Genetic Optimization Techniques for Enhancing Web Attacks Classification in Machine Learning

Ameera Jaradat, Ahmad Nasayreh, Qais Al-Na'Amneh, Hasan Gharaibeh and Rabia Al Mamlook

October 1, 2023

Genetic Optimization Techniques for Enhancing Web Attacks Classification in Machine Learning

1st Ameera S. Jaradat
Dep. of Computer Science
Yarmouk University
Irbid, Jordan
ameera@yu.edu.jo

2nd Ahmad Nasayreh
Dep. of Computer Science
Yarmouk University
Irbid 211633, Jordan
nasayrahahmad@gmail.com

3rd Qais Al-Na'amneh
Dep. of Cyber Security and Cloud Computing
Applied Science Private University
Amman ,Jordan
q_naamneh@asu.edu.jo

4th Hasan Gharaibeh
Dep. of Computer Science
Yarmouk University
Irbid 211633, Jordan
hasangharaibeh87@gmail.com

5th Rabia Emhamed Al Mamlook
Dep. of Business Administration
Trine University,Indiana, USA
Department of Industrial Engineering
University Zawia Tripoli, Libia
almamlookr@trine.edu

***Abstract*—Web-based applications are now the preferred approach for delivering a variety of services via the Internet. As a result of the globalization of commerce, web applications have been growing quickly and becoming increasingly complicated. Such applications have a significant security vulnerability in the online environment since they were developed with little experience and without testing or validation. Web application vulnerability is an issue brought on by the way the program was created. Numerous attackers use this security vulnerability to take control of the program, modify the data, and steal the most crucial information. They may also access all internal, unauthorized items. In this study, we present a hybrid model that classifies website attacks as benign through the integration of four gradient machine learning algorithms: Boost (GB), Multi-Layer Perceptron (MLP), and Boost. The study employed optimization algorithms such as K Nearest Neighbor (KNN), Logistic Regression, and Genetic algorithm (GA) to extract the optimal parameter. The model underwent evaluation utilizing a data set from the Canadian Institute 2023 that contains various types of attacks on the Internet of Things. Among these algorithms, GB achieved the best accuracy, with accuracy scores of 95%, and a score of 94% and 95% for accuracy, recall and F1-score, respectively.**

***Index Terms*—Machine learning; Genetic Algorithm Optimization; Web applications, web vulnerability, web attack**

I. Introduction

The importance of the security system increases with the growth of the number of Internet users. Therefore, web application security is essential to protect information, customers, and companies from information theft, interference with trade progress, and other malicious cybercrime actions [1]. Cyberattacks can affect any website on the Internet. Among the threats is human error in sophisticated cyberattacks carried out by an organized group of criminals. According to Verizon's Data Breach Investigation Report, the primary motivation for cybercriminals is financial [2]. We are vulnerable to cyberattacks whether we are running an e-commerce website, or a simple corporate website and many different types of attacks are spread around, it becomes difficult to defend against them all. However, there is a lot that can be done to protect websites from these attacks and reduce the possibility of them being targeted by dangerous hackers. Web application security and protection approaches strive to ensure application security using measures such as WAFs, multi-factor authentication for clients, user security, and threat approval to preserve client states. In this paper, we are going to discuss the various most important attacks affecting web applications such as SQL injection attack, brute force attacks, dictionary attack, browser hijacking, backdoor malware, load attack and command injection, therefore, web developers should include security measures like input validation, output encryption, WAF, secure encryption methods, strong authentication schemes, and constantly update their online applications to address any known vulnerabilities. Where we intend to use machine learning algorithms to detect these attacks for what these algorithms have achieved in [3] and among these algorithms are gradient boost (GB), multilayer perceptron (MLP), logistic regression (LR) and K nearest neighbor (KNN). We also decided to use a well-known optimization algorithm to improve the results, the genetic optimization algorithm (GA) [4], which contributes to selecting the best parameters in the model and extracting important features that also contribute to obtaining the best performance. In the second section, we discuss relevant studies and compare them, in the third section, we describe the proposed algorithm and how it works, in the fourth section, we analyze and discuss the results, and in the last section, we describe the conclusion of this paper.

