University of Ibadan
Journal of Science and Logics in ICT Research

Comparative Performance Evaluation of Random Forest on Web-based Attacks

Oluwaseye Abayomi Adeyemi[1], Azeez Ajani Waheed[2] and Ogunsanwo Olajide Damilola[3]

[1]*Department of Health Sciences, Bamidele Olumilua University of Education, Science and Technology, Ikere, Ekiti – State*
[2]*Department of Mathematics, Lead City University, Ibadan, Oyo – State*
[3]*Department of Computer Science, Lead City University, Ibadan, Oyo – State*
sheyee4u2@gmail.com, waheed.azeez@lcu.edu.ng, ogunsanwo.olajide@lcu.edu.ng

Abstract
The majority of typical online attack methods are thoroughly researched and documented. Countries, corporations, people, and vital infrastructures that depend on information technology for daily operations have suffered financial losses, the loss of personal information, and economic harm as a result of web-based intrusion. However, foreseeing an attack before it happens can aid in its prevention. This research proposes a predictive model for web-based attacks and a performance comparison of random forest with and without feature selection to secure the availability, integrity, and secrecy of networks, computer systems, and their data. The CIC-Bell-IDS2017 dataset, which includes typical and contemporary intrusion attacks, served as the raw data source for the proposed model. A python-based programming environment and interface for Anaconda Navigator, Jupyter Notebook, was used to create the predictive models. Performance evaluation and comparative analysis were conducted, and the results demonstrate that, once big data analytics (feature scaling and feature selection) were applied to the dataset, the models' prediction accuracies improved, creating a potential intrusion detection system. The outcome yielded excellent accuracy and model development times in both cases, with 97% and 98% precision for both sets and model development times of 35 seconds for the raw set and 15 seconds for the reduced set, which is an important factor when deploying machine learning models in a real-time setting. Random Forest is more computationally expensive than Correlation feature Selection-based classifiers, but having higher predictive accuracy, according to a comparison. Both of these methods work well and each has advantages and disadvantages. The use of big data analytics (PySpark) was found to help machine learning models perform better, resulting in better intrusion detection system.

Keywords: *Web Based Attacks, Random Forest, Correlation Feature Selection, machine learning*

1. Introduction

Information Technology is becoming a significant part of our daily life, and one used to a great extent, to help solve difficulties since we live in a dynamic and fast changing environment. Web-based systems and applications have always been a significant part of this Information Technology [1]. Generally, Information Technology is becoming more crucial now that many of critical infrastructures, i.e. health sector, banking sector, business sector, commutations sector, and networking are depending on them. The World Wide Web is expanding daily and the internet has become a necessity for everyone. The use of the internet by individuals, academia, government, business and organizations across a variety of industries, has drastically expanded during the past decade [1]. Internet information resources are rapidly expanding and are now present in many aspects of daily life. Web applications can provide excellent digital experiences, but only those that are secure can properly deliver this service [2].

Web-based systems provide organization and businesses with quick and easy operation through digitalization and automation of process. However, web-based applications vulnerabilities in coding and databases can now be exploited by attackers to gain illegal access to user system and personal information [2]. Attackers now target web applications explicitly

Oluwaseye Abayomi Adeyemi, Azeez Ajani Waheed and Ogunsanwo Olajide Damilola (2024). Comparative Performance Evaluation of Random Forest on Web-based Attacks, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 11 No. 2, pp. 14 – 23

and create variety of malicious web content in an effort to obtain information from a specific company or individual who is in possession of valuable information [3].

Even though most of these organizations install firewall, antimalware, and other traditional security systems, they are still unable to stop the vast majority of Web attacks. The Identification and prevention of various forms of cyber-attacks are becoming more difficult due to the limits of traditional security methods. It is getting harder to secure the integrity confidentiality and availability of web-based systems even though users rely more and more on them.

Due to the sensitive, valuable, and sufficient information collected from users and held by these systems, web-based systems and applications are frequently the target of hackers who want to steal information, make money, or engage in other illegal actions, leading to disastrous effects on the finances and reputation of system users [3]. Web-based attacks are transmitted as internet and web-based system usage grows. Between 2020 and 2021, the number of fraudulent web application requests increased by 88% with more than 75% been injection and broken access control attacks. Banking and finance sector, along with Software as a Service providers, were the most frequently attacked sectors in 2021, combining for more than 28% of web-based attacks.

Information-Technology Promotion Agency (IPA) estimates that more than 75% of attacks are now targeted against web applications [4]. And Over 80% of online applications on the internet have at least one significant vulnerability [5]. According to the 2017 Internet Security Treat Report (ISTR) more than 76% of the websites that were scanned were deemed to be vulnerable [4].

This vulnerability is of various categories; however, web-based attacks mostly employ SQL injection and Cross-site scripting. SQL Injection is a type of web attack that target database-driven website and involve an intruder inserting malicious SQL queries into the database using data sent from the client to the server [4]. These attacks have grown as more web applications are made available in the cloud, posing a serious danger to web-based services and various web application programs. Cross-site scripting is a form of attack in which the victims' web browser is used to run or download malicious script from remote online pages [5].

Due to the increase in web-based applications and systems and the inability of traditional system to defend against this attack efficiently, attackers are now focusing more on exploiting web-based vulnerability. However, machine learning techniques which have been employed by several researcher/study in the field of Cybersecurity and Data mining for the task of predicting attack before they occur based on knowledge acquired from data [6]. This approach can also be used against web-attack by predicting this attack before they actually occur. In terms of web-based attack security, intrusion detection methodology is still relatively new. IDS are generally made to monitor and find intrusive online activity. However, network-based assaults differ significantly from web-based attacks in terms of their properties [7].

Machine learning predictive models learn by looking for patterns in a set of input data. It uses classification algorithm to classify data and foretell future events. It is an essential part of predictive analytics, a sort of data analytics that makes use of both recent and old data to predict activity, behavior, and trends [5]. Predictive modeling is a statistical method that uses data mining and machine learning to predict and anticipate likely future outcomes using historical and existing data [6].

Attackers carefully target organizations by exploring its vulnerabilities and infiltrating their network and control systems using many different types of attack: Trojan horse, Viruses, Worms, Ransomware, Man-in-the-Middle Attack, Denial-of-Service Attack (DoS), Distributed Denial-of-Service Attack (DDoS), SQL Injection, Cross-Site Scripting, User/Root Access Compromise, Phishing Attacks and Zero-Day-Attack [6]. They are indeed ready to incur great costs, time, and expertise in order to accomplish their goals.

Users have belief that the private and secure handling of their sensitive personal information on the website like their credit card, social security, medical information becomes public

due to intrusion in the form of Web- based attacks with potentially serious repercussions. Cyber-intrusion and cyber-security continue to be a serious issue for any sector in the cyberspace as a number of security breaches keep increasing, and modern attacks keep evading traditional Cybersecurity procedures and blacklist approach by using impressive highly sophisticated techniques, they can be difficult or even impossible to detect even though most of these attacks are variants of previously known attacks with known signatures. It is known that thousands of zero-day attacks are continuously emerging because of the addition of various protocols, mainly from the field of Internet of Things (IoT).

In 2020 Cybersecurity Ventures predict that global cybercrime costs will grow by 15% each year for the next five years, reaching at least 10.5 trillion USD annually by 2025, up from $3 trillion USD in 2015 [5]. The United State Government alone invested over 16 billion USD for cyber security and defense in the 2019 fiscal year's budget, which increase to almost 19 billion USD in 2020 fiscal year's budget [8]. This now leads to many researchers and study, leveraging different data science techniques and artificial intelligence to create anomaly-based IDSs that can detect unknown attacks, using different architecture, methods, approaches, and algorithms such as statistics, data mining, machine learning techniques, advance analytics, and hybridization of system.

Therefore, this study proposes a predictive model for intrusion detection, that uses Data Analytics for feature scaling and feature selection, in order to select the most relevant features and to improve the model performance, and machine learning techniques for classification and prediction of attack.

2. Related Works

This section discusses the literature utilized in this culminating study. The literature spotlights knowledge gaps for discussion and techniques seen in later section.

Alongside the rise in cyberattacks, the internet and web-based applications have been expanding quickly. Requests that are seen as normal or odd are used to carry out these assaults (attack requests). As a result, an intrusion attempt could be a classification issue. In order to improve the security of web services, machine learning algorithms are employed to train models to categorize this request.

An organization's network can be attacked by hackers through unprotected Web applications. According to statistics, 42% of web apps are vulnerable to threats and hackers. Although the widespread use of web-based apps and the emphasis on data storage on the internet have been productive and beneficial, they have also made the system's flaws more obvious [1].

Due to the ongoing increase in web threats, web application security is currently one of the most important challenges in information security. Over 76% of the websites scanned were determined to be vulnerable, according to the Internet Security Treat Report (ISTR) 2017. Additionally, according to the research, there were 35% more web-based security breaches in the first quarter of 2017 than there were in the same period in 2016 [5]. Researchers have suggested many Intrusion Detection strategies over the years to deal with the complexity and number of threats to computer systems that have grown over time. In the domain of behavior-based intrusion detection systems, Random Forest models have been delivering a noteworthy performance on their applications. Classification, feature choice, and proximity metrics are provided using particulars of the Random Forest model [8].

In order to discriminate between regular and abnormal traffic, cleaned and labeled the CSIC HTTP 2010 dataset before conducting the experiments. In order to find missing features for a typical attack, the data was finely preprocessed using a Python script. Additionally, by using various Machine Learning classifiers like J48, Naive Bayes, OneR, and Decision tables that use evaluation metrics to find the accuracy using Weka 3.8, feature extraction. Dataset played a key role in identifying malicious behavior and the attack types like SQL injection (SQLi), Cross-Site Scripting (XSS), and Buffer Overflow [24]. Additionally, 20 features were extracted with enhanced web-based attack detection thanks to the use of fine-tuned feature set engineering, raising the true positive rate. Last but not least,

the J48 decision tree algorithm was shown to be the top performing algorithm with the best attack detection rate of 94.5% in testing results using three machine learning algorithms (J48, Naive Bayes, and OneR).

This study reviewed existing literatures to understand predictive models, cyberattacks, web attacks, machine learning techniques and algorithms. The reviews show that making use of data science, data mining, and machine learning has been directed towards the development for predictive models for different types of cyberattacks including web-based attack by several researchers. It also shows the different dataset and classification algorithm, mostly used for predictive models and their merit and demerit. Furthermore the researchers also called for further researches on the subjects. This study make use of a fraction of CIC-Bell-IDS2017 dataset that contains only web attack to create a predictive model specifically for those types of attacks, instead of making use of a general attack dataset like most other study and comparatively evaluate the performance of random forest with or without feature selection.

3. Methodology

This section elaborates the research techniques and method that are used to achieve the stated aim and objectives of the study. This study proposes an architecture that uses Machine Learning Techniques, (Random Forest) and Feature Selection to predict network intrusions and comparatively evaluate the performance of the Random Forest with and without Feature Selection. It also intends to elaborate on all the phases and stages involved in the development of a predictive model for Web-Based attacks which includes planning, organizing and building up of every stage require making the system model functional. The methods of data collection adopted for this research work is secondary data collection method which is the dataset created by the Canadian Institute of Cybersecurity (CIC) and Bell Canada (BC) Cyber Threat Intelligence (CTI). The experiments were done on a 64-bitWindows7 operating system with 4GB of RAM and a Intel Pentium (R) Dual core CPU at 1.90GHz per core.

The step by step guide taking cognizance of the research aims and objectives is summarized in a work process diagram/conceptual framework/system architecture in Figure 1.

3.1 Data Collection and Description of Dataset

This research work used a fraction of a modern big intrusion dataset created by the Canadian Institute of Cybersecurity (CIC) and Bell Canada (BC) Cyber Threat Intelligence (CTI), generally refer to as the CIC-Bell-IDS2017 dataset, which contains common and modern attacks.

This dataset specifically consists of 2,830,743 records created on 8 files, each of which has 78 unique attributes and a label. For this study the fraction of this dataset which contains 170,366 records of benign and common web attacks such as Cross-site scripting (XSS), Brute Force and SQL Injection was used to train and test the machine learning model for predicting web-based attacks.

3.2 Pre-Processing / Data Cleaning

This is the process of removing irrelevant or superfluous information from data and keep only the most crucial and significant information [3]. The most time-consuming and crucial phase in data mining is data pretreatment. Realistic data can be noisy, redundant, partial, and inconsistent and is frequently derived from diverse platforms.

Therefore, the preprocessing/data cleaning stage in this study includes Dropping Unwanted Column, Replacing NaN value or Whitespace, removing infinite value, and Label Encoding (for binary classification where all attacks are grouped as “Attack” and good-ware as “Benign).

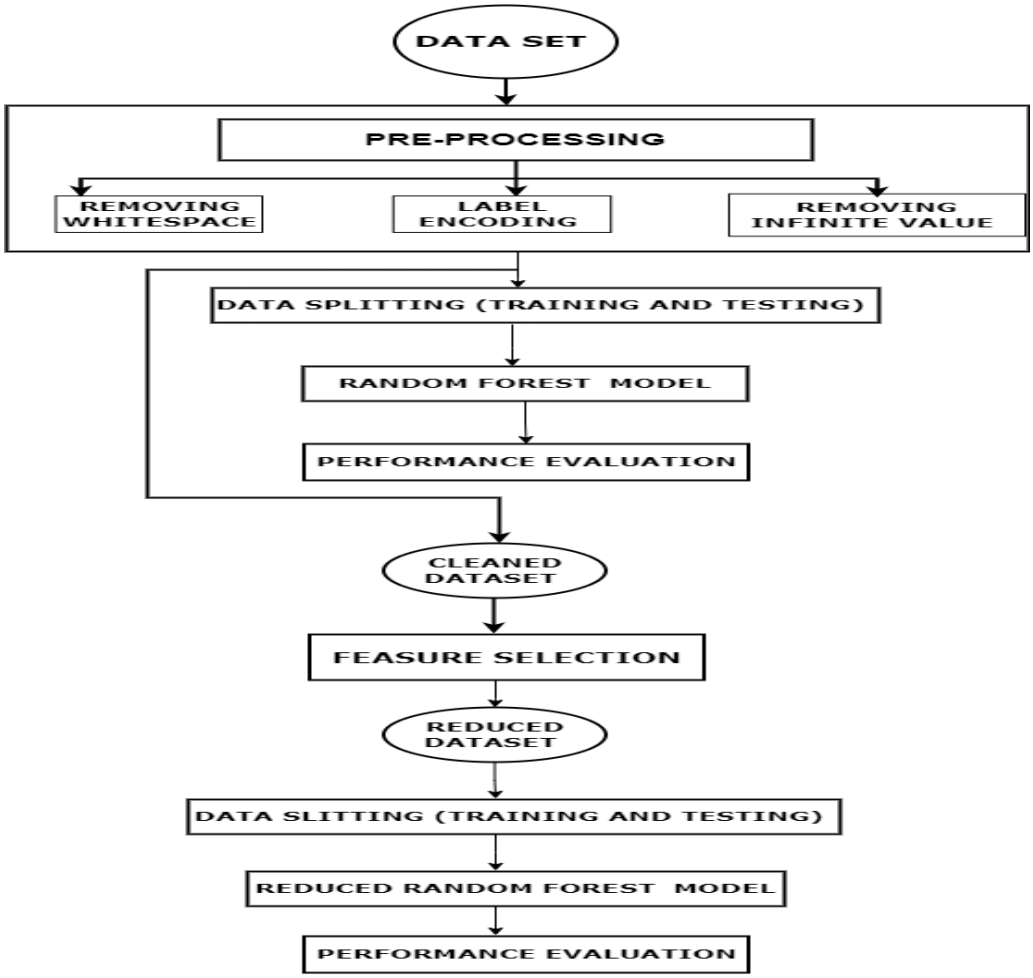

Figure 1: Conceptual Framework

.

Table 1: Distribution Attacks

S/N	ATTACK TYPE	COUNT	DATA TYPE
0	Benign (Good Ware)	149113	Int64
1	Denial of Service (Dos Hulk)	22,380	Int64
2	Denial of Service (Golden Eye)	10,293	Int64
3	Distributed Denial of Service (Ddos)	20,317	Int64
4	Port Scan	11,002	Int64
5	Ftp-Patator	7,938	Int64
6	Ssh-Patator	5,897	Int64
7	Brute Force (Web Attack)	1,470	Int64
8	Cross Site Scripting (Web Attack)	652	Int64
9	Infiltration	36	Int64
10	Sql Injection (Web Attack)	21	Int64
11	Heartbleed	11	Int64

3.3 Data Splitting for Raw Dataset

This is the process of dividing data into two, where a certain percentage was used for training the ML algorithm 80% training and 20% for testing

3.4 Development of Predictive Model for Clean Dataset

At this stage of this work Random Forest that operates by constructing a multiple Decision Trees, is used on the cleaned dataset to generate a predictive model, using data balancing and 80:20 split tests. The model was evaluated using confusion matrix to determine the accuracy, precision, recall, and F-1 score of each model.

Random Forest provides us with guidance on which important features should be kept and which ones should be dropped from the dataset [5]. An assembly of independent decision trees is called Random Forest (RF). Every individual decision tree first categorizes each instance, and the instance is then finally classified by the collective wisdom of all the individual trees [8].

The Gini index is a function that assesses the impureness of data and event uncertainty. By calculating the Gini of each branch on a node using class and probability, this method can predict which branch is most likely to occur. The formula for GINI's is:

Gini (t) = 1- $\sum_{i=1}^{N} P(C_i|t)^2$ or

Gini = 1- $\sum_{i=1}^{N} (P_i)^2$

Another way to describe this is the use of entropy to determine how nodes branch in a decision tree based on the probability of a specific outcome.

Entropy = $\sum_{i=1}^{N} - P_i * log_2(P_i)$

- P and P_i represents the relative frequency of the class.
- T is a condition,
- N the number of classes in the data set, and
- C_i is the i^{th} class label in the data set.

3.5 Description of Evaluation Tool: Confusion Matrix

A confusion matrix, which using true and false detection of the model, the classification accuracy, precision, recall, and F-1 score can be determined. A common structure for evaluating accuracy is the confusion matrix, commonly referred to as the error matrix. It primarily serves as a means of contrasting classification outcomes with actual measured values. In a confusion matrix, it can show the classification findings' accuracy [1].

Table 2: Confusion matrix for binary classification

PARAMETER	ATTACK	BENIGN
Attack	True Positive	False Negative
Benign	False Negative	True Negative

Source:https://www.sciencedirect.com/topics/engineering/confusion-matrix#:~:text=A%20confusion%20matrix%20is%20a,performance%20of%20a%20classification%20algorithm.

- Where True Positive (TP) is the correctly predicted Attack
- True Negative (TN) is the correctly predicted Benign.
- False Negative (FN) is Attack that failed to be identified or predicted as Benign.
- False Positive (FP) are Benign that failed to be identified or predicted as Attack

Accuracy determines the percentage of correctly classified instances. Out of all the samples in the dataset, it is the proportion of properly predicted samples. Given as;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision gives the percentage of true positive instance that are correctly classified. Finding the probability that a positive forecast will come true requires precision. Given as;

$$Precision = \frac{TP}{TP + FP}$$

Recall is used to calculate the model's ability to predict positive value. Given as;

$$Recall = \frac{TP}{TP + FN}$$

F-measure is defined as the harmonic mean of Precision and Recall (when both considered)

$$F\text{-}Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.6 Feature Selection

The process of evaluating each feature individually to decide which ones within the dataset have the most impact on the outcome is known as feature selection [7]. The goal is to reduce the dimensions of high dimensional data while maintaining the same accuracy, if not higher.
The concept of correlation is used to compare two different features. For example, if the features are uncorrelated, the correlation will be zero; if not, it will be 1. To calculate the correlation between the two distinct variables, two complete modules—classical linear correlation and correlation on the basis of information theory—were put into use [9].

Correlation can be calculated via few methods
To calculate the correlation between two variable x and y, using Pearson correlation coefficient:

$$r = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]}{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}$$

- Where r is the correlation coefficient
- x and y are the variables
- $\bar{x}$ is mean of x,
- $\bar{y}$ is the mean of y

NOTE:
- if r is between 0.6 and 1 then a Positive correlation exist
- if r is between -0.6 and -1 then a Negative correlation exist
- if r is 0 then no correlation whatsoever (Neutral)
- if r is closer to 0 than 1 (<=0.5) then weak correlation exit

4. Result

The experiments were done on a 64-bitWindows7 operating system with 4GB of RAM and a Intel Pentium (R) Dual core CPU at 1.90GHz per core. The predictive models were developed with python programming language, using Jupyter Notebook, which is a python-based programming environment and interface for Anaconda Navigator. Anaconda Navigator is a desktop graphical user interface (GUI) that allows users to launch applications and easily manage conda packages, libraries, and environments without using command-line interface. Conda is a package and environment management framework that helps data scientist and programmers to find and install many data science libraries required for machine learning tasks and data analysis.

This section shows the performance of Random Forest (RF) for the task of detecting web attack. In the experiments the fraction of CIC-IDS2017 dataset described in methodology was balanced and used to train and test the model. The model generated used 80% training set and 20% testing set, before and after feature selection. The results of the two different models are shown in Table 3 for the raw dataset and Table 4 for the reduced dataset.

Performance Evaluation without Feature Selection for the Raw Dataset

This section shows the results of the performance analysis (Accuracy, Attack Precision, Attack Recall, Model Development time and Attack F-1 score of Random Forest (RF), on the raw dataset.

Table 3: Random Forest (RF) without Feature Selection for the Raw Dataset

Performance Matrix	Prediction Score
Precision	97%
Recall	96%
F1-Score	98%
Classification Accuracy	94%
Model Development time	35sec

Performance Evaluation with Feature Selection for the Reduced Dataset

This section shows the results of the performance analysis (Accuracy, Precision, Recall, and F-1 score) of Random Forest (RF), for the reduced dataset.

Table 4: Random Forest (RF) with Feature Selection for the Reduced Dataset

Performance Matrix	Prediction Score
Precision	98%
Recall	97%
F1-Score	99%
Classification Accuracy	99%
Model Development time	15 sec

Table 5: Comparative Analysis of the Performance Evaluation for both Raw and Reduced Dataset.

Classifiers	Classification Accuracy	Precision	Recall	F1-Score	Model Development Time
Raw set	94%	100%	96%	98%	35 sec
Reduced set	99%	100%	97%	99%	15 sec

4.1 Discussion of Results

From the results of the experiment carried out, it was discovered that there were improvements in the performance of the Random forest classifier, in all categories of evaluation: Accuracy, Precision, Recall, F1-score, and Model Development Time, after correlation technique was used for feature selection, except for precision that produce 100% for both full and reduced set. A good recall was also generated for both the raw and the reduced dataset, achieving 96% and 97% respectively.

The result also produced a great precision and model development time in both cases, with precision of 100% for both set, and model development time of 35sec for raw set and 15sec for reduced set, which is a vital point in deploying machine learning model in a real-time environment. Furthermore, even with the ensemble learning ability of the Random Forest classifier, the experimental result shows an increase of 5% in the classification accuracy after feature selection, which is a shred of evidence that even the much-celebrated ensemble learning algorithm can be improved with feature selection.

Other areas of comparison between the two approaches include the quantity of features chosen, the length of training and testing, bias toward particular characteristics, and other performance metric. Comparison shows that even though Random Forest has higher predictive accuracy than CFS based classifier, it is computationally expensive. Both of these techniques are suitable with their own merits and demerits.

5. Conclusions

As a result of traditional security systems' repeated failures to identify complex and novel attacks, the security industry has recently come under harsh criticism. However, the Anomaly Intrusion Detection System (AIDS) that employs machine learning methodology is an efficient tool in identifying these attacks and comparing the performance evaluation of Random Forest with and without Feature Selection, as demonstrated in this research. This study has demonstrated how the predictive power of machine learning models can be increased by using the big data analytics tool (PySpark). The findings of the classification algorithms Random Forest for both raw and reduced dataset were examined in terms of their appropriateness for identifying intrusions from a sizable dataset comprising numerous contemporary attacks using the Python programming language. The use of big data analytics (PySpark) was found to help machine learning models perform better, resulting in a better intrusion detection system.

References

[1] Singh A., Kumar A. & Bharti A.K. (2021). Identification and Prevention approaches for Web-based Attacks using Machine Learning Techniques. *International Journal of Creative Research Thoughts (IJCRT)* 2, vol. 9, 4558-4563.

[2] Chowdhurry R., Banerjee P., Deep Dey S., Saha B. & Bandyopadhyay S.K. (2020). A Decision Tree Based Intrusion Detection System For Identification of Malicious Web Based Attacks," *Preprints (*www.preprints.org*)*, vol. 1, 10-15.

[3] Acar G., Huang D.Y., Li F., Narayanan A., & Feamster N. (2018). Web Based Attacks to Discover and Control Local IoT Devices. *In IoT S&P: ACM SIGCOMM*, 29-35 https://doi.org/10.1145/3229565.3229568

[4] Riera T.S., Higuera J.B., Herraiz J.M., & Montalvo J.S. (2022). A New Multi-label Dataset for Web Attacks CAPEC Classification using Machine Learning Techniques. *Journal of Science Direct Computer and Security*, 1-18

[5] Sharif M.H.U. (2022). Web Attacks Analysis and Mitigation Techniques. *International Journal of Engineering Research and Technology (IJERT***),** 23-27, www.ijert.org

[6] Zwilling M., Klien G., Lesjak D., Wiechetek L., Cetin F., & Basim H.N. (2020). CyberSecurity Awareness, Knowledge and Behavior: A Comparative Study. *Journal of Computer Information System*, 1-16

[7] Agarwal N., & Hussain S. Z. (2018). A Closer Look at Intrusion Detection System for Web Applications. *Hindawi Security and Communication Networks***,** 27-31 https://doi.org/10.1155/2018/9601357

[8] Karuparthi B., & Mahesh A. (2022). Comparative Study between Random Forest and Support Vector Machine Algorithm in Classifying Cervical Cancer. *International Journal of Engineering Research & Technology (IJERT)*, Vol. 11 Issue 01, 434-437

[9] Li W. (2022). Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty. *Hindawi Security and Communication Networks*, 9-14 https://doi.org/10.1155/2022/113199

[10] Singh S., Choudhary S.S., & Bhavishya S. (2018). Feature Selection Effects on Classification Algorithms: Laconic description of Machine Learning Algorithms. *International Journal of Engineering Research & Technology (IJERT)*, Vol. 7 Issue 02, 183-185

[11] Al-Shalabi L. (2017). New Feature Selection Algorithm Based on Feature Stability and Correlation. *Journal of Information Technology and Computing*, 1-16 DOI 10.1109/ACCESS.2022.3140209, IEEE Access

[12] Avkurova Z. (2022). Models for Early Web – Attacks Detection and Intruders Identification Based on Fuzzy Logic. *Journal of Science Direct: Procedia Computer Scienec*, 694-699

[13] Avkurova Z., Gnatyuk S., Abduraimova B., Fedushko S., Syerov Y., & Trach O. (2021). Detecting web Attacks using Random Under sampling and Ensemble Learners. *Journal of Big Data*, 1-20 https://doi.org/10.1186/s40537-021-00460-8

[14] M. S. Hasan & M. Nosonovsky (2022). Triboinformatics: machine learning algorithms and data topology methods for tribology. *Surface Innovations* 10(4–5), 229–242, https://doi.org/10.1680/jsuin.22.00027

[15] I. H. Sarker, (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 1-21 https://doi.org/10.1007/s42979-021-00592-

[16] K. Akram, G. Banu, M. Basthikodi & A. R. Faizabadi, (2019). Defense Mechanism Using Multilayered Approach and SQL Injection Methods for Web Based Attacks. *Journal of Emerging Technologies and Innovative Research (JETIR)* Volume 6, Issue 5, 122-129

[17] D. Truong, D. Tran, L. Nguyen, H. Mac, H. A. Tran & T. Bui, (2019). Detecting Web Attacks using Stacked Denoising Autoencoder and Ensemble Learning Methods. *In The Tenth International Symposium on Information and Communication*, 1-6

[18] Y. Pan, F. Sun, Z. Teng, J. White, D. C. Schmidt, J. Staples & L. Krause, (2019). Detecting web attacks with end-to-end deep learning. *Journal of Internet Services and Applications*, 10:16, 1-22 https://doi.org/10.1186/s13174-019-0115-x

[19] R. A. Katole, S. S. Sherekar & V. M. Thakare, (2018). Detection of SQL Injection Attacks by Removing the Parameter Values of SQL Query. *Proceedings of the Second International Conference on Inventive Systems and Control (ICISC)*, 736-741

[20] M. H. AL-Maliki & M. N. Jasim, (2022). Review of SQL injection attacks: Detection, to enhance the security of the website from client-side attacks. *Int. J. Nonlinear Anal. Appl*, 3773-3782 http://dx.doi.org/10.22075/ijnaa.2022.6152

[21] M. Alsaffar, S. Aljaloud, B. A. Mohammed, Z. G. Al-Mekhlafi, T. S. Almurayziq, G. Alshammari & A. Alshammari, (2022). Detection of Web Cross-Site Scripting (XSS) Attacks. *Journal of Electronics*, 1-13 https://doi.org/10.3390/electronics11142212

[22] S. Chavan & S. Tamane, (2021). Enhancement in Cloud Security for Web Application Attacks. *IEEE Xplore*, 91-95

[23] B. Kapoor & B. Nagpal,(2022). Ensemble Modelling for Predicting the Relation between Biopsychosocial Signals and Seizures using the Gradient Boosting Method. *Research Square*, 1-17 DOI: https://doi.org/10.21203/rs.3.rs-1810072/v1

[24] S. Sharma, P. Zavarsky & S. Butakov, (2020). Machine Learning based Intrusion Detection System for Web-Based Attacks. *IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS),*1-4 DOI 10.1109/BigDataSecurity-HPSC-IDS49724.2020.00048

[25] A. K. Keshri, A. Sharma, A. Chowdhury, S. S. Rawat & K. Kiran, (2022). SQL – Attacks, Modes, Prevention. *International*

Journal of Research in Engineering, Science and Management, Volume 5, Issue 1, 162-165

[26] G. Battineni, G. G. Sagaro, N. Chinatalapudi & F. Amenta, (2020). Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *Journal of Personalized Medicine*, volume 10, issue 21, 1-11 doi:10.3390/jpm10020021

[27] A. M. Vartouni, S. S. Kashi & M. Teshnehlab, (2018). An Anomaly Detection Method to Detect Web Attacks Using Stacked Auto-Encoder. *6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, 131-134

[28] A. N. Nazarov, D. V. Pantiukhin, I. M. Voronkov & M. A. Nazarov, (2020). Approach To Intelligent Monitoring Of Cyber Attacks. *Synchroinfo Journal*, No. 6, 1-8 DOI: 10.36724/2664-066X-2020-6-6-2-9

[29] J. Diaz-Verdejo, J. Munoz-Calle, A. E. Alonso, R. E. Alonso & G. Madinabeitia, (2022). On the Detection Capabilities of Signature-Based Intrusion Detection Systems in the Context of Web Attacks. *Applied Sciences*, 12, 852, 1-16 http://doi.org/10.3390/app12020852

[30] M. Indushree, M. Kaur, R. Manish, R. Shashihara & L. Heung-No, (2022). Cross Channel Scripting and Code Injection Attacks on Web and Cloud-Based Applications: A Comprehensive Review. *Journal of Sensors*, 22, 1959, 1-20 https://doi.org/10.3390/s22051959