JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 01
2026

Research Article
https://doi.org/10.56979/1101/2026/1309

An Effective Rapid Threat Detection Framework for Detecting Malicious Domains Using Supervised Machine Learning Approach

Amir Haris Bin Ahmad[1], Hadhrami Bin Ab Ghani[1], Muhammad Muzzammal Mirza[1&2], and Muhammad Anwar[1&3*]

[1]Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kota Bahru 16100, Kelantan, Malaysia.
[2]Department of Computer Science and Information Technology, Lahore Leads University, Lahore 54000, Pakistan.
[3]Department of Information Sciences, Division of Science and Technology, University of Education, Lahore 54000, Pakistan.
[*]Corresponding Author: Muhammad Anwar, Email: anwar.muhammad@ue.edu.pk

Received: January 10, 2026 **Accepted:** April 16, 2026

Abstract: The Domain Name System (DNS) that translates human readable domain names into IP addresses is important in Internet communication. The critical importance of DNS and its widespread trust has made it a typical target of cybercrimes such as cache poisoning and DNS spoofing, which redirect users to dangerous websites where they lose a lot of money and data. Conventional blacklist-based detection methods are becoming less and less effective against threats that are changing quickly. This research investigates the effectiveness of the supervised machine learning techniques to identify malicious domains using DNS data. Three models (RF TFIDF Lex, SAE Stacking, and CharCNN BiLSTM) are developed and evaluated using the CAIDA DNS dataset. The standard performance measures of accuracy, precision, recall, and F1-score are used to evaluate the performance of the models in a multi-class classification scenario. The experimental data indicate that CharCNN_BiLSTM model is superior to the rest of the approaches, particularly in detecting more complex patterns in the domain, such as DGA and phishing domains. These findings show that deep learning models constructed using sequences are better at reflecting the structural characteristics of malicious domains. The research indicates that in the case of DNS-based threat identification, machine learning-based methods will provide a more credible alternative to traditional methods.

Keywords: Malicious Domain Detection; Supervised Machine Learning; DNS Security

1. Introduction

In 1983, Paul Mockapetris developed a distributed, hierarchical system to allocate memorable domain names (such as www.google.com) to numerical IP addresses. In 1986, this system was made a standard of the internet (RFC 1043, 1035) and allowed managing domains in a scalable way with easy navigation [1]. Domain Name System Domain Name System was developed based on the necessity to substitute the manual hosts.txt file on ARPANET in the 1970s with the Internet expanding. The internet was made to be usable because it shifted to a decentralized system and had security features such as zone transfer, BIND and DNSSEC. The internet is the largest computer network worldwide with more than 580 million users. To a user, each node or resource on this network is uniquely named, the domain name. Internet resources are things such as web servers, used to access websites; mail servers, used to deliver emails; and application servers, used to remotely access databases and software systems [2].

The most typical ways that DNS-based risks manifest are malicious domain names that do phishing, malware distribution, and domain generation algorithms (DGA). Since attackers keep on changing patterns of domains to avoid the traditional forms of detection, detecting such malicious domains is a very important cyber security problem.

The machine learning and deep learning models have been largely utilized in detection of these tunnels using handcrafted statistical features, flow-based features, and representation learning techniques such as CNNs, RNNs. These models tend to be very accurate at the point of testing data generated in the laboratory or in a particular region, but differ significantly in their construction, the data they rely on, the kinds of network traffic they take into account, how they quantify performance, and how they respond to new tools or slow, obscure data leaks. This renders it difficult to equate such models and select the most appropriate to be used in the real world [3].

The current machine learning-based DNS security framework however has the disadvantage of lack of generalization and flexibility to the emerging threats. Depending on the pre-defined features and rule-based heuristics, as Ayo et al. report[3][4], will be too risky to the dynamic nature of the attacks. This weakness highlight the need of more dynamic and proactive security systems that can identify the new and emerging threats before they can cause greater destruction. This research can contribute to these problems in the following ways:

- Creation of a supervised machine learning system that uses DNS data to identify malicious domains.
- Three classification models are compared: RF_TFIDF_Lex, SAE_Stacking, and CharCNN_BiLSTM. Models' performance is examined in several domain categories, such as malware, phishing, DGA, and benign.
- Finding deep learning-based strategies that work better for intricate DNS threat patterns.

This paper does not cover any additional DNS-based security such as DNS tunneling or traffic-level abnormalities. However, it only focuses on malicious domain classification.

2. Literature Review

Several supervised machine learning models are proposed in the literature for DNS security. The study by [5] introduced FlowLens which is a proposed framework abides by directed approach showcasing how programmable network switches can be employed to mine linearly independent flow features directly from network traffic allowing parallel and fast machine learning based classification. It focuses on important security cases of covert channel detection, website fingerprinting, and botnet identification using supervised models such as XGBoost, Random Forest and Naïve Bayes respectively. FlowLens lowers the cost and enhances scalability of low-latency feature extraction by offloading it on the data plane so that real-time analytics can be achieved at high throughput. The paper demonstrates the effectiveness of their system across a few different datasets and shows that they are able to achieve high accuracy while adding minimal latency on average. While the findings are positive, it may be that depending on rather specialized switch capabilities and based on labeled datasets will suppress generalizability in diverse or evolving network environments.

In the study conducted by [6] investigates the ability of ensemble machine learning to decide between malicious and benign DNS over HTTPS traffic. Many models such as Random Forest, KNN, and Logistic Regression are tested by utilizing the CIRA-CIC-DoHBrw-2020 dataset and thirty certain characteristics. Since ensemble methods are employed, the ensemble trials yield an accuracy of 100 percent, thus proving the effectiveness of ensemble similarly. Given the focus on offline static data and an omission real-world encrypted data traffic condition, the research also demands further study. Moreover, the consideration of the pattern of development of attacker actions is excluded, reducing their advantage in preventing emerging threats.

The study by Jerabek et al., [7] used flow-level statistical features extracted from NetFlow and IPFIX records and apply several supervised machine learning algorithms like Random Forest, Decision Tree, Gradient Boosting. The study implements their approach on real-world datasets and demonstrates a strong

attack generalization in differential classification of DoH from regular HTTPS traffic. In this work, author proposes a privacy-preserving detection scheme that does not depend on the payloads and hence can be applied in encrypted domains. Nonetheless, the methodology might be limited in its adaptability to such adversaries which are dynamically changing their traffic patterns to bypass statistical detection (calling for frequent model re-training).

The study by Calle et al., [8] address large scale DNS tampering, especially by authoritarian states such as China's Great Firewall. The study uses both supervised and unsupervised models on OONI's global DNS data as shown in figure 2.2, achieving high precision to learn tampering heuristics in a wide sense and detect previously unknown manipulations. Moreover, the study examines the performance in various time windows (1-24 months) to provide a clue on the dynamics of censorship over time. Strong scalability and cross-regional deployment capability, but no specific interpretability or a false-positive cost in benign anomalies (e.g., load balancing). Further refinement is required in order to support resilience and active response mechanisms to adversarial.

The article [9] discusses the different machine learning models (i.e. XGBoost, MLP, SVM) in terms of DNS flooding DDoS attack-based CIC-DDoS2019 data. The current study enhances past restrictions of model feature selection (filter, wrapper and embedded) methods. Although the technique has been demonstrated in emulated attack situations, this paper does not address zero-day or polymorphic attacks and does not address how this can be integrated with real-world DNS infrastructure. Neither does it examine the robustness of models to attack or examine the performance of detection to evolve with time.

Schummer et al. [10] introduces a complete approach to detection of network anomalies as well as DNS-specific irregularities using different supervised and unsupervised machine learning mechanisms, clustering, change-point detection and classification models. This is not limited to DNS attacks, however the system's ability to watch for packet loss, congestion and behavioral anomalies can apply to DNS anomaly settings. First, by allowing the interpretation of model confidence, paper address one of the principal drawbacks in ML-based cybersecurity systems the interpretability of such models. The paper is not comprehensive with respect to considerable DNS protocol nuances like tunneling, DoH or DGA-based evasion that are abstracted away in this study and therefore restrict its direct applicability for deployment in specialized DNS security systems. Although without tuning for the DNS-specific deployment, this remains suitable as a general-purpose anomaly detector.

3. Proposed Methodology

In order to identify malicious domain names in DNS traffic, this study suggests a supervised machine learning-based approach. Accurately classifying domain names into many categories, such as malware, phishing, benign, and domain generation algorithm (DGA) based domains, is the major goal [11] The three primary phases of the suggested framework are feature extraction, classification, and data pre-processing.

The DNS data from CAIDA dataset is cleaned and normalized during the data pre-processing step to eliminate noise and inconsistencies. After that, extraneous characters are eliminated and formats are standardized to get domain names ready for feature extraction. Both lexical and sequence-based features are obtained during the feature extraction phase. While sequence-based representations are created for deep learning models to capture the structural and morphological elements of domain names, lexical features include statistical trends, character distribution, and domain length. Three supervised machine learning models are used and assessed during the classification phase:

- RF_TFIDF_Lex: A Random Forest model with manually created lexical features and TF-IDF.
- SAE_Stacking: An ensemble model that uses stacked autoencoders to capture high-level feature representations.
- CharCNN_BiLSTM: This hybrid deep learning model learns sequential patterns in domain names by fusi ng bidirectional LSTM layers with convolutional neural networks.

To guarantee robustness, these models are trained and assessed using several experimental runs. Standard metrics such as accuracy, precision, recall, and F1-score are used to evaluate performance.[12].

3.1. ata Acquisition

The publicly accessible CAIDA DNS dataset, which includes both malicious and valid DNS traffic, is the source of the experimental dataset used in this investigation. There are about 4,800 domain samples in the collection. Benign, malware, phishing, and domain generation algorithm (DGA)-based domains are the four classes that make up the dataset. The dataset's inherent imbalance reflects the features of DNS traffic in the actual world. For instance, the bnign class has substantially fewer samples (134 instances) than the DGA class, which makes up the majority of samples (3,259 instances). The remaining samples are split between the categories of phishing and malware. The confusion matrices and the class distribution utilized in the experiments match.

3.2. Data Pre-processing

There were redundant entries, noise, and discrepancies in the DNS data that was gathered. Pre-processing is therefore carried out to enhance model performance and data quality. Among the preprocessing procedures are:

- Elimination of invalid and duplicate domain entries
- Domain name normalization
- Removing unique and unnecessary characters.
- Categorical class label encoding for supervised learning

The cleanliness, consistency, and suitability of the dataset for feature extraction and model training were guaranteed by these procedures.

3.3. Feature Extraction

Both lexical features and sequence-based representations are extracted in order to efficiently capture the features of domain names. The structural characteristics of domain names are the source of lexical features. Among them are:

- Domain length
- Frequency of vowels, consonants, and digits
- Character entropy
- The existence of unique patterns, such as hyphens and repeated characters

Furthermore, domain strings are converted into numerical feature vectors using TF-IDF (Term Frequency–Inverse Document Frequency), which allows conventional machine learning models to discover discriminative patterns. The domain names are represented by character sequences. Each domain is initially encoded into a string of letters in order to construct numerical embeddings. This approach allows the model to learn:

- Morphological patterns
- Sequential dependencies
- Hidden structures in malicious domain generation

3.4. Classification Models

This study examines the efficacy of three supervised machine learning models in DNS-based threat detection.

RF_TFIDF_Lex: This model model consists of TF-IDF, random forest classification, and manually developed lexical features. Random Forest was chosen due to its strength, interpretability and ability to handle high-dimensional feature spaces.

SAE_Stacking: Based on the input data, the hierarchical feature representations are learned with the help of Stacked Autoencoder (SAE) model. The stacking ensemble classifier is fed with the encoded features and it boosts the performance in terms of classification and generalization.

CharCNN_BiLSTM: This hybrid deep learning model is based on the combination of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. BiLSTM layers are used to identify long-range relationships in domain sequences and CNN layers identify local character-level patterns. The model can successfully identify intricate patterns linked to malicious domains, especially DGA and phishing attempts, thanks to this combination.

3.5. Model Training and Evaluation

To ensure strength and stability, all the models are trained and evaluated on a number of experimental runs. A typical split approach is used to separate the dataset into training and testing sets. Every model is trained using its corresponding setup. Standard categorization measures are used to assess performance, such as:

- Accuracy
- Precision
- Recall
- F1-score

Appropriate evaluation techniques are used because of the dataset's inherent class imbalance [13]. To account for the different class distributions, weighted averaging is specifically applied to precision, recall, and F1-score. Confusion matrices are also examined to evaluate misclassification behavior and class-wise performance. As covered in the results section, the comparative analysis of the three models sheds light on how well they identify various DNS-based threats.

The general architecture of the suggested harmful domain detection system is shown in Figure 1. Data pre-processing, lexical and character sequence-based feature extraction, and supervised machine learning model classification are all part of the pipeline. The models for multi-class classification of domain names into benign, malware, phishing, and DGA categories are trained and assessed using the processed data.

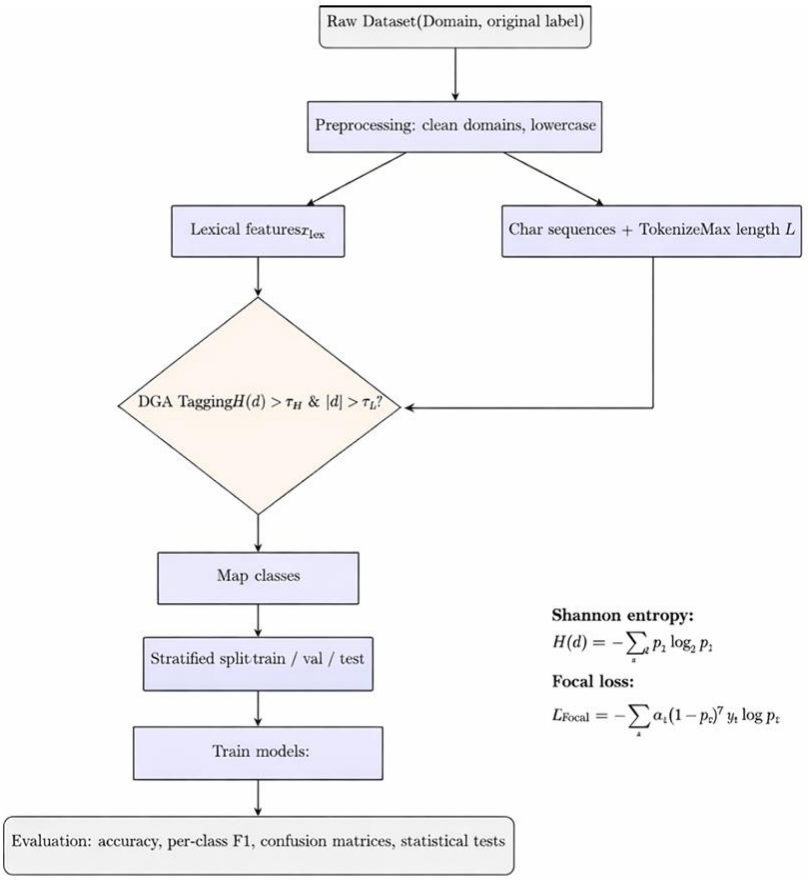

Figure 1. Overall framework of the proposed malicious domain detection system

4. Performance Evaluation and Results

Three supervised machine learning models outlined in the technique are compared and evaluated in this section. The dataset used in the study (consisting of approximately 4,800 domain samples) is based on publicly available CAIDA DNS dataset. Benign, malware, phishing, and domain generation algorithm (DGA)-based domains are the four classes that make up the dataset. The intrinsic imbalance of the dataset is

the characteristics of DNS traffic in reality. An example is that the innocuous class contains well below the samples of the DGA class (134 and 3,259 instances, respectively). The rest of the samples are divided into the following categories: phishing and malware.

This study implements three supervised models, RF_TFIDF LEX, SAE Stacking and CharCNN BiLSTM. Each model is evaluated against a series of iterations of tests to measure the accuracy, precision, recall, F1-score and confusion matrix analysis [14]. An unbalanced dataset that reflects the distributions of DNS traffic in the real world is used for the evaluation. Figure 2 through Figure 4 of the confusion matrices helps to comprehend the behavior of the three hybrid DNS detection models RF_TFIDF_Lex, SAE_Stacking, and CharCNN_BiLSTM at the class-level. Despite the great overall predictive performance of all three models, they have vastly different misclassification behaviors, particularly of minority classes like phishing and benign that reflect imbalances in real life that are commonly seen in DNS traffic data.

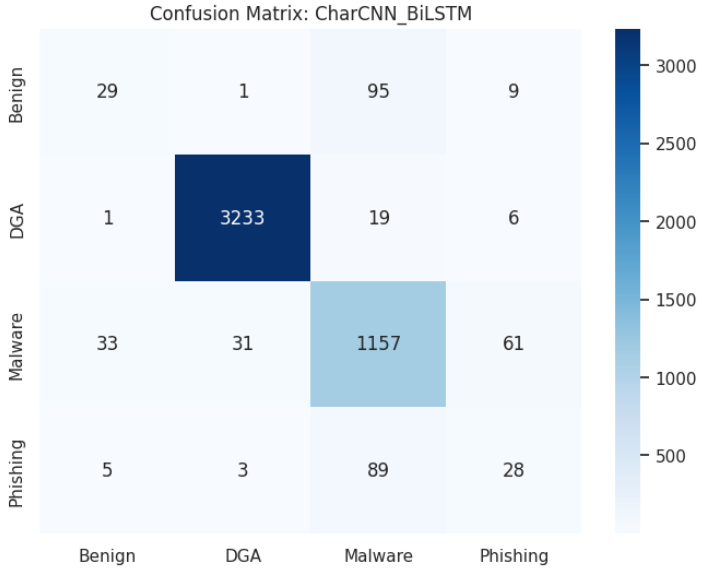

Figure 2. Confusion Matrix of CharCNN_BiLSTM

In all models, the DGA class has the largest true positive rates.

- CharCNN_BiLSTM has the highest performance with 3,233 correct samples, and overall, it misclassifies 26 samples.
- RF_TFIDF_Lex and SAE_Stackingare also high-accuracy (3126 and 3119 true positives respectively), but CharCNN_BiLSTM is more stable, and has fewer false positives to other classes.

The performance of the three models in the DGA class can be attributed to its high representation (3259 samples) that supports the previous research findings that deep and ensemble models generally perform well in classes that are of the majority domain [15].

A clear separation in model capability emerges for the *Malware* class:

- CharCNN_BiLSTM attains the highest recall with 1,157 correct predictions, misclassifying fewer samples (125 mislabels) compared to the other models.
- SAE_Stacking comes second best (1,215 correct predictions) but its errors are distributed more equally between Benign and Phishing, which may indicate that it is sensitive to lexical ambiguity in domain names.
- RF_TFIDF_Lex is poor compared to the deep learning models with only 1,024 true positives and a large amount of Malware samples spilling over to the DGA category (253 misclassifications).

This implies that lexical-based Random Forest models cannot easily differentiate between DGA-like malware generated domains, in line with previous research that demonstrated character-level deep models to be more powerful in identifying malware-related DNS patterns [16].

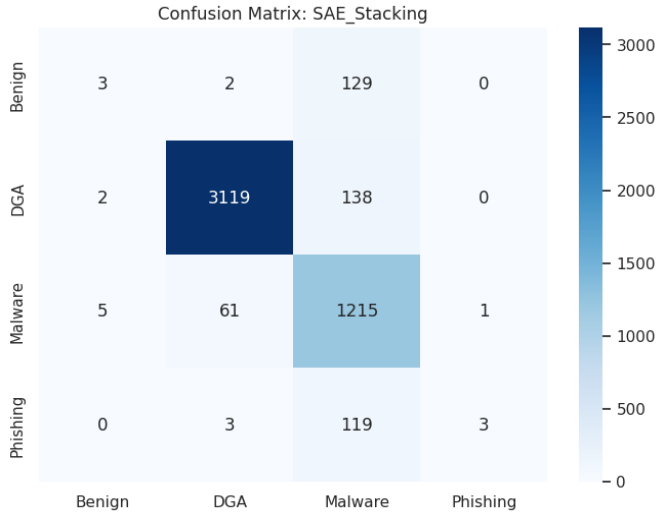

Figure 3. Confusion Matrix of SAE-Stacking Ensemble

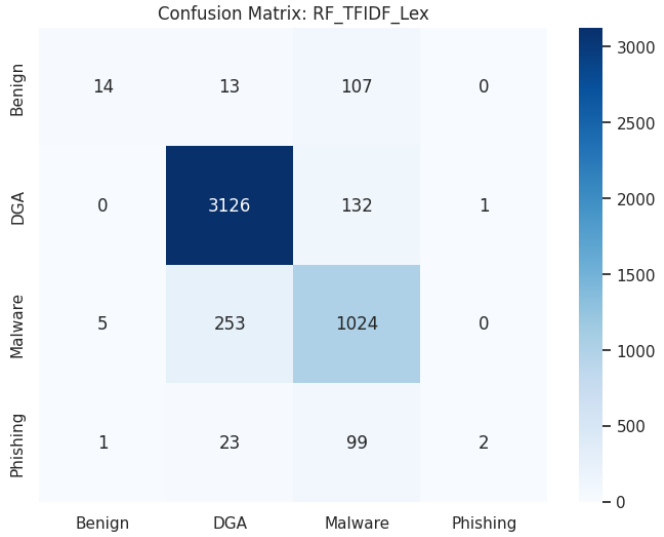

Figure 4. Confusion Matrix of RF_TFIDF_Lex

The Benign category is the most challenging to the classical and hybrid feature-based models because of low support (134 samples) and duplication of lexical features between malicious domains.

- CharCNN_BiLSTM provides the strongest performance, correctly classifying 29 samples, significantly higher than SAE_Stacking (3) and RF_TFIDF_Lex (14).
- SAE_Stacking presents significant challenge, with almost all Benign samples falsely classified as Malware (129 errors), which means that the model is very sensitive but not specific to benign behaviour.
- RF_TFIDF_Lex also confuses *Benign* for *Malware* (107 mislabels), reflecting its struggle to distinguish natural language domain patterns from algorithmically generated ones.

This underscores the significance of contextual and sequential character encoding-which is well employed in CharCNN_BiLSTM- to encode subtle structural information that differentiates benign traffic [13].

Phishing detection remains the hardest across all models.

- CharCNN_BiLSTM has the best recall on Phishing with recall of 28 samples, doing better than SAE_Stacking(3) and RF_TFIDF_Lex(2).
- Most malicious domains are mislabeled as Malware or Benign and this implies that there is structural similarity of domain patterns that are social engineered (e.g., typosquatting, homoglyph attacks).

The enhanced performance of CharCNN_BiLSTM can be explained by the fact that it manages to capture a deeper morphological information, which is in agreement with recent reports that character-level deep sequence models are better than feature-engineered models in detecting phishing domains [14].

Table 1. Performance Comparison of Three Hybrid DNS Detection Models (Weighted Average Metrics)

Model	Accuracy	Precision	Recall	F1-Score
Model A: CharCNN_BiLSTM	0.9265	0.63	0.58	0.60
Model B: SAE_Stacking	0.9042	0.70	0.49	0.47
Model C: RF_TFIDF_Lex	0.8679	0.76	0.47	0.48

Table 1 summarizes the performance of the three hybrid models: CharCNN-BiLSTM, SAE_Stacking and RF_TFIDF_Lex. Based on the distribution of the classes in the dataset, the given precision, recall, and F1-score values are weighted averages of all classes. This method guarantees that the assessment takes into account the effects of class imbalance, especially for underrepresented classes like phishing and benign domains.

5. Conclusion

Applications for cyber security are seriously threatened by malicious domains. DNS communication is considered to be always trustworthy and, therefore, conventional security measures cannot detect fraudulent domains. Machine learning techniques come in handy to deal with such issues. We proposed in this work a supervised machine learning-based framework of DNS data-based domain name classification. Three models namely RF_TFIDF_Lex, SAE_Stacking, and CharCNN_BiLSTM were put into practice and assessed using a multi-class dataset that included four categories: DGA, Phishing, Malware, and Benign. According to the results of the experiment, methods based on deep learning, specifically, CharCNN_BiLSTM, are more effective in detecting complex patterns in malicious areas. The findings reveal the extent to which lexical and sequence-based features can be used successfully in detecting hostile domains. Also, the analysis demonstrates that supervised learning models, which are not based on pre-established signatures, can significantly enhance the detection of emerging domain-based attacks.

References:

1. Alshammari A and Alqarni A (2026). Cyberattack detection and prevention framework for the healthcare sector using machine learning techniques. International Journal of Advanced and Applied Sciences, 13(1): 1-12.
2. Sadiq, A., Anwar, M., Butt, R. A., Masud, F., Shahzad, M. K., Naseem, S., & Younas, M. (2021). A review of phishing attacks and countermeasures for internet of things-based smart business applications in industry 4.0. Human behavior and emerging technologies, 3(5), 854-864.
3. Altaimimi M (2025). Improved network traffic classification using hashing techniques in machine and deep learning. International Journal of Advanced and Applied Sciences, 12(5): 255-261.
4. Awais, Z., Hussain, M., Elshenawy, A., Arsalan, A., Anwar, M., Habib, M. A., ... & Ahmad, M. (2025). ISCC: Intelligent Semantic Caching and Control for NDN-Enabled Industrial IoT Networks. IEEE Access.
5. Alkhliwi S (2022). An efficient dynamic access control and security sharing scheme using blockchain. International Journal of Advanced and Applied Sciences, 9(8): 28-40.
6. Singh SK, Roy PK. Malicious traffic detection of DNS over https using ensemble machine learning. International Journal of Computing and Digital Systems. 2022 Feb 15;11(1):189-97.
7. Jerabek K, Hynek K, Rysavy O, Burgetova I. Dns over https detection using standard flow telemetry. IEEE Access. 2023 May 12;11:50000-12.
8. Alanazi, S. S., & Alanazi, A. A. (2022). Knowing the Unknown: The Hunting Loop. Int. j. adv. appl. sci, 1(9), 8-19.
9. El Attar A, Khatoun R, Chbib F, Fadlallah A, Serhrouchni A. DNS flooding attack detection scheme through Machine Learning. In2024 International Wireless Communications and Mobile Computing (IWCMC) 2024 May 27 (pp. 132-137). IEEE.
10. Alshammari, N., Shahzadi, S., Alanazi, S. A., Naseem, S., Anwar, M., Alruwaili, M., ... & Ahmad, F. (2024). Security monitoring and management for the network services in the orchestration of SDN-NFV environment using machine learning techniques. Computer Systems Science and Engineering, 48(2), 363-394.
11. O. Abualghanam, H. Alazzam, B. Elshqeirat, M. Qatawneh, and M. A. Almaiah, "Real-time detection system for data exfiltration over DNS tunneling using machine learning," Electronics, vol. 12, no. 6, p. 1467, 2023.
12. Z. Weifang, L. Zhong, S. Yu, K. Ningning, and S. Peng, "Attention-based Mechanism for Anomaly Time Slice Detection in DNS Tunnel Communication," in 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2023, pp. 767–771.
13. M. Luo, Q. Wang, Y. Yao, X. Wang, P. Yang, and Z. Jiang, "Towards comprehensive detection of DNS tunnels," in 2020 IEEE Symposium on Computers and Communications (ISCC), 2020, pp. 1–7.
14. Hussien AS and Mounes HM (2024). Cybersecurity in international commercial arbitration (Legal vision). International Journal of Advanced and Applied Sciences, 11(2): 219-229.
15. Rehman, F., Sharif, H., Anwar, M., & Riaz, N. (2023). Big Data Analytics for Cybersecurity in IoE Networks. In Cybersecurity Vigilance and Security Engineering of Internet of Everything (pp. 163-176). Cham: Springer Nature Switzerland.
16. Muneer, M., Rehman, F., Sajjad, M. H., Anwar, M., & Qureshi, K. N. (2024). Security and privacy concerns in AI models. In Next Generation AI Language Models in Research (pp. 293-326). CRC Press.