achieved 100% accuracy in detecting phishing URL attacks. This underscores the efficacy of feature selection methods in enhancing to a deeper understanding of feature selection's role in bolstering the effectiveness of phishing detection system across diverse datasets, highlighting the importance of leveraging techniques such as PCA for optimal results.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Preeti

Department of Computer Science and Applications, Maharshi Dayanand University

Rohtak Haryana, India

Email: miskhokhar121@gmail.com

1. INTRODUCTION

Phishing attacks, utilizing deceptive uniform resource locators (URLs), epitomize sophisticated cyber deception, employing imitation web servers to illicitly acquire sensitive information. This study examined how machine learning (ML) algorithm impact the detection of phishing attacks [1]. Although previous research has looked into traditional detection methods, it has not specifically analyzed how advanced ML techniques can enhance the accuracy and efficiency of phishing prevention. The finding reveals that integrating ML can significantly improve the effectiveness of detecting and preventing phishing attempts.

This research develops a cost-effective phishing detection sensor using deep learning (DL) techniques [2]. Traditional methods rely on user reports, but recent advancements in DL have led to improved detection methods. This paper introduces a lightweight DL algorithm for real-time, energy-efficient phishing detection, demonstrating enhanced accuracy and practical viability on embedded systems. Microsoft's attention-mechanism-based method excels in detection phishing attacks by dynamically weighting URL component, Emphasizing nuanced analysis [3].

This study investigated the effects of using a double deep Q-Network (DDQN) for web phishing detection [4]. While previous research has examined DL techniques for this task, it has not specifically addressed the integration of deep reinforcement learning with an imbalanced classification Markov decision

process (ICMDP). The study demonstrates that the DDQN-based classifier significantly outperforms traditional methods in handling data imbalance and improving detection accuracy.

This research explored XGBoost for phishing detection using 22,000 URLs and 22 features with NLP evaluation. While previous studies examined various techniques, they didn't focus on text-as-image representation's impact. The model achieved 94% accuracy and 91% precision [5]. The paper explores ML for phishing domain detection, developing and comparing models based on support vector machine (SVM), DT, ANN, and random forest (RF). Using the UCI phishing domains dataset, the findings reveal that the RF model outperforms others, proving its superiority over existing solutions [6].

Presents an innovative approach for phishing websites classification using convolutional neural networks (CNN) with URL-based features. The CNNs, employing entropy loss function and ReLU to address vanishing gradient, achieve an 86.5% classification accuracy on a dataset of 1,353 URLs [7]. This survey evaluated multiple ML algorithms for phishing detection, such as SVM, Naïve Bayes (NB), decision tree (DT) and RF [8]. Although previous studies employed these methods, they did not specifically cover newer systems like PhishScore and PhishChecker. The research summarized effective techniques and emphasized updating features to address emerging phishing threats.

Explores modern ML techniques for detecting phishing attacks with high accuracy. Utilizing a Kaggle dataset of over 11,000 websites, the study assesses 30 website features, using neural network, NB, and AdaBoost models. Results show accuracies of 90.23%, 92.97%, and 95.43%, correspondingly [9]. Presented a recognition way utilizing nine lexical features, reaching a 99.57% accuracy utilizing the RF model on the ISCXURL -2016 dataset comprising 11,964 instances [10]. Used ML techniques (XGBoost, DT, logistic regression (LR), RF, and SVM), the research identifies phishing websites. RF achieves 98.90% and 97.87% accuracy on Phish-Tank and UCI datasets [11].

In the conducted study, a pioneering phishing detection method achieved accuracy, surpassing previous models. A dataset of 10,000 malicious URLs and legitimate sites was utilized, and our integrated CNN-based model excelled with a 98.77% accuracy, benefiting from deeper layers and additional features for enhanced performance [12]. Presented emphasizing techniques for prevention rather than mitigation. Providing a general overview, it highlights DL as a key strategy in effective phishing attack detection [13]. Introduced three DL techniques for phishing website identification. The experimental outcomes indicate impressive accuracy rates of 96.8%, 99.2%, and 97.6% for long short-term memory (LSTM), CNN and LSTM-CNN models, correspondingly. The CNN-based system proves superior in phishing detection [14].

Presented a "Phish Derby" competition at U.S. university to gamify phishing security awareness training. Finding highlighted relationship between demographics, personality traits, goal orientation, and phishing detection performance. Insights stress fostering positive cyber behaviors beyond click rates in organizational training cultures [15].

Proposed a novel client-side technique for effortless phishing website identification using a redesigned browser architecture. A RF classification model analyzes 30 URL properties extracted through a rule of extraction framework. The 'embedded phishing detection browser' (EPDB) integrates phishing detection without compromising user experience, achieving a real-time accuracy of 99.36% [16].

This study investigates the efficacy of hybrid LSTM and CNN DL models for fake website URL detection, leveraging the strengths of both approaches [17]. Using two publicly existing datasets, the hybrid model achieved significantly higher accuracies compared to standalone CNN and LSTM models. These conclusions suggest the latent of hybrid DL techniques in mitigating losses from spoofing attacks.

This article presents an experimental study enhancing ML model performance for malicious dataset [18]. It explores hyperparameter optimization, data balancing and feature selection, showing significant accuracy improvement. Combining tuned factors enhances algorithm efficiency, with gradient boosting and extreme gradient boosting achieving high accuracy rates for both datasets. Introduced a hybrid feature-based anti-phishing strategy, achieved 99.17% detection accuracy using XGBoost on client-side URL and hyperlink data [19].

This paper showcases a systematic approach to constructing detection models employing three DL architectures [20]. Utilizing fully connected deep neural networks (DNNs), CNNs and LSTM, it achieved a peak accuracy of 97.37% across four phishing website datasets. Additionally, comparison of optimization algorithms led to accuracy enhancements of 0.1%-1%.

Almomani *et al.* [21] employ 16 ML models with ten features to detect fake webpage from 2 datasets. Gradient boosting classifier (GBC) achieves highest accuracy (97%) while GaussianNB and SGD classifier exhibit lowest accuracy (84% and 81% respectively) among classifiers. Introduces an innovative malicious URL detection technique combining DL and BERT feature extraction. BERT extracts text from URLs, NLP algorithms extract meaningful features, and a CNN method detects phishing URLs. With 96.66% accuracy, it proves efficient in detecting phishing websites' URLs, validated against existing literature [22].

This research focuses on detecting three-stage phishing attacks via content analysis. Input values include URLs, traffic, and web content features [23]. Real phishing cases dataset yields high accuracy

(95.18%) with NN, outperforming SVM (85.45%), and RF (78.89%). ML proves effective for phishing detection. This study introduces a novel method developed by researchers, integrating DL for URL classification with a genetic algorithm for feature selection [24]. Their approach significantly improves recall in URL classification, outperforming recent DL methods with notable accuracy and recall enhancements. This study introduces a supervised learning method for Android malware detection, leveraging a full labeled dataset of over 18,000 samples across 5 categories [25]. Validation against established datasets demonstrates outperformance in specific parameters, contributing to advanced techniques for Android device security amidst evolving threats.

2. SUMMARIZING KEY FINDING

Recent studies highlight significant advancements in phishing detection through ML and DL techniques. Integrating DL models, such as lightweight algorithms and hybrid approaches like LSTM-CNN, has markedly improved accuracy and practical applicability in real-time detection. For instance, RF and XGBoost methods achieved high accuracy rates, with RF reaching up to 99.57% in identifying phishing URLs. The use of advanced techniques like BERT for feature extraction and genetic algorithm for optimizing DL models has future enhanced detection capabilities. These findings demonstrate that modern ML and DL methods significantly outperform traditional approaches, offering robust solutions for preventing and detecting phishing attacks. This underscores the potential of these technologies to address evolving cybersecurity threats effectively.

3. INTERPRETING RESULTS

Our methodology encompassed a judicious selection of algorithms and methodologies aimed at maximizing accuracy and robustness. We leveraged a diverse array of ML and DL, including RF, XGBoost, multi layer perceptron (MLP) and DT. These models were chosen for their versatility and effectiveness in handling classification tasks across various domains. By training these models on five distinct datasets, we aimed to capture a broad spectrum of patterns and intricacies inherent in phishing URLs.

Moreover, feature selection played a pivotal role in enhancing the models' performance. We employed a two-pronged approach, utilizing RF select from model along side a combination of SelectKBest with the Chi-square (χ^2) statistic, recursive feature elimination (RFE) and principal component analysis (PCA). This meticulous feature selection process aimed to distill the most informative features while mitigating the risk of overfitting and improving model interpretability.

To facilitate a comprehensive understanding of our findings, we organized our results into a structured format, as exemplified in Table 1 (in Appendix). This table provided a succinct yet informative overview of the models' performance metrics across different datasets, enabling readers to discern trends and draw insights. Furthermore, we honed in on the analysis of the fourth datasets to showcase the efficacy of our proposed model. Note by, the application of PCA for feature selection of 100%. Such finding underscored the robustness and reliability of our approach in identifying phishing URLs.

In addition to quantitative analysis, we supplemented our findings with visual representations, as depicted Figure 1 and Figure 2. These figures elucidated the comparative performance of the models and highlighted the tangible enhancements observed in both existing and proposed models. Notably, our proposed models exhibited a significant improvement, achieving 100% accuracy and affirming their suitability for real-world deployment.

Figure 1 presents the accuracy of four models: DT, RF, XGBoost, and MLP by employing PCA for feature selection. Each model achieved a perfect accuracy of 100%, highlighting the effectiveness of PCA in improving model performance. This figure demonstrates the robustness of our feature selection method and confirms the high predictive capabilities of the models. The consistently high accuracy across all models underscores the importance of PCA in optimizing feature selection for better ML results.

Figure 2 presents a horizontal bar graph comparing the accuracy of various models by different authors from 2022 to 2024. The graph highlights that the proposed model achieves the highest accuracy at 100.0%, outperforming others. Notably, Gupta *et al.* [19] achieved 99.17% and Ujah-Ogbuagu [17] reached 98.90%. This Chart effectively highlights advancements in model performance and illustrates the variations in accuracy across recent studies.

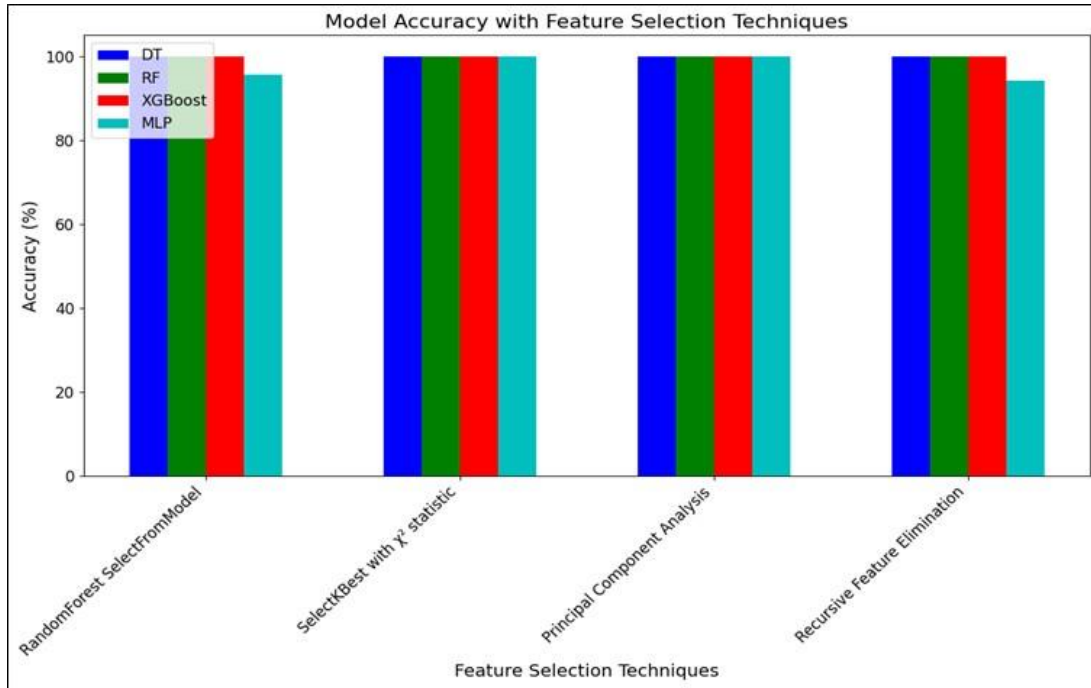


Figure 1. Comparative analysis of model performance across four datasets

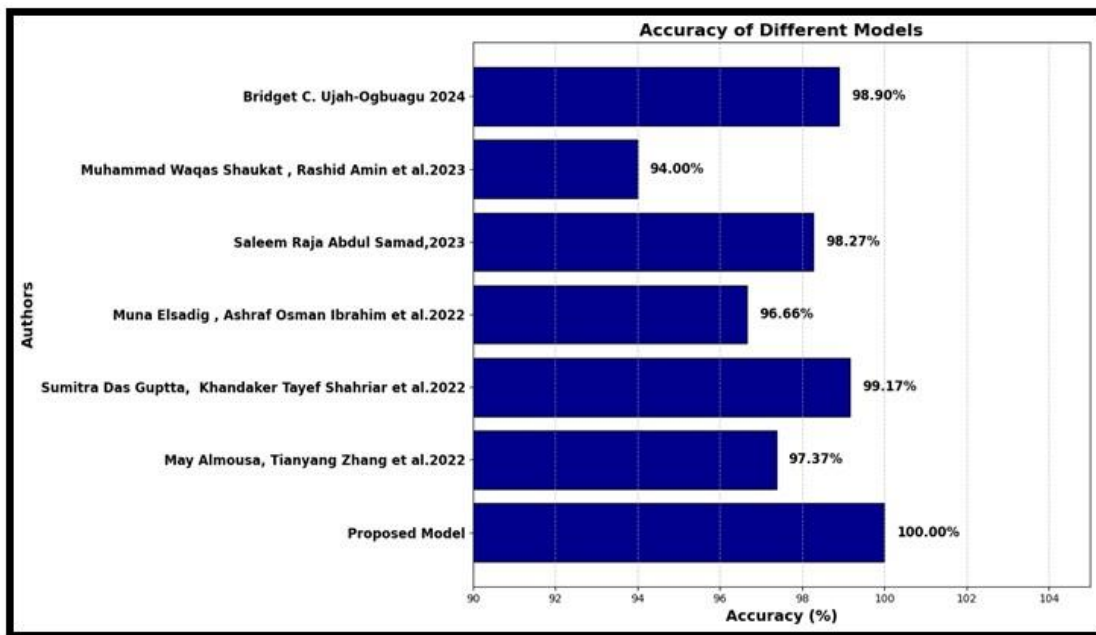


Figure 2. Accuracy comparisons between existing and proposed models for phishing URL attack detection

Crucially, our experimentation was underpinned by the utilization of five multidimensional datasets, ensuring the models' exposure to diverse and representative data samples. This strategic approach bolstered the models' generalization capability. This strategic approach bolstered the models' generalization capability and engendered confidence in their real-world applicability. In essence, our research endeavors aimed to contribute valuable insights into the optimization of ML and DL approaches for detecting phishing URL attacks. By meticulously select models, employing advanced features selection techniques, and conducting rigorous experimentation, we strived to push the boundaries of accuracy and efficacy in combating cyber threats.

4. ADDRESSING LIMITATIONS

While our research demonstrates promising results, several limitations must be acknowledged. Firstly, achieving 100% accuracy across all models may indicate overfitting, particularly with the dataset used. This performance might not generalize to other datasets or real-world scenarios. Secondary, the datasets used may not fully capture the diversity and evolving nature of phishing URLs. Moreover, our study focused on a limited number of feature selection techniques; exploring additional methods could yield further insights. Lastly, the computational efficiency and scalability of the models were not extensively evaluated, which is crucial for real-time browser integration and practical deployment.

5. IMPLICATIONS FOR FUTURE RESEARCH

Building on our finding, future research should focus on validating the models with more diverse and dynamic datasets to ensure robustness and generalizability. Investigating additional feature selection techniques and hybrid approaches could enhance model performance and adaptability. Furthermore, addressing computational efficiency and scalability will be essential for real-time applications and integration into web browsers. Exploring the models' effectiveness against emerging phishing tactic will also be critical. Finally, interdisciplinary collaboration with cybersecurity experts and practitioners can provide practical insights, ensuring the models' efficiency and reliability in real-world environments.

6. CONCLUSION

In this research paper, we conducted a thorough analysis of feature selection techniques across five datasets, employing methods such as RF select from model, Select KBest with chi-square statistic, RFE, PCA. Our experiments, particularly focusing on PCA and the fourth dataset, demonstrated that all four models (DTs, RFs, XGBoost, and MLP) achieved 100% accuracy in detecting phishing URLs attacks. This exceptional performance suggests the robustness of our proposed models. Despite some limitations, including potential overfitting and the need for broader dataset validation, these findings indicate the viability and reliability of our approach for real-life implementation in addressing phishing threats.

APPENDIX

Table 1. Comprehensive analysis of model performance metrics across datasets and efficacy of proposed model on fourth dataset with PCA feature selection (100%)

Datasets	Feature selection techniques	ML and DL models	Accuracy (%)
1 st datasets contain 39656 instances with 14 Multidimensional features	RF select from model feature selection	DT	94.02
		RF	95.62
		XGBoost	94.35
		MLP	73.09
	Select KBest with the chi-square (χ^2) statistic	DT	92.35
		RF	93.02
		XGBoost	92.6
		MLP	89.97
	PCA	DT	93.71
		RF	95.89
		XGBoost	95.02
		MLP	46.25
RFE	DT	94.04	
	RF	95.62	
	XGBoost	94.4	
	MLP	67.52	
2 nd datasets contain 11430 instances with 89 Multidimensional features	RF select from model feature selection	DT	93.03
		RF	95.62
		XGBoost	96.12
		MLP	77.19
	Select KBest with the chi-square (χ^2) statistic	DT	92.00
		RF	92.73
		XGBoost	92.91
		MLP	91.8
	PCA	DT	85.85
		RF	90.95
		XGBoost	90.93
		MLP	70.31
RFE	DT	93.49	
	RF	95.12	
	XGBoost	95.94	
	MLP	76.78	

Table 1. Comprehensive analysis of model performance metrics across datasets and efficacy of proposed model on fourth dataset with PCA feature selection (100%) (Continued)

Datasets	Feature selection techniques	ML and DL models	Accuracy (%)
3 rd datasets contain 88646 instances with 111 multidimensional features	RF select from model feature selection	DT	95.19
		RF	96.78
		XGBoost	96.73
	Select KBest with the chi-square (χ^2) statistic	MLP	95.76
		DT	86.58
		RF	86.58
		XGBoost	86.57
		MLP	86.57
		DT	94.05
	PCA	RF	96.02
		XGBoost	95.78
		MLP	94.33
	RFE	DT	94.75
		RF	96.52
		XGBoost	96.56
MLP		86.98	
DT		100	
RF		100	
4 th datasets contain 10000 instances with 50 multidimensional features	RF select from model feature selection	XGBoost	99.96
		MLP	95.56
		DT	100
	Select KBest with the chi-square (χ^2) statistic	RF	100
		XGBoost	99.96
		MLP	99.86
		DT	100
		RF	100
		XGBoost	100
	PCA	MLP	100
		DT	100
		RF	100
	RFE	XGBoost	100
		MLP	100
		DT	100
RF		100	
XGBoost		99.96	
MLP		94.1	
5 th datasets contain 11055 instances with 31 multidimensional features	RF select from model feature selection	DT	93.48
		RF	93.48
		XGBoost	93.49
	Select KBest with the chi-square (χ^2) statistic	MLP	93.57
		DT	94.15
		RF	94.33
		XGBoost	94.42
		MLP	93.97
		DT	93.91
	PCA	RF	95.41
		XGBoost	95.62
		MLP	94.27
	RFE	DT	95.14
		RF	95.2
		XGBoost	95.11
MLP		94.93	

REFERENCES

[1] V. Bhavsar, A. Kadlak, and S. Sharma, "Study on phishing attacks," *International Journal of Computer Applications*, vol. 182, no. 33, pp. 27–29, Dec. 2018, doi: 10.5120/ijca2018918286.

[2] B. Wei *et al.*, "A deep-learning-driven light-weight phishing detection sensor," *Sensors (Switzerland)*, vol. 19, no. 19, Oct. 2019, doi: 10.3390/s19194258.

[3] M. Alazab and S. Fellow, "Malicious URL detection using deep learning," *TechRxiv*, 2023, doi: 10.36227/techrxiv.11492622.v1.

[4] A. Maci, A. Santorsola, A. Coscia, and A. Iannacone, "Unbalanced web phishing classification through deep reinforcement learning," *Computers*, vol. 12, no. 6, Jun. 2023, doi: 10.3390/computers12060118.

[5] M. W. Shaukat, R. Amin, M. M. A. Muslam, A. H. Alshehri, and J. Xie, "A hybrid approach for alluring ads phishing attack detection using machine learning," *Sensors*, vol. 23, no. 19, Oct. 2023, doi: 10.3390/s23198070.

[6] S. Alnemari and M. Alshammari, "Detecting phishing domains using machine learning," *Applied Sciences (Switzerland)*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084649.




[7] A. D. Kulkarni, "Convolution neural networks for phishing detection convolution neural networks for phishing detection convolution neural networks for phishing detection," [Online]. Available: <http://hdl.handle.net/10950/4224www.ijacsa.thesai.org>

[8] R. Kiruthiga and D. Akila, "Phishing websites detection using machine learning," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 11, pp. 111–114, Sep. 2019, doi: 10.35940/ijrte.B1018.0982S1119.




- [9] D. T. Mosa, M. Y. Shams, A. A. Abohany, E. S. M. El-Kenawy, and M. Thabet, "Machine learning techniques for detecting phishing URL attacks," *Computers, Materials and Continua*, vol. 75, no. 1, pp. 1271–1290, 2023, doi: 10.32604/cmc.2023.036422.
- [10] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.
- [11] T. Choudhary, S. Mhapankar, R. Bhddha, A. Kharuk, and R. Patil, "A machine learning approach for phishing attack detection," *Journal of Artificial Intelligence and Technology*, vol. 3, no. 3, pp. 108–113, Jul. 2023, doi: 10.37965/jait.2023.0197.
- [12] E. A. Aldakheel, M. Zakariah, G. A. Gashgari, F. A. Almarshad, and A. I. A. Alzahrani, "A deep learning-based innovative technique for phishing detection in modern security with uniform resource locators," *Sensors*, vol. 23, no. 9, May 2023, doi: 10.3390/s23094403.
- [13] L. M. Abdulrahman, S. H. Ahmed, Z. N. Rashid, Y. S. Jghaf, T. M. Ghazi, and U. H. Jader, "Web phishing detection using web crawling, cloud infrastructure and deep learning framework," *Journal of Applied Science and Technology Trends*, vol. 4, no. 01, pp. 54–71, Mar. 2023, doi: 10.38094/jastt401144.
- [14] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN," *Electronics (Switzerland)*, vol. 12, no. 1, Jan. 2023, doi: 10.3390/electronics12010232.
- [15] M. Canham, C. Posey, and M. Constantino, "Phish derby: shoring the human shield through gamified phishing attacks," *Front Educ (Lausanne)*, vol. 6, Jan. 2022, doi: 10.3389/educ.2021.807277.
- [16] M. G. Hr, A. Mv, S. G. Prasad, and S. Vinay, "Development of anti-phishing browser based on random forest and rule of extraction framework," *Cybersecurity*, vol. 3, no. 1, Dec. 2020, doi: 10.1186/s42400-020-00059-1.
- [17] B. C. Ujah-Ogbuagu, O. N. Akande, and E. Ogbuju, "A hybrid deep learning technique for spoofing website URL detection in real-time applications," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, Jan. 2024, doi: 10.1186/s43067-023-00128-8.
- [18] S. R. A. Samad *et al.*, "Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection," *Electronics (Switzerland)*, vol. 12, no. 7, Apr. 2023, doi: 10.3390/electronics12071642.
- [19] S. D. Gupta, K. T. Shahriar, H. Alqahtani, D. Alsaman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Annals of Data Science*, vol. 11, no. 1, pp. 217–242, Feb. 2024, doi: 10.1007/s40745-022-00379-8.
- [20] M. Almousa, T. Zhang, A. Sarrafzadeh, and M. Anwar, "Phishing website detection: How effective are deep learning-based models and hyperparameter optimization?," *Security and Privacy*, vol. 5, no. 6, Nov. 2022, doi: 10.1002/spy2.256.
- [21] A. Almomani *et al.*, "Phishing website detection with semantic features based on machine learning classifiers: a comparative study," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, Jan. 2022, doi: 10.4018/IJSWIS.297032.
- [22] M. Elsadig *et al.*, "Intelligent deep machine learning cyber phishing URL detection based on BERT features extraction," *Electronics (Switzerland)*, vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223647.
- [23] G. Mohamed, J. Visumathi, M. Mahdal, J. Anand, and M. Elangovan, "An effective and secure mechanism for phishing attacks using a machine learning approach," *Processes*, vol. 10, no. 7, Jul. 2022, doi: 10.3390/pr10071356.
- [24] S. J. Bu and H. J. Kim, "Optimized URL feature selection based on genetic-algorithm-embedded deep learning for phishing website detection," *Electronics (Switzerland)*, vol. 11, no. 7, Apr. 2022, doi: 10.3390/electronics11071090.
- [25] A. Gómez and A. Muñoz, "Deep learning-based attack detection and classification in android devices," *Electronics (Switzerland)*, vol. 12, no. 15, Aug. 2023, doi: 10.3390/electronics12153253.

BIOGRAPHIES OF AUTHORS



Ms. Preeti    M.Tech., Ph.D. Pursuing from Department of Computer Science and Applications, M.D.University, Rohtak. She has published more than 16 publications in various journals /magazines of national and international repute. Her area of research includes machine learning, deep learning, and cyber security. She can be contacted at email: miskhokhar121@gmail.com.



Dr. Priti Sharma    MCA, Ph.D. (Compute Science) is working as an Assistant Professor in the Department of Computer Science and Applications, M. D. University, Rohtak. She has published more than 60 publications in various journals/magazines of national and international repute. She is engaged in teaching and research from the last 15 years. Her area of research includes data mining, big data, software engineering, machine learning, and deep learning. She can be contacted at email: Priti@mdurohtak.ac.in.