II. Related work

The study [5] introduced an innovative Ensemble Deep Learning based web attack detection System (EDL-WADS) for (IoT) networks, comprising three deep learning models: LSTM,MRN, and CNN with hyperparameter tuning, this a novel system underwent evaluation using two distinct datasets - HTTP CSIC 2010 and a real-world dataset. Remarkably, the EDL-WADS demonstrated exceptional performance, an astounding accuracy of 99.9% and 99.8% on the respectively. Furthermore, [6], an innovative pretraining methodology was devised, leveraging a deep autoencoder (PTDAE) in conjunction with a deep neural network (DNN) to enhance the intrusion detection system's (IDS) capability in identifying different types of attacks. To achieve optimal performance, hyperparameter tuning was conducted employing both grid search and random search techniques. The evaluation of this approach on the NSL-KDD and CSE-CIC-IDS2018 datasets yielded impressive overall accuracies of 83.33% and 95.79%, respectively. [7] underscored the importance of employing advanced machine learning and deep learning algorithms to detect SQL injection attacks. A thorough examination of previous studies utilizing these algorithms to identify attacks, especially those related to SQL, was undertaken. This meticulous survey encompassed a comprehensive questionnaire, addressing diverse facets of SQL attacks. Additionally, the study explored cutting-edge innovations and proposed remedies that utilize machine learning techniques to effectively counter SQL injection attacks. [8] introduced a novel Robust Software Modeling Tool (RSMT) designed to identify web-targeted attacks using an unsupervised/semi-supervised approach. The technique involved utilizing a stacked denoising auto-encoder to encode and reconstruct the call graph, enabling end-to-end deep learning. The evaluation encompassed synthetic datasets and real-world applications intentionally containing vulnerabilities. Remarkably, the proposed model achieved an impressive f1-score of 0.918, demonstrating its effectiveness in detecting attacks. In addition, [9] proposed a model consisting of two Recurrent Neural Networks (RNNs) with LSTM or GRU units to detect suspicious requests on web applications and used HTTP dataset CSIC and WAF logs dataset to evaluate the model which consisted of 36,000 normal requests and more than 25,000 anomalous requests and the model with LSTM achieved the highest accuracy of 0.984 using HTTP dataset CSIC and 0.985 with GRU using WAF logs dataset. [10] introduced attention-based deep neural networks for the efficient detection of real-time web attacks, particularly focusing on the Payload Locating Network (PLN). To evaluate the model's efficacy, a RealDataset was assembled, encompassing a vast collection of 3 million real-world web traffic instances. Impressively, the model demonstrated remarkable detection capabilities, boasting an outstanding accuracy of 99.84

III. Materials and methods

In this section, Fig 1 shows the proposed technique which we used four machine learning algorithms were used and linked with the Tree-based Pipeline Optimization Tool (TPOT), which relies on genetic algorithms that contribute to improving the accuracy and performance of the algorithms used by several processes.

A. Dataset description and collection

In [11] A new realistic web-based attack dataset is proposed using the DVWA, and utilizes a sophisticated architecture comprised of several real technologies to help researchers develop powerful security models against web-based assaults to classify and detect request traffic as malicious or benign. Using many tools to perform all attacks, such as Hping3, Burp Suite, Hydra, Remot3d, Vulscan, Beef, NMAP, Fping, Netcat, and Angry-IP-Scanner, to discuss many attacks that affect the web environment:

SQL Injection Attack: is a code injection technique used to compromise websites to gain administrator privileges. This attack targets websites that are unsecured. The attacker can inject SQL commands and gain access to the database to collect data [12]. SQL injection attacks cannot be prevented by firewalls and intrusion detection systems.

Brute Force Attacks: In general, protecting against brute force attacks is tough. As a result, brute force attacks are conducted on a massive number of key combinations on a trial-and-error basis [13]. Unlike a dictionary attack, it can target unknown combinations. When the key size is minimal, passwords can be readily broken up. When the key size is huge and the password is strong, a brute force assault takes a long time. For brute force attacks, a computer program or ready-made software is usually employed. To perform a brute force attack considerably quickly and efficiently, the computer setup must be high.

Dictionary Attack: The attack on authentication data attempts to use every available word in a dictionary. Dictionary attacks are limited to a specific list of weak passwords or a small number of key combinations with a high chance of success [14]. As a result, dictionary attacks are always faster than brute-force attacks. A dictionary attack is simple when the password is short, weak, or common, but it becomes exceedingly complicated and fails when unusual characters are used as passwords before attempting.

Browser Hijacking: An attack in which the goal is to modify a web browser setting such as the home page and bookmark to redirect the client to a different unwanted website [15].

Backdoor malware involves the installation of malware on a targeted system that allows the attacker to later obtain unauthorized access to the system. The malware, referred to as a "backdoor," generates a covert entry point into the system that can be used to circumvent security measures, get access to sensitive information, or perform destructive acts [16].

Cross-Site Scripting (XSS): allows an intruder to inject malicious code into a web page. The inject script can then be performed by any user with access to the page's web browser, allowing the attacker to steal personal data [17].

Uploading Attack: An uploading attack attempts to obtain

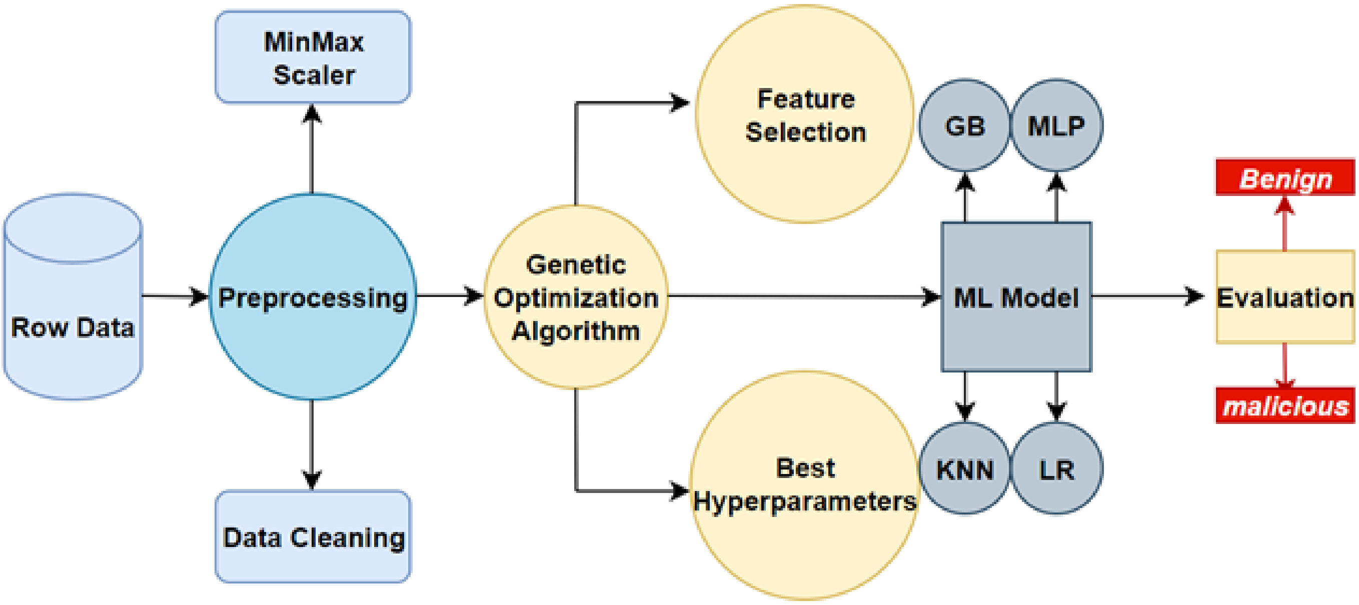

Fig. 1. Flowchart of Proposed Technique

unauthorized access or execute arbitrary code by uploading harmful files, [18] such as malware, to a targeted system .

Command injection: attack on online applications that involves inserting malicious commands into an input field with the objective of gaining unauthorized access to a system, stealing private information [19].

B. Data Pre-Processing

The data must be checked before entering it into the model so that it checks if there are missing values or NAN so that the data is entered appropriately for the models, and also to reduce the extreme values and the high dimensions between the values, we used normalization Min Max scaling to limit the values between (0,1), which is a statistical method represented in the equation 1:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Where x represents an initial value, x^* denotes the standardized value. X min corresponds to the minimum value, and X max represents the maximum value.

C. Tree-based Pipeline Optimization Tool (TPOT)

Tree-based Pipeline Optimization Tool (TPOT) is an automatic library that uses machine learning algorithms [20] based on its use Genetic algorithm, where its goal is to find the best combination of pre-processing steps and feature selection methods in our models Gradient Boosting, MLP, Support Vector Machine, and Logistic Regression, where the genetic optimization process works with an initial production of a set of tubes consisting of pre-processing steps and feature selection. The best method is found using the fitness function based on the training dataset. Where genetic algorithms are applied using several factors such as selection, intersection, and mutation to improve and develop a group of pipelines in the long term, where the pipeline is chosen best in terms of performance and then we make it multiply and its materials are inherited to the next generations, in return the pipes are discarded because of poor performance. We used four machine learning algorithms to detect and classify malicious attacks from benign attacks on websites.

Gradient boosting: It is a reinforced machine learning algorithm [21] that depends in its work on building several weak decision trees, and choosing a strong predictor based on training these weak trees and making them learn from previous mistakes and not repeating these mistakes in the next iterations, The gradient boosting algorithm is one of the algorithms that works On complex data, especially in cybersecurity, in its ability to detect and classify attacks

Multi-Layer Perceptron (MLP): It is considered an artificial neural network consisting of several layers [22], the most important of which are the input layer, the hidden layers, and the output layer, as it works on complex relationships and discovers complex patterns and non-linear relationships, In the context of web attacks, it has been used to detect attacks on websites and classify them as benign attacks.

K Nearest Neighbor (KNN): The K-Nearest Neighbor (KNN) classification algorithm [23] is a well-established data mining technique known for its theoretical maturity and computational efficiency. Its fundamental concept is based on identifying the category of a sample in a given sample

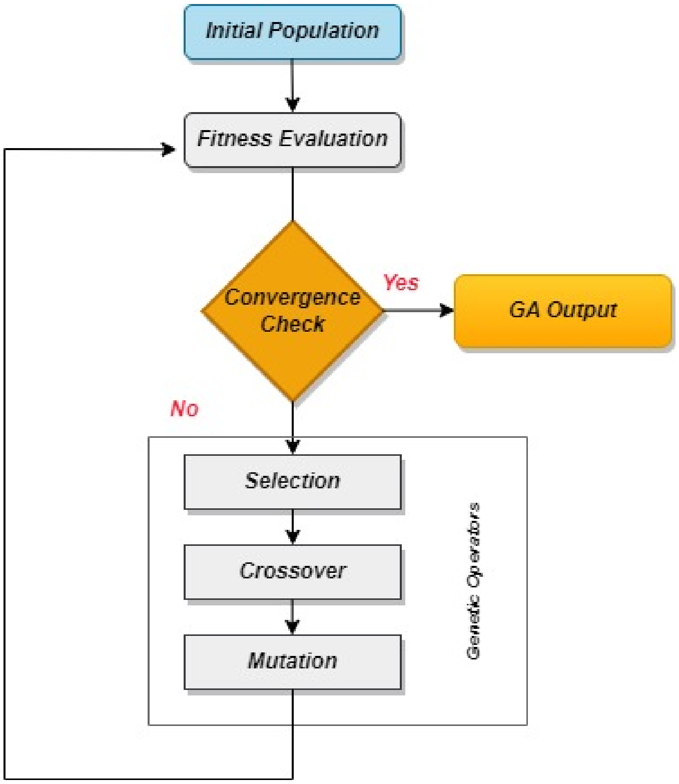

Fig. 2. Flow-Chart of a Genetic Algorithm

space by looking at the majority category among its K nearest neighboring samples. These neighbors are determined based on the Euclidean distance of the feature vectors, which represent the single or multidimensional characteristics describing the sample [24].

Logistic Regression: is a discriminative supervised learning algorithm widely employed for detecting web attacks. By establishing a mathematical connection between input features and binary outcomes (attack occurrence), it computes probabilities and applies a threshold to categorize instances into normal or malicious classes, thus proving to be a potent technique in identifying web-based security vulnerabilities [25].

IV. Performance Evaluation

Metrics such as accuracy, accuracy, recall, and F1-Score in evaluating machine learning algorithms or deep learning models play an important and vital role [26].

Accuracy measures the overall accuracy of the forecast as shown in equation 2:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision pays attention to correct positives and reduces false positives as shown in equation 3:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall is concerned with reducing false negatives as shown in equation 4:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-Score integrates precision and recall, as it balances false answers and false negatives as shown in equation 5:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

Confusion Matrix is also a tool for evaluating model performance, as it provides a comprehensive view of model predictions and real results, showing false negative (FN), true positive (TP), true positive (TP) and false positive (FP), as it is a comprehensive analysis of accuracy and recall. and precision and other metrics.Also, AUC-ROC Curve: a graphical representation of the performance of a binary classifier, plotting the true positive rate against the false positive rate at different classification thresholds. A perfect classifier has an AUC-ROC of 1, while a random classifier has an AUC-ROC of 0.5. A classifier with an AUC-ROC less than 0.5 is considered a poor classifier.

V. Results and Discussion

In this section, we discuss the results as shown in the Table I, which we got from four machine learning models, as we note in the Table I the Gradient Boost (GB) had the highest accuracy of 0.95 and AVG 0.95 for accuracy and memorization and the result f1 which outperformed other machine learning algorithms, and the MLP model comes with an accuracy of 0.83 and then the logistic regression model achieved an accuracy of 0.81 and in the other the KNN model where its accuracy was a little m Compared to other models, it achieved an accuracy of 0.76.

TABLE I
EVALUATION OF MODELS WITH DEFAULT PARAMETERS

ML Model	Accuracy	Precision	Recall	F1-Score
Gradient Boost	0.95	0.94 (1) 0.95 (0)	0.95 (1) 0.94 (0)	0.94 (1) 0.95 (0)
MLP	0.83	0.83 (1) 0.84 (0)	0.85 (1) 0.82 (0)	0.84 (1) 0.83 (0)
Logistic Regression	0.81	0.81 (1) 0.81 (0)	0.82 (1) 0.78 (0)	0.80 (1) 0.83 (0)
KNN	0.76	0.78 (1) 0.73 (0)	0.74 (1) 0.78 (0)	0.75 (1) 0.76 (0)

Table II shows the evaluation of the algorithms using the genetic algorithm to improve the optimal performance of each model, where the Gradient Boost (GB) model also outperformed the rest of the other models with an accuracy of 0.95,

TABLE II
EVALUATION OF MODELS TECHNIQUES USING GA

ML Model	Accuracy	Precision	Recall	F1-Score
Gradient Boost	0.95	0.94 (1) 0.95 (0)	0.95 (1) 0.94 (0)	0.94 (1) 0.95 (0)
MLP	0.86	0.87(1) 0.84 (0)	0.86 (1) 0.86 (0)	0.87 (1) 0.85 (0)
Logistic Regression	0.87	0.85 (1) 0.90 (0)	0.92 (1) 0.82 (0)	0.86 (1) 0.88 (0)
KNN	0.78	0.79 (1) 0.77 (0)	0.79 (1) 0.77 (0)	0.79 (1) 0.77 (0)

and the accuracy of the logistic regression algorithm increased, as the accuracy increased to 0.87. be suitable to achieve the highest performance of the model. The chart in Figure 3 shows the accuracy for each model, and the comparison between using Genetic optimization and without using it, as it turns out that the improved algorithm contributes significantly to improving the performance of each model.

It is necessary to clarify the classification process between the malicious and benign attack, where the confusion matrix shows the actual value and the expected value for each category, as it becomes clear to us the confusion matrix of the four models used in this study using the genetic optimization algorithm as shown in Figure 4, where it is found that we have the GB model superiority over the other models, as it predicted significantly TP and TN and a small number of false expectations FP and FN, and other models show wrong expectations and correct expectations, but there are some false expectations compared to the GB model, which superior to all other models, Figure 5 shows the receiver operating characteristic curve (ROC curve), where the GB model with the genetic optimization algorithm shows a clear superiority in terms of Curve, where the x axis curve represents the false positive rate (FPR) and the y axis represents the true positive rate (TPR). And the lowest AUC was for the KNN model, where it achieved AUC 0.87, and we clarify in the end that the GB model is superior to other models through confusion matrix and ROC Curve.

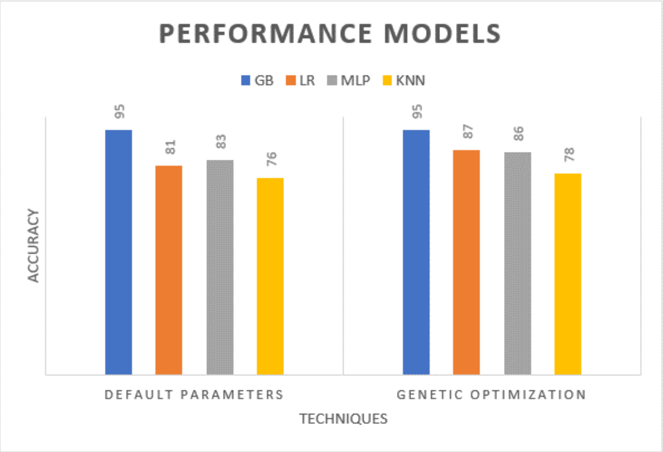

Fig. 3. Chart Represents Comparison Performance of Models

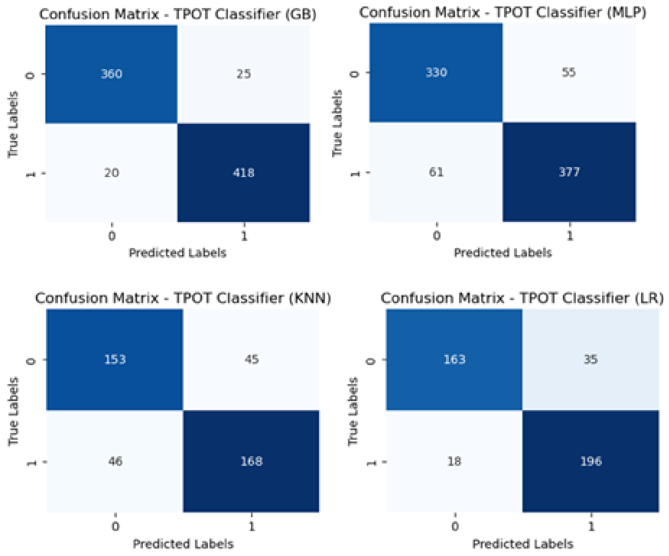

Fig. 4. Confusion Matrix of Models Using Genetic Optimization

CONCLUSION AND FUTURE WORK

Web application attacks pose significant risks to both individuals and organizations. Organizations may reduce the likelihood of successful attacks and secure sensitive data by taking a proactive, layered approach to security. Furthermore, user awareness and education is crucial in preventing successful attacks and mitigating the consequences of any potential breaches. Regular communication, information sharing, and

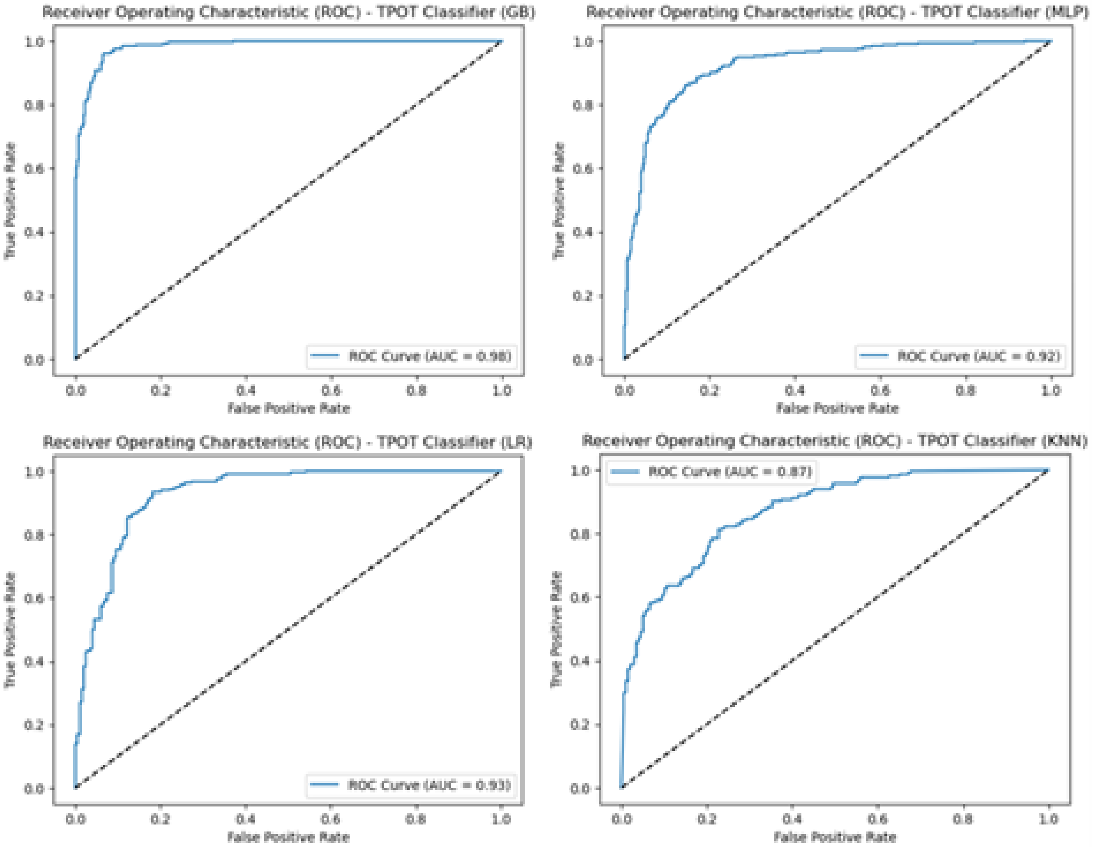

Fig. 5. ROC Curve Demonstrate AUC for Models Using Genetic Optimization

continuous monitoring are important components of a complete security plan to defend against web application attacks. Through machine learning and deep learning algorithms that contribute to the discovery of web attacks in the IoT system, we decided to use an approach based on Genetic Optimization Algorithm with the use of machine learning classifiers that include: Gradient Boost, MLP, Logistic Regression, KNN, where promising results were achieved, and higher The Gradient Boost machine learning model achieved an accuracy of 95%, precision 95%, recall 95%, and f1-score 95% and outperformed other models, demonstrating model efficiency with improved GA to select features that give the best performance. In the future, we will use other algorithms, especially in deep learning, with modern optimization algorithms, to achieve high results compared to other studies, which ensures the detection of cyber attacks on websites and provides security for users of this site.

References

[1] H.-C. Huang, Z.-K. Zhang, H.-W. Cheng, and S. W. Shieh, "Web application security: Threats, countermeasures, and pitfalls," *Computer*, vol. 50, no. 6, pp. 81–85, 2017.

[2] T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, "A comprehensive survey on the cyber-security of smart grids: Cyber-attacks, detection, countermeasure techniques, and future directions," *arXiv e-prints*, pp. arXiv–2207, 2022.

[3] A. Bansal and S. Kaur, "Extreme gradient boosting based tuning for classification in intrusion detection systems," in *Advances in Computing and Data Sciences: Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I 2*, pp. 372–380, Springer, 2018.

[4] H. Bhasin and S. Bhatia, "Application of genetic algorithms in machine learning," *IJCSIT*, vol. 2, no. 5, pp. 2412–2415, 2011.

[5] C. Luo, Z. Tan, G. Min, J. Gan, W. Shi, and Z. Tian, "A novel web attack detection system for internet of things via ensemble classification," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5810–5818, 2020.

[6] Y. N. Kunang, S. Nurmaini, D. Stiawan, and B. Y. Suprapto, "Attack classification of an intrusion detection system using deep learning and hyperparameter optimization," *Journal of Information Security and Applications*, vol. 58, p. 102804, 2021.

[7] I. Jemal, O. Cheikhrouhou, H. Hamam, and A. Mahfoudhi, "Sql injection attack detection and prevention techniques using machine learning," *International Journal of Applied Engineering Research*, vol. 15, no. 6, pp. 569–580, 2020.

[8] Y. Pan, F. Sun, Z. Teng, J. White, D. C. Schmidt, J. Staples, and L. Krause, "Detecting web attacks with end-to-end deep learning," *Journal of Internet Services and Applications*, vol. 10, no. 1, pp. 1–22, 2019.

[9] J. Liang, W. Zhao, and W. Ye, "Anomaly-based web attack detection: a deep learning approach," in *Proceedings of the 2017 VI International Conference on Network, Communication and Computing*, pp. 80–85, 2017.

[10] T. Liu, Y. Qi, L. Shi, and J. Yan, "Locate-then-detect: Real-time web attack detection via attention-based deep neural networks.," in *IJCAI*, pp. 4725–4731, 2019.

[11] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment," 2023.

[12] V. Srivastava, A. Majumdar, *et al.*, “Prevention of sql injection attacks in web applications,” *Journal of Survey in Fisheries Sciences*, vol. 10, no. 2S, pp. 1113–1119, 2023.

[13] R. P. Aji, Y. Prayudi, and A. Luthfi, “Analysis of brute force attack logs toward nginx web server on dashboard improved log logging system using forensic investigation method,” *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 1, pp. 39–48, 2023.

[14] J. Jayashree, J. Vijayashree, N. