AL-Bahir
Journal for Engineering and Pure Sciences
BJEPS

Volume 6 | Issue 2 Article 9

A Comprehensive Analysis of Feature Selection and Classification Techniques for Web Attack Classification

Zaman Ahmed Razak Abdul Ali
Islamic Azad University, Faculty of Engineering, Department of Artificial Intelligence and Robotics Engineering, Tehran, Iran

Recommended Citation
Ali, Zaman Ahmed Razak Abdul (2025) "A Comprehensive Analysis of Feature Selection and Classification Techniques for Web Attack Classification," *Al-Bahir*: Vol. 6: Iss. 2, Article 9.
Available at: https://doi.org/10.55810/2313-0083.1091

A Comprehensive Analysis of Feature Selection and Classification Techniques for Web Attack Classification

Source of Funding
No

Conflict of Interest
No

ORIGINAL STUDY

A Comprehensive Analysis of Feature Selection and Classification Techniques for Web Attack Classification

Zaman A. Razak Abdul Ali

Islamic Azad University, Faculty of Engineering, Department of Artificial Intelligence and Robotics Engineering, Tehran, Iran

Abstract

Cybersecurity is seriously threatened by web attacks, which makes it necessary to create strong classification models as soon as possible for early detection and prevention. Networked system security is now a crucial global concern affecting people, businesses, and governments. Attacks on networked systems are occurring far more often, and the attackers' strategies are always changing. One defense against these attacks is intrusion detection. Machine learning is a popular and useful method for creating intrusion detection systems (IDS). Having more representative and discriminative traits greatly enhances an IDS's performance. We do a thorough review of feature selection and classification methods for the purpose of classifying web attacks in this work.

This work aims to use Decision Tree, Random Forest, and Logistic Regression as classifiers and evaluate their accuracy with the phishing website dataset. To improve the performance of the selected classifiers, subsequently, various feature selection methods, including Recursive Feature Elimination (RFE) and the Chi-square test, were used to identify the most relevant features for classification. Feature selection is a key topic for dimension reduction and classification in high-dimensional datasets. During the feature selection procedure, only the most relevant attributes from the datasets will be chosen.

Logistic Regression, Decision Tree, and Random Forest. We assess the models' performance based on metrics such as accuracy and classification reports to determine their effectiveness in classifying web attacks. Our findings provide insights into the effectiveness of different feature selection and classification techniques for web attack classification, contributing to the advancement of cybersecurity research.

Keywords: Cybersecurity, Web attacks, Intrusion detection system (IDS), Machine learning, Feature selection, Decision tree, Random forest, Recursive Feature Elimination (RFE), Chi-square test, Phishing website dataset, Network security, Attack detection

1. Introduction

We use the Internet for nearly everything in our everyday lives these days. However, cyberattacks are a challenging issue to handle since they have cost us a great lot of money and hassle. Furthermore, since the Internet of Things (IoT) has become more and more integrated into our daily lives, IoT-enabled networks are now more susceptible to cyberattacks due to the usage of unreliable wireless connections, resource-constrained designs, and a variety of IoT devices.

Every day, tens of thousands of new malware programs are developed. For instance, according to reports, Kaspersky's detection systems found 400 thousand new viruses and malware on average every day in 2022 [1]. In 2022, Kaspersky's systems found around 122 million dangerous files overall, which is six million more than the previous year [1]. In November 2022, the Azure DDoS Protection team at Microsoft witnessed a large DDoS assault that broke past records by achieving a maximum speed of 3.47 Tbps and an average packet speed of 340 billion packets per second (pps) [2].

Received 20 November 2024; revised 18 January 2025; accepted 9 February 2025.
Available online 28 April 2025

E-mail address: abo.mosa.lara@gmail.com.

https://doi.org/10.55810/2313-0083.1091

Web attacks, including phishing, malware injection, and cross-site scripting, continue to pose serious threats to internet security, leading to data breaches, financial losses, and reputational damage. Timely detection and classification of web attacks are crucial for preventing their adverse impacts on individuals and organizations. In this research paper, we investigate various feature selection and classification techniques to develop effective models for web attack classification.

The most popular method for identifying cyberattacks is an intrusion detection system, or IDS. IDS systems frequently employ strategies based on statistics and signatures [3,4]. Nevertheless, the detection efficiency of intrusion detection systems (IDS) is significantly impacted by the inability of signature-based techniques to identify novel attack types or variations, and the challenge of meticulously predefining the patterns of several extant assaults, even for proficient professionals. In the meanwhile, statistical information-based intrusion detection systems (IDSs) consistently presume that normal or aberrant communications conform to a specific distribution. This is plainly untrue, and defining the elements of the assumed distribution is not straightforward. Specifically, the methods employed in cyberattacks are also become more complex.

Both benign and phishing websites are included in the well-rounded, representative sample of online traffic that the phishing website dataset is designed to give. By using feature selection methods like RFE and Chi-Square and machine learning models, this preparation guarantees that the dataset is appropriate for assessing intrusion detection systems. Researchers may create and evaluate efficient IDS systems that can reliably and accurately identify phishing and other web-based threats by using this dataset. Many researchers and developers have given the Artificial Intelligence (AI)-based IDS a lot of attention in this area. From basic ANN models to intricate deep learning models [5], traditional machine learning methods including SVM (Support Vector Machine) [6], DT (Decision Tree) [8], RF (Random Forest) [9], and ANN (Artificial Neural Network) are also beginning to be employed in IDSs. In actuality, many application systems must take cyberattack resistance into account during design [7].

Although the study's results on feature selection and classification models show promise for classifying online assaults, it is difficult to apply these strategies to other kinds of web attacks or datasets. The challenge of expanding this strategy is exacerbated by the variety of attack kinds, dataset characteristics, feature overlap, class imbalance, and the always changing environment of cyberthreats. Effective generalization would need careful model adaption, tuning, and retraining, and domain-specific modifications would probably be needed to maximize performance for novel assaults or datasets.

Combining Random Forest, Decision Trees, and Logistic Regression offers a flexible and efficient method for categorizing online threats. Random Forests improve predictive accuracy by combining the advantages of many trees, Decision Trees give flexibility and interpretability for complicated, non-linear connections, and Logistic Regression delivers efficiency and simplicity for linear issues. When combined with feature selection methods such as RFE and Chi-Square, these models help guarantee that the most pertinent characteristics are selected to increase the precision and resilience of web attack detection. The proliferation of web-based applications and services has transformed the way we interact, communicate, and conduct business. However, alongside the benefits of online connectivity comes the looming threat of web attacks, which exploit vulnerabilities in web systems to compromise data integrity, confidentiality, and availability. Web attacks encompass a wide range of malicious activities, including but not limited to phishing, SQL injection, cross-site scripting (XSS), and distributed denial of service (DDoS) attacks. These attacks pose significant challenges to cybersecurity professionals, as they can lead to financial losses, reputational damage, and legal liabilities for individuals and organizations [8].

Addressing the evolving landscape of web attacks requires proactive measures to detect, classify, and mitigate potential threats. Machine learning techniques have emerged as powerful tools for cybersecurity, offering the ability to analyze vast amounts of data and identify patterns indicative of malicious behavior. In particular, classification models trained on labeled datasets can distinguish between benign and malicious web traffic, enabling automated detection and response to security incidents [5]. RFE and the Chi-square test were chosen above other feature selection algorithms due to their exceptional suitability for the classification issues of online attacks. The Chi-square test gives a quick and reliable way to choose pertinent categorical features, whereas RFE enables model-driven, iterative feature selection that enhances performance and interpretability. These techniques are a logical option for this study as they work together to improve the classifier's capacity to identify and categorize online threats.

In this research paper, we focus on the development and evaluation of classification models for web attack classification. We aim to explore the

effectiveness of various feature selection and classification techniques in accurately identifying and categorizing different types of web attacks. By leveraging machine learning algorithms and advanced data analysis techniques, we seek to enhance our understanding of web attack patterns and improve the efficacy of cybersecurity defenses in mitigating cyber threats [7].

Our study contributes to the broader field of cybersecurity research by providing insights into the performance of different machine learning approaches for web attack classification. Through empirical experimentation and analysis, we aim to identify optimal strategies for feature selection, feature extraction, and model training that maximize classification accuracy and efficiency. Ultimately, our findings can inform the development of robust and adaptive cybersecurity solutions capable of defending against emerging web threats in an increasingly interconnected digital landscape.

The main contribution of the paper lies in its comprehensive analysis of feature selection and classification techniques for web attack classification. By systematically evaluating various methods for feature selection, including Recursive Feature Elimination (RFE) and Chi-square test, the paper identifies the most relevant features for accurate classification of web attacks.

Furthermore, the paper evaluates the performance of three widely-used classification algorithms − Logistic Regression, Decision Tree, and Random Forest − in classifying web attacks. Through empirical experimentation and thorough evaluation, the paper provides insights into the effectiveness of these algorithms and their suitability for cybersecurity applications.

Overall, the paper's contribution lies in its empirical findings, which provide valuable insights into the effectiveness of different feature selection and classification techniques for web attack classification. These insights can inform the development of more robust and efficient cybersecurity solutions, thereby enhancing the overall resilience of digital systems against web-based threats.

2. Literature review

In order to improve network infrastructure in an industrial setting [9], reviewed a few efficient features selection strategies based on correlation measurement techniques and developed a novel method for adding functions to the Fast Based Correlation Features FCBF algorithm. Nevertheless, they modify the FCBF method to the FCBFiP algorithm in their research. Dividing the feature space into equal-sized sections was the primary goal. They enhanced the machine learning and correlation apps that operate on each node by putting forward this strategy. Nonetheless, their suggested model produces better outcomes in terms of execution time and model accuracy.

In [10] developed a novel technique for identifying attacks that originate from Internet of Things devices. They also proposed and empirically evaluated the anomaly detection method, which uses autoencoders to detect anomalies in network traffic from IoT devices and extracts network performance. However, they compromised a few commercial IoT-connected devices and utilized two well-known IoT-based botnet assaults, Bashlite and Mirai, to test the suggested approach. According to experimental data, their suggested method is capable of detecting IoT device threats.

The UNSW-N15 and KDDCup99 datasets were used by the researchers in Ref. [11] to test the Genetic Algorithm (GA) and Logistic Regression (LR) wrapper-based feature selection approach. The Weka simulation tool was employed in this study. Following several simulations, the GA-LR in conjunction with the DT classifier achieved a detection score of 81.42 % and a FAR of 6.39 % using 20 of the 42 features in the UNSW-NB15 feature set. When applied to the KDDCup99 dataset, the GA-LR and DT classifier achieved a 99.90 % detection score and a 0.105 % FAR rate using 18 features.

The authors in Ref. [12] introduced a filter-based method for Distributed Denial of Service (DDoS) detection that makes use of several filters. Information Gain, Chi-Square and Gain Ratio, and ReliefF are some of the filter techniques that were employed. The researchers used the NSL-KDD attack detection dataset to try to show off this system's capabilities. The authors used the Decision Tree (DT) algorithm for the classification process, which was trained and verified using the k-fold cross-validation approach, where $k = 10$. According to the testing data, the DT classifier achieved a detection accuracy score of 99.67 % and a false alarm rate (FAR) of 0.42 % using just 13 features out of the 42 features (full feature space). But this study didn't go into great detail on the study of the multiclass classification problem of the NSL-KDD. In order to reduce the quantity of input characteristics (features) needed for their model's training and testing, the authors in Ref. [13] constructed an IDS utilizing a filter-based approach. A correlation input selection method was used with the DT classifier. The NSL-KDD dataset is used in the experimental procedures. After applying the filter to the feature space, 14 features were chosen. Additionally, the

author took into account both the binary classification configuration and the multiclass classification setup, which encompassed all five kinds of assaults inside the NSL-KDD. The results of the trial showed that the system produced an accuracy of 90.30 % for the binary setting and 83.66 % for the multiclass configuration. In Ref. [10] developed a novel technique for identifying attacks that originate from Internet of Things devices. They also proposed and empirically evaluated the anomaly detection method, which uses autoencoders to detect anomalies in network traffic from IoT devices and extracts network performance. However, they compromised a few commercial IoT-connected devices and utilized two well-known IoT-based botnet assaults, Bashlite and Mirai, to test the suggested approach. According to experimental data, their suggested method is capable of detecting IoT device threats. In an attempt to choose the best feature space, Janarthanan and Zargari [14] used the UNSW-NB15 to build many feature selection algorithms. The Ranker Method, Greedy Stepwise, Information Gain, and attribute evaluator (CfsSubsetEval) were all built using the Weka tool. Two subsets were taken into consideration following a variety of simulations. The effectiveness of each subgroup was assessed by the authors using the Kappa Statistic metric. During the studies, a variety of classifiers were taken into consideration; nevertheless, the RF classifier was chosen as the most effective approach based on its overall performance. The first subset, which contained eight significant characteristics, had an accuracy of 75.6617 % across the test dataset and a Kappa Score of 0.6891. With just five significant characteristics, the second group obtained a Kappa value of 0.7639 and an accuracy of 81.6175 %.

To determine the Feature Importance (FI) score for each attribute in the UNSW-NB15 dataset, Maajid and Nalina [15] used a feature reduction technique based on the RF algorithm. The feature with the greatest FI is the most crucial property in the classification process with respect to the target variable (class), according to the FI algorithm, which is a ranking system. A feature subset with 11 properties was chosen following a number of studies. Regarding the classification process, the writers took into the following machine learning techniques: RF, kNN, DT, XGBoost, and Bagging Meta Estimator (BME). The accuracy and F-Measure scores derived from test data were employed as the primary measures to evaluate the effectiveness of various approaches. The best results were produced by the RF algorithm using an AC of 75.56 % and an FM of 73.00 %.

Using the UNSW-NB15 dataset, Vikash and Ditipriya [16] created an Inference Detection System (IDS). 22 important features were chosen using a feature reduction process that was influenced by the Information Gain methodology. The system combined many tree-based classifications with an integrated rule-based approach. Attack Accuracy (AAc), F-Measure (FM), and False Alarm Rate (FAR) were used to gauge performance. The findings revealed a FAR of 2.01 %, an FM of 90 %, and an AAc of 57.01 %. According to the study, rigid tree-based approaches might be replaced with different machine learning techniques.

[17] proposed an intrusion detection system for the UNSW-NB15 dataset. They performed feature selection in two stages. In the first stage, they computed the best features using information gain and Random Forest methods and selected the common features received from these algorithms. In the second stage, they gave these features to the RFE algorithm to eliminate the low-ranked features. For the classification, they used a multilayer perceptron (MLP) classifier. They used only one classifier; however, experimentation with various classifiers could have potentially yielded improved classification metrics for this research. In Ref. [18] presents a comprehensive review of feature selection techniques for intrusion detection systems (IDSs). The authors survey various feature selection methods, including filter, wrapper, and embedded techniques, and evaluate their effectiveness in enhancing the performance of IDSs. The review highlights the importance of feature selection in reducing the dimensionality of high-dimensional data and improving the accuracy and efficiency of intrusion detection models. By synthesizing the findings from existing research, the paper provides valuable insights into the state-of-the-art feature selection techniques and their applications in the field of cybersecurity.

Kasongo [19] proposed an intrusion detection system for industrial IoT networks. This researcher used the UNSW-NB15 dataset in his work. Feature selection is performed using a Genetic Algorithm where a Random Forest is used as a fitness function. Random Forest, Logistic Regression, Naive Bayes, Decision Tree, Extra-Trees (ET), and Extreme Gradient Boosting (XGB) classifiers are used for binary classification. This research achieved the best classification accuracy when using the RF classifier with 16 selected features. This research did employ different classifiers but did not leverage an ensemble model.

[20] This reviews article explores the application of ensemble learning techniques in intrusion detection systems (IDSs). The authors provide an overview of ensemble methods such as bagging, boosting, and

stacking, and discuss their strengths and weaknesses in the context of detecting network intrusions. By combining multiple base classifiers, ensemble learning models can effectively leverage the collective intelligence of diverse algorithms to improve classification accuracy and robustness. The review evaluates the performance of different ensemble approaches on benchmark datasets and identifies promising directions for future research in the field of intrusion detection.

[21] This research paper investigates adversarial attacks and defenses in deep learning models, with a focus on their implications for cybersecurity applications. The authors analyze the vulnerabilities of deep learning models to adversarial examples, which are carefully crafted input samples designed to deceive the model's predictions. The paper explores various adversarial attack techniques, including gradient-based attacks and transferability attacks, and examines strategies for defending against them, such as adversarial training and input preprocessing. By shedding light on the security risks associated with deep learning models, the study underscores the importance of developing robust and resilient machine learning algorithms for cybersecurity applications.

[22] intelligent agent system uses automated feature extraction and selection to detect DDoS attacks. By a ratio of 99.7, the system surpassed the DDoS assault detection methods in our study using the custom-generated dataset CICDDoS2019. We have created a device for this system that combines sequential feature selection with machine learning techniques. The learning phase of the system selected the best qualities and rebuilt the DDoS detection agent when DDoS attack traffic was dynamically discovered. Using the most recent technologies, our recommended method provides faster dispensation than the current standard while maintaining state-of-the-art detection accuracy.

An effective method for MANET intrusion detection was proposed by Ninu [23], using a Gas Solubility Optimization (EHGSO). The recently created EHGSO algorithm is used in the early phases of safe routing to choose the optimal pathways. These method's fitness criteria include energy, distance, quality of the neighborhood, and connection quality. The suggested EHGSO, which combines HGSO and EWMA (Exponential Weighted Moving Average), is more efficient. During the second phase, when the transmitted data packets are altered and the Knowledge discovery in databases (KDD) characteristics are retrieved, the intrusion detection phase starts at the base station. After extracting the KDD features, the data is enhanced. The suggested EHGSO technique is used to train the Deep Neuro Fuzzy Network prior to performing intrusion detection. The recommended strategy definitely works better than other already used methods. With respect to energy, throughput, packet loss, jitter, (PDR), accuracy, and recall, the proposed method obtains 95.877 %, 0.950, and 0.924 in the absence of assaults.

In order to choose the optimal features with the highest fitness value, the Grey Wolf Optimisation (GWO) approach is applied, hence streamlining the IDS procedure overall. The (DCNN) approach is used to predict whether or not the findings are harmful once the data has been categorized. Throughout the examination, we assessed several performance indicators [24], want to apply integrated optimization and classification algorithms in order to correctly predict the given label.

Its functional components include pre-processing, feature extraction, optimization, and classification. First, the input datasets are pre-processed, meaning that duplicates are eliminated and voids are filled in. Next, the (PCA) approach is used to choose a set of features in order to improve classification accuracy even more.

In order to choose the optimal features with the highest fitness value, the Grey Wolf Optimisation (GWO) approach is applied, hence streamlining the IDS procedure overall. The (DCNN) approach is used to predict whether or not the findings are harmful once the data has been categorized. Throughout the examination, we assessed several performance indicators.

3. Research methodology

The proposed method consists of several stages, the first of which is the acquisition of digital data from a dataset. The data is then cleaned, checked for missing values, and divided into training and testing data after being assessed and pre-processed to remove any outliers. The design of proposed system for data crawling and classifier generation satisfies the following requirements:

(1) As misclassifying a genuine website could result in serious repercussions like legal action and financial loss, the classifier should operate with few false positives. The administrator of such systems may even be ready to give up a high true positive rate in exchange for a low false positive rate in a real commercial application.
(2) The classifier must be resistant to criminals' evasive strategies. To achieve this, the feature set

must include a sizeable proportion of features that are difficult enough to forge.

(3) Regarding modifications to the feature set and data collecting procedure, the feature collection system should be adaptable. Given how quickly phishing websites change, the feature set—which is what determines classification—might occasionally need to be adjusted. The technology must therefore be adaptable to modifications in a seamless manner.

Based on the aforementioned goals, Fig. 1 represents the general organization and elements of the data collection and classifier training system. Each component's specifics are covered one after the other.

Adaptability, real-time detection, scalability, and resistance to evasive strategies are key considerations for modifying the suggested system into a practical intrusion detection system. The system can continue to function effectively in the face of dynamic and changing attack plans by utilizing multi-layered detection methodologies, utilizing powerful machine learning techniques, and making sure that feature selection is ongoing and efficient. To keep ahead of the rapidly evolving cybersecurity scene, it is imperative to have the flexibility to update and improve both the models and the features.

3.1. Data acquisition

The suggested approach made use of a phishing website detector. The relevant data that can be utilized as inputs for model creation are provided by the data set, which is accessible as both text and.csv files. A list of more than 11,000 websites' URLs. Every sample has 30 site attributes and a class value that designates whether it is or is not a phishing website (1 or −1). The data set seeks to define the non-functional and functional needs for the project as well as acting as an input for project scoping [23].

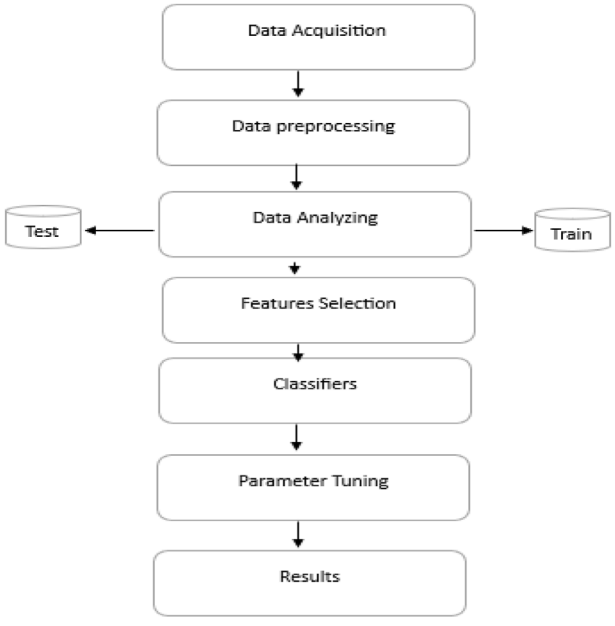

Fig. 1. Proposed system.

Dataset contain the following features{ 'UsingIP', 'LongURL', 'ShortURL', 'Symbol@', 'Redirecting//', 'PrefixSuffix-', 'SubDomains', 'HTTPS', 'DomainRegLen', 'Favicon', 'NonStdPort', 'HTTPSDomainURL', 'RequestURL', 'AnchorURL', 'LinksInScriptTags', 'ServerFormHandler', 'InfoEmail', 'AbnormalURL', 'WebsiteForwarding', 'StatusBarCust', 'DisableRightClick', 'UsingPopupWindow', 'IframeRedirection', 'AgeofDomain', 'DNSRecording', 'WebsiteTraffic', 'PageRank', 'GoogleIndex', 'LinksPointingToPage', 'StatsReport',}.

Proposed system begins by loading the web attack dataset and conducting exploratory data analysis (EDA) to understand its structure and characteristics. EDA involves visualizing the class distribution. The distribution of values in the class property is given in Fig. 2, and a sort of value balancing is apparent when the ratio of phishing websites to legitimate websites is shown [23].

3.2. Data preprocessing

There are a lot of random and noisy numbers in real-world data. These datasets are pre-processed to get around these problems and yield reliable results. Noise and missing values are frequently produced during the data cleaning process. Instead of just one source, data might be gathered from many, and it needs to be merged before processing. After classifying the data and dividing it into training and test sets, several algorithms are applied to each set to get accuracy score results. Finding, fixing, or eliminating erroneous and corrupted information from a dataset or database is known as data cleansing or cleaning. Additionally, it detects incomplete or erroneous data, filling in the gaps and eliminating noise and outliers. Usually, transmission errors, storing inadequacies, or human typing errors have resulted in missing data or conflicts. This procedure reduces training time, enhances the dataset's effectiveness and quality, and makes it easier to recognize the crucial components of the exploration process. Additionally, it improves the outcomes of the machine learning model. We eliminate any redundant and incomplete claims in this work in order to get rid of unimportant claims.Data cleaning techniques such as

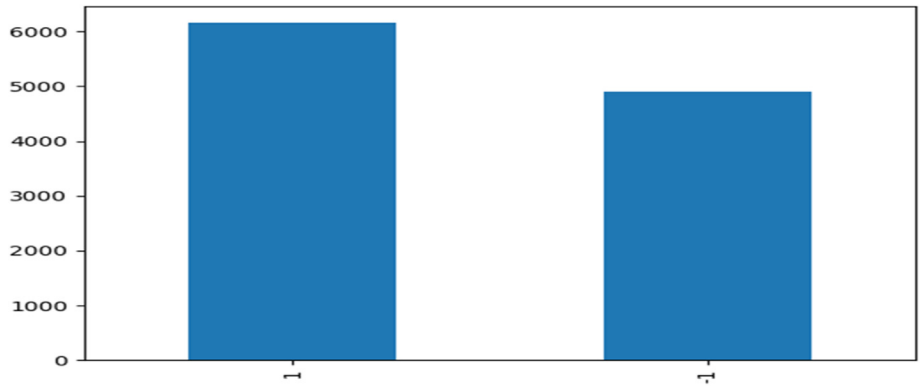

Fig. 2. Classes attribute distribution.

imputation, removal of duplicate entries, and outlier detection are employed to ensure data quality and integrity. Additionally, preprocessing techniques such as feature scaling, encoding categorical variables, and handling class imbalance may be applied to prepare the dataset for model training. Feature scaling techniques such as Min-Max scaling or Standardization are often used to normalize the range of numeric features, ensuring that no single feature dominates the model training process. Furthermore, encoding categorical variables into numerical representations en99ables machine learning algorithms to process them effectively. Class imbalance, a common challenge in binary classification tasks such as web attack classification, may be addressed using techniques such as oversampling, under sampling, or synthetic data generation to achieve a balanced distribution of classes [23].

3.3. Dataset

A well-known set of online traffic statistics, the CSIC 2010v2 dataset is widely utilized in the field of intrusion detection systems (IDS) and network security. This dataset was created as an improved version of the CSIC 2010 dataset with the goal of providing representative regular traffic as well as more varied and realistic attack situations. Helping academics and practitioners assess and compare intrusion detection systems and associated technologies for online applications is its main goal.

To replicate the traffic of an actual online application, the CSIC 2010v2 dataset was created. It is made especially to meet the requirement for realistic datasets for assessing intrusion detection and web application security methods. It encompasses a variety of web-based vulnerabilities and attack vectors and includes both normal and attack traffic data.

3.4. Feature selection

Feature selection plays a vital role in building accurate classification models by identifying the most relevant features for prediction. In this study, we employ two feature selection methods:

3.4.1. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a feature selection technique used in machine learning to enhance model performance and interpretability by selecting the most important features from a dataset. RFE operates by iteratively fitting a model, such as a linear regression, support vector machine, or decision tree, and ranking features based on their importance or contribution to the model's predictive power. In each iteration, the least important features, as determined by their coefficients or feature importance scores, are removed from the dataset. The model is then refitted on the reduced dataset, and this process repeats until the desired number of features remains. The iterative elimination helps in refining the feature set to those that most significantly impact the model's performance, thereby reducing overfitting and computational complexity. RFE can be combined with cross-validation to ensure robustness and generalizability of the selected features, making it a powerful tool for creating parsimonious models that maintain high predictive accuracy Table 1.

3.4.2. Chi-square test

The Chi-square test is a statistical method used to determine whether there is a significant association between categorical variables. It compares the observed frequencies in each category of a contingency table to the frequencies expected if there were no association between the variables. The test calculates the Chi-square statistic, which is

Table 1. Comparison between different algorithms.

Criteria	RFE [24]	Chi-Square Test [25]	BOC-PDO [26]	IBJA [27]	BIWSO3 [28]	IBCSA3 [29]
Type of method	Wrapper-based	Statistical/Filter-based	Optimization/Hybrid	Optimization/Hybrid	Optimization/Hybrid	Optimization/Hybrid
Approach	Recursive elimination of features	Statistical significance of features	Biologically inspired optimization	Swarm intelligence-based optimization	Swarm intelligence-based optimization	Swarm intelligence-based optimization
Suitability	Supervised learning	Categorical feature selection	Complex systems, optimization	Feature selection, multi-objective optimization	Feature selection, multi-objective optimization	Feature selection, multi-objective optimization
Computational complexity	High (model retraining required)	Low (fast and simple)	Moderate to high	Moderate to high	Moderate to high	Moderate to high
Handling of feature interactions	Considers feature interactions through model performance	Assesses features independently	Considers global and local search for optimal solutions	Can capture interactions via swarm behavior	Can capture interactions via swarm behavior	Can capture interactions via swarm behavior
Suitability for high-dimensional data	Less efficient for very high dimensions	Effective for categorical, low-dimensional data	Effective in complex, high-dimensional data	Effective for high-dimensional problems	Effective for high-dimensional problems	Effective for high-dimensional problems
Type of features	Any type (continuous or categorical)	Categorical	Continuous and categorical	Continuous and categorical	Continuous and categorical	Continuous and categorical
Model dependency	Depends on model for feature importance	Independent of model (no model needed)	Independent of model, optimization-driven	Independent of model, optimization-driven	Independent of model, optimization-driven	Independent of model, optimization-driven
Strengths	Can improve model performance, flexible	Fast, simple, and easy to implement	Suitable for optimization of complex problems	Effective in dynamic environments with high flexibility	Effective in dynamic environments with high flexibility	Effective in dynamic environments with high flexibility
Weaknesses	Computationally expensive, prone to overfitting	Assumes independence, limited to categorical features	Requires careful tuning of parameters	May suffer from local optima, computationally intensive	May suffer from local optima, computationally intensive	May suffer from local optima, computationally intensive
Application areas	Feature selection in classification tasks	Classification tasks with categorical data	Complex optimization, machine learning, cybersecurity	Feature selection in classification, optimization problems	Feature selection in classification, optimization problems	Feature selection in classification, optimization problems

the sum of the squared difference between observed and expected frequencies, divided by the expected frequencies for each category. This statistic follows a Chi-square distribution with degrees of freedom equal to the number of categories minus one. A high Chi-square value indicates a significant discrepancy between observed and expected frequencies, suggesting a potential association between the variables. The significance of the result is determined by comparing the Chi-square statistic to a critical value from the Chi-square statistic to a critical value from the Chi-square distribution table or by calculating the p-value, with a lower p-value indicating stronger evidence against the null hypothesis of no association Table 1.

By concentrating on the most pertinent characteristics and minimizing noise, feature selection enhances model performance, which has a substantial effect on classifier accuracy and computational efficiency. Faster training durations, increased accuracy, and greater interpretability are just a few of the advantages our study found, which highlight how crucial feature selection is. There is a trade-off, though, since too strict feature selection might result in the loss of crucial characteristics, which could impair model performance. Building effective and dependable web attack categorization models requires careful evaluation of the feature selection method and striking a balance between feature reduction and accuracy. Feature selection plays a crucial role in building accurate and efficient classification models by identifying the most informative features relevant to the target variable. Various feature selection techniques, including filter, wrapper, and embedded methods, may be employed to evaluate the importance of each feature and select a subset of features that contribute most to the classification task. By selecting the most relevant features and reducing dimensionality, researchers can build more effective classification models for web attack classification, leading to improved detection and mitigation of cyber threats (Table 1).

3.5. Classification

To evaluate the performance of the proposed system, three classification algorithms were used: Logistic Regression, Decision Tree, and Random Forest. These algorithms are widely used in machine learning for binary classification tasks and have demonstrated effectiveness in various domains. We train each model on the preprocessed dataset and evaluate its performance using metrics such as accuracy and classification report.

3.5.1. Logistic Regression

Logistic Regression is a widely-used statistical technique for binary classification tasks, where the goal is to predict the probability that an instance belongs to a particular class. In the context of web attack classification, Logistic Regression models the relationship between the independent variables (features) and the binary target variable (class label) using the logistic function. The logistic function transforms the output of a linear combination of features into a probability score between 0 and 1, representing the likelihood of an instance belonging to the positive class. Logistic Regression estimates the model parameters (coefficients) using optimization algorithms such as gradient descent or Newton–Raphson, maximizing the likelihood of the observed data. The resulting model can then be used to predict class labels for new instances based on their feature values. Logistic Regression offers several advantages, including simplicity, interpretability, and efficiency, making it a popular choice for binary classification tasks [30].

3.5.2. Decision tree

Decision Tree is a non-parametric supervised learning algorithm used for both classification and regression tasks. In the context of classification, a Decision Tree recursively splits the feature space into subsets based on the values of individual features, with each split maximizing the homogeneity of the target variable within the resulting subsets. At each node of the tree, the algorithm selects the feature that best separates the data into distinct classes, based on criteria such as Gini impurity or information gain. The tree continues to grow until a stopping criterion is met, such as reaching a maximum depth or minimum number of instances per leaf node. Decision Trees offer several advantages, including interpretability, ease of visualization, and the ability to handle non-linear relationships between features and the target variable. However, they are prone to overfitting, especially when the tree depth is not properly constrained [30].

3.5.3. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve classification performance and robustness. In a Random Forest, each tree is trained on a random subset of the training data and a random subset of the features, introducing variability and diversity into the ensemble. During prediction, each tree in the forest independently classifies the input instance, and the final prediction is determined by aggregating the individual tree's predictions

through voting (for classification) or averaging (for regression). Random Forest mitigates overfitting and variance by averaging the predictions of multiple trees, leading to more robust and generalizable models. Additionally, Random Forest can handle high-dimensional data and automatically select the most informative features, making it well-suited for classification tasks with complex feature interactions. Despite its effectiveness, Random Forest may suffer from increased computational complexity and reduced interpretability compared to individual Decision Trees [30].

4. Results and discussion

Experimental results reveal the effectiveness of different feature selection and classification techniques for web attack classification. The study conducted an empirical evaluation of different feature selection techniques and classification models for web attack classification. Through a series of experiments, the performance of various methods was assessed based on classification accuracy, interpretability, and computational efficiency. The results indicate that feature selection techniques such as Recursive Feature Elimination (RFE) and Chi-square test were effective in identifying relevant features for web attack classification. By selecting a subset of informative features, these techniques improved model performance and reduced computational overhead [30].

4.1. Evaluation metrics and result analysis

To determine the categorization process' average performance. The labeled data set is divided into training and testing at random. Once the classifier has been trained, one subset is picked to serve as the test set. First, four commonly utilized metrics —precision, recall, accuracy, and F-1 score are used to assess how well a phishing detection system performs.

Precision is calculated by dividing the total number of positive examples predicted by the classifier (i.e., the sum of true positives and false positives) by the number of true positives (i.e., the instances expected to be positive by the classifier that are in fact positive). It is, in other words, the proportion of all successfully categorized positive examples to all positive examples predicted by the classifier. Therefore, a low precision means that the system is producing a lot more false alarms than true issue cases that are being successfully detected. Equation 1 can be used to calculate precision.

$$\text{Precision} = \frac{Tp}{Tp + Fp} \tag{1}$$

On the other hand, recall, or true positive rate (TPR), refers to the proportion of accurate positive predictions made by the classifier out of all the positive cases in the dataset. The quantity of true positives divided by the total of true positives and false negatives is another way to put it. A low recall means that the classifier misclassifies many real problematic cases as benign. Equation 2 can be utilized to determine recall.

$$\text{Recall} = \frac{Tp}{TN + Fn} \tag{2}$$

A classifier's **accuracy** may be computed using equation 3 and is a very basic indicator of how well it is performing by calculating the total number of examples that are correctly classified across all classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + Fp + FN} \tag{3}$$

The F-1 score combines Precision and Recall, demonstrating the classifier's capacity to forecast both positive and negative classes. The formula for the F-1 score is given in equation 4.

$$\text{F} - \text{measure} = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

Support is the proportion of actual occurrences of the class in the dataset. Unbalanced support in the training data, which may imply structural issues with the classifier's reported scores, may suggest the need for stratified sampling or rebalancing. Support is the same across models, but diagnosis is on the evaluation process.

Precision, recall, and F1-score are crucial for assessing model performance in an unbalanced dataset scenario because they take into account the model's capacity to accurately identify the minority class (phishing websites) while avoiding incorrect classifications. Compared to standard accuracy, these measures provide a more thorough and nuanced knowledge of a model's efficacy, guaranteeing that intrusion detection systems can identify and lessen the impact of assaults such as phishing.

4.1.1. Feature selection analysis

Feature selection is an essential process in machine learning that aims to enhance model performance by identifying the most significant features for the task. In this analysis, two methods were

employed: Recursive Feature Elimination (RFE) and Chi-Square Test.

Recursive Feature Elimination (RFE): RFE is an iterative approach that fits a model and removes the least important features until the desired number of features is achieved. The features selected by RFE are:

- Short URL: Indicates whether a URL is shortened, which is often used in phishing to disguise the true destination.
- Prefix Suffix-: Checks for hyphens in the domain name, commonly found in phishing URLs.
- HTTPS: Determines if the website uses HTTPS, as phishing sites may lack secure connections.
- Anchor URL: Analyzes the proportion of URLs in anchor tags that link to different domains, with higher proportions potentially indicating phishing.
- Links In Script Tags: Considers the number of links within script tags, which can be used for malicious purposes.
- Server Form Handler: Examines if the form data is handled securely, a key indicator of phishing.
- Website Forwarding: Detects if the website forwards to other URLs, a common phishing tactic.
- WebsiteTraffic: Assesses the amount of traffic, with legitimate sites generally having more traffic.
- Google Index: Checks if the site is indexed by Google, as phishing sites might not be.
- Links Pointing to Page: Evaluates the number of links pointing to the webpage, with fewer links often seen in phishing sites.

Chi-Square Test: The Chi-Square Test is used for categorical data to evaluate the relationship between features and the target variable. The features selected by Chi-Square Test are:

- Prefix Suffix-: Same as in RFE, indicating its significance.
- Sub Domains: Evaluates the number of subdomains, with more subdomains potentially indicating phishing.
- HTTPS: Also selected in RFE, reinforcing its importance.
- Domain Reg Len: Considers the length of domain registration, with shorter periods sometimes linked to phishing.
- Request URL: Looks at the percentage of URLs that request external resources, which can indicate phishing.
- Anchor URL: Also, in RFE, highlighting its relevance.
- Links In Script Tags: Selected in both methods, showing its importance.
- Server Form Handler: Commonly selected, indicating its significance in detecting phishing.
- Website Traffic: Again selected, reinforcing its relevance.
- PageRank: Considers the page rank of the site, with lower ranks possibly indicating phishing.

The analysis highlights that feature selection methods like RFE and Chi-Square Test are effective in identifying relevant features for phishing detection. Both methods highlighted important features such as HTTPS, Anchor URL, Links in Script Tags, and Server Form Handler, underscoring their significance in distinguishing phishing websites from legitimate ones.

4.1.2. Model evaluation

Three machine learning models were evaluated: Logistic Regression, Decision Tree, and Random Forest. Each model's performance was measured using accuracy and a detailed classification report comprising precision, recall, and F1-score.

For phishing detection, Decision Tree, Logistic Regression, and Random Forest classifiers were employed. The performance results of the suggested system are shown in Fig. 3.

The Logistic Regression model demonstrated strong performance with an accuracy of 92.63 %, indicating that it correctly predicted the legitimacy of websites in approximately 92.63 % of instances. The model exhibited balanced precision and recall metrics, with a precision of 0.92 for legitimate websites (class −1) and 0.93 for phishing websites (class 1). This means that 92 % of the legitimate websites and 93 % of the phishing websites predicted by the model were correct. The recall scores were similarly robust, with 91 % of actual legitimate websites and 94 % of actual phishing websites being correctly identified. The F1-scores, which balance precision and recall, were 0.92 for legitimate websites and 0.93 for phishing websites, indicating a well-rounded performance for this classification task.

The Decision Tree model outperformed Logistic Regression with an accuracy of 95.57 %. Its precision for class −1 (legitimate) is 0.95 and for class 1 (phishing) is 0.96, demonstrating improved precision over Logistic Regression. The recall rates are 0.95 for legitimate websites and 0.96 for phishing websites, indicating the model's high sensitivity and ability to correctly identify both types of websites. The F1-scores for the Decision Tree are 0.95 for legitimate websites and 0.96 for phishing websites,

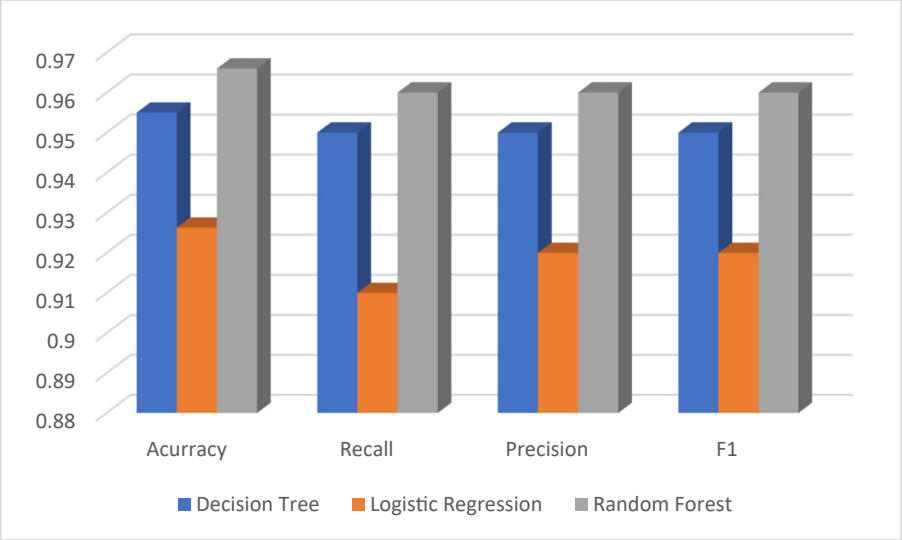

Fig. 3. Phishing detection system results.

highlighting the model's robust performance. The higher accuracy and F1-scores compared to Logistic Regression suggest that the Decision Tree model is more effective in this classification task.

The Random Forest model achieved remarkable performance in distinguishing between legitimate and phishing websites, with an accuracy of 96.70 %. This indicates that the model correctly predicted the legitimacy of websites in nearly all instances. The precision was 0.96 for legitimate websites and 0.97 for phishing websites, demonstrating the model's high accuracy in identifying true positives while minimizing false positives. The recall scores were equally impressive at 0.96 for legitimate and 0.97 for phishing websites, highlighting the model's effectiveness in capturing the vast majority of actual cases. The F1-scores, which balance precision and recall, were 0.96 for legitimate and 0.97 for phishing websites, reflecting a robust performance across both classes. Additionally, the macro and weighted averages for precision, recall, and F1-scores were 0.97, underscoring the model's balanced and superior performance across all metrics. This comprehensive performance makes Random Forest the most reliable and effective model for phishing detection in this analysis.

Logistic Regression, Decision Tree, and Random Forest were evaluated for their effectiveness in classifying web attacks. Logistic Regression demonstrated simplicity and interpretability, making it a suitable choice for scenarios where model transparency is crucial. Decision Tree models offered intuitive decision rules and were capable of capturing non-linear relationships between features and the target variable. However, Decision Trees were prone to overfitting, especially when the tree depth was not properly constrained. Random Forest, on the other hand, mitigated overfitting by aggregating the predictions of multiple trees, leading to improved generalization performance and robustness. Overall, the study's findings suggest that a combination of feature selection techniques and ensemble learning models such as Random Forest can yield the best results for web attack classification. By selecting informative features and leveraging the collective intelligence of multiple trees, researchers can build more accurate and robust classification models capable of effectively detecting and mitigating web-based threats. Additionally, the study highlights the importance of interpretability, computational efficiency, and scalability in designing practical cybersecurity solutions for real-world applications.

5. Conclusions and future works

In conclusion, the empirical evaluation of feature selection techniques and classification models for web attack classification provides valuable insights into the effectiveness of different approaches in addressing cybersecurity challenges. Through systematic experimentation and analysis, the study demonstrates the importance of feature selection in improving classification performance and reducing computational complexity. Techniques such as Recursive Feature Elimination (RFE) and the Chi-square test enable the identification of relevant features. However, the dataset used in this study, its size, structure, and specific characteristics were not discussed, which limits the context and

generalizability of the findings. A detailed analysis of the dataset would have provided a clearer understanding of how these feature selection techniques and classification models perform in different scenarios, especially when applied to other types of web-based attacks or datasets with varying features.

Moreover, the evaluation of classification models highlights the strengths and limitations of various algorithms in classifying web attacks. Logistic Regression offers simplicity and interpretability, making it suitable for scenarios where model transparency is paramount. Decision Tree models provide intuitive decision rules but are susceptible to overfitting, especially with complex datasets. Random Forest, an ensemble learning algorithm, addresses the overfitting issue by aggregating the predictions of multiple trees, leading to improved generalization performance and robustness. Although these three classifiers were evaluated, the study did not consider other models such as Support Vector Machines (SVMs) or Neural Networks, which could have potentially enhanced the evaluation. Including a broader set of classifiers could provide a more comprehensive understanding of the model's generalizability and the potential of more complex, nonlinear models in detecting web attacks.

The findings of this study have practical implications for cybersecurity practitioners and researchers, providing guidance on selecting appropriate feature selection techniques and classification models for web attack detection and mitigation. By leveraging the insights gained from this research, organizations can enhance their cybersecurity posture and effectively defend against a wide range of web-based threats. However, it is important to acknowledge the limitations of this study. The Chi-Square test and Recursive Feature Elimination (RFE) are both useful feature selection methods; however, each has advantages and disadvantages. Though computationally costly and requiring a model to assess feature significance, RFE is a flexible technique that works well with complicated models and captures feature interactions. However, the Chi-Square test implies independence and ignores feature interactions, while being quick, easy, and effective with categorical information. The model being utilized, the computing resources available, and the data properties all influence the choice of feature selection technique. The study chose RFE and Chi-Square due to their ability to handle high-dimensional feature spaces and their straightforward interpretability, but alternatives such as Principal Component Analysis (PCA) or Mutual Information could have been explored to better handle feature correlations or provide additional perspectives on feature importance. A detailed discussion of the reasons for choosing RFE and Chi-Square over these alternatives would provide better insights into their advantages for this specific application.

The evaluation was conducted on a specific dataset, and the results may vary with different datasets and real-world scenarios. The computational feasibility of models for real-time use was not addressed in the current study, which is an important factor for deployment in cybersecurity systems. While the models performed well during training and evaluation, their performance in a real-time environment—especially under dynamic and evolving attack conditions—requires further investigation. This includes evaluating computational complexity, response time, and scalability to ensure they can operate efficiently in real-world settings.

In conclusion, the suggested model has a number of drawbacks that may affect its functionality and suitability for use in actual cybersecurity situations, even if it presents a viable strategy for categorizing online assaults using well-known classifiers and feature selection techniques. These drawbacks underscore the necessity for more improvement, which includes investigating more sophisticated feature selection strategies, improving model generalization, and resolving issues with scalability and interpretability. Future research should explore the generalizability of the findings across diverse datasets and investigate advanced feature selection methods and ensemble learning techniques to further improve classification performance. There are several ways to include innovative and cutting-edge techniques that would greatly improve the state of the art in intrusion detection systems, even though the study provides a helpful comparison of conventional feature selection techniques and classification models. The study may be able to better address the changing terrain of web-based assaults by integrating deep learning methods, hybrid models, real-time flexibility, and more sophisticated feature selection algorithms. These developments would increase IDS systems' resilience and robustness in real-world situations in addition to improving detection efficiency and accuracy. Future work could investigate the development and application of more advanced feature selection methods and ensemble learning techniques to further optimize classification performance. These approaches have the potential to enhance the robustness and accuracy of models, particularly in the context of

complex, high-dimensional datasets. Moreover, exploring hybrid models that integrate multiple feature selection strategies and ensemble methods could offer promising solutions to address the dynamic and ever-evolving challenges in cybersecurity, such as the detection of novel threats, reduction of false positives, and improved adaptability to emerging attack strategies.

Source of Funding

This work receives no funding.

Conflict of interest

Authors declare no conflict of interest.

Ethical approval

All protocols adopted follows standard ethical procedure.

Data availability

No other dataset to declare.

Author contribution

All authors contributed equally to the manuscript.

Acknowledgement

Authors wish to acknowledge personnel at the herbarium of the Plant Biology Department, University of Ilorin, Nigeria for the voucher authentication.

References

[1] Report K. Cybercriminals attack users with 400,000 new malicious files daily – that is 5% more than in 2021 [Internet]. 2022. Available from: https://www.kaspersky.com/about/press-releases/cybercriminals-attack-users-with-400000-new-malicious-files-daily-that-is-5-more-than-in-2021.
[2] The Hacker News [Internet]. The hacker news. 2023. Available from: https://thehackernews.com/.
[3] Ravale U, Marathe N, Padiya P. Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function. Procedia Comput Sci 2015;45:428–35.
[4] Chen CM, Chen YL, Lin HC. An efficient network intrusion detection. Comput Commun 2010;33(4):477–84.
[5] Ashiku L, Dagli C. Network intrusion detection system using deep learning. Procedia Comput Sci 2021;185:239–47.
[6] Shams EA, Rizaner A. A novel support vector machine based intrusion detection system for mobile ad hoc networks. Wirel Netw 2018;24:1821–9.
[7] Al-Zubaidie M, Zhang Z, Zhang J. RAMHU: a new robust lightweight scheme for mutual users authentication in healthcare applications. Secur Commun Network 2019;2019(1):3263902.
[8] Farnaaz N, Jabbar MA. Random forest modeling for network intrusion detection system. Procedia Comput Sci 2016;89:213–7.
[9] Egea S, Rego Mañez A, Carro B, Sánchez-Esguevillas A, Lloret J. Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments. IEEE Internet Things J June 2018;5(3):1616–24. https://doi.org/10.1109/JIOT.2017.2787959.
[10] Meidan Y, Bohadana M, Shocher A, Oren Y, Ovadia Y, Shabtai A, et al. N-BaIoT—network-Based detection of IoT botnet attacks using deep autoencoders. IEEE Pervasive Computing Jul.-Sep. 2018;17(3):12–22. https://doi.org/10.1109/MPRV.2018.03367731.
[11] Khammassi C, Krichen S. A GA-LR wrapper approach for feature selection in network intrusion detection. Comput Secur 2017;70:255–77.
[12] Osanaiye O, Cai H, Choo K-KR, Dehghantanha A, Xu Z, Dlodlo M. Ensemble-based multi-fliter feature selection method for DDOS detection in cloud computing. EURASIP J Wirel Commun Netw 2016;2016(1):130.
[13] Ingre B, Yadav A. Performance analysis of NSL-KDD dataset using ANN. In: 2015 international conference on signal processing and communication engineering systems. IEEE; 2015. p. 92–6.
[14] Janarthanan T, Zargari S. Feature selection in UNSW-NB15 and KDDCUP'99 datasets. In: 2017 IEEE 26th international symposium on industrial electronics (ISIE). IEEE; 2017. p. 1881–6.
[15] Khan NM, Negi A, Thaseen IS, Anwar W, Singh P, Sharma R, et al. Analysis on improving the performance of machine learning models using feature selection technique. In: International conference on intelligent systems design and applications. Springer; 2018. p. 69–77.
[16] Kumar V, Sinha D, Das AK, Pandey SC, Goswami RT. An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset. Clust Comput 2020;23(2):1397–418.
[17] Yin Y, Jang-Jaccard J, Xu W, Singh A, Zhu J, Sabrina F, et al. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. J Big Data 2023;10(1):15.
[18] Osanaiye O, Choo KKR, Dlodlo M. Analysing feature selection and classification techniques for DDoS detection in cloud. In: Proceedings of Southern Africa telecommunication; 2016. p. 198–203.
[19] Sankaran A, Vatsa M, Singh R, Majumdar A. Group sparse autoencoder. Image Vis Comput 2017;60:64–74.
[20] Tama BA, Lim S. Ensemble learning for intrusion detection systems: a systematic mapping study and cross-benchmark evaluation. Comput Sci Rev 2021;39:100357.
[21] Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D. Adversarial attacks and defences: a survey. arXiv preprint arXiv:181000069 2018.
[22] Abu Bakar R, Huang X, Javed MS, Hussain S, Majeed MF. An intelligent agent-based detection system for DDoS attacks using automatic feature extraction and selection. Sensors 2023;23(6):3333.
[23] Ninu SB. An intrusion detection system using exponential Henry gas solubility optimization based deep neuro fuzzy network in MANET. Eng Appl Artif Intell 2023;123:105969.
[24] Mathebula Solani D. Biochemical changes in diabetic retinopathy triggered by hyperglycaemia: a review. Aveh Journal 2017.
[25] Chawla A, Chawla R, Jaggi S. Microvascular and macrovascular complications in diabetes mellitus: distinct or continu? NCBI 2016.
[26] Maza S, Touahria M. Feature selection algorithms in intrusion detection system: a survey. KSII Trans Internet Inform Syst (TIIS) 2018;12(10):5079–99.
[27] Karimi F, Sadoghi Yazdi H, Abasi AK, Safavi AA, Abbasi M, Bahrami A, et al. SemiACO: a semi-supervised feature selection based on ant colony optimization. Expert Syst Appl 2023.

[28] Alawad NA, Abed-Alguni BH, Al-Betar MA, Jaradat A. Binary improved white shark algorithm for intrusion detection systems. Neural Comput Appl 2023;35(26):19427−51. https://doi.org/10.1007/s00521-023-08772-x.
[29] Abed-alguni B, Al-Betar MA, Abualigah L, Abd Elaziz M, Mirjalili S, Alawad NA, et al. Opposition-based sine cosine optimizer utilizing refraction learning and variable neighborhood search for feature selection. Appl Intell 2023;53:13224−60. https://doi.org/10.1007/s10489-022-04201-z.
[30] Prashanth SK, Iqbal H, Illuri B. An enhanced grey wolf optimisation−deterministic convolutional neural network (GWO−DCNN) model-based IDS in MANET. J Inf Knowl Manag 2023;22(4):2350010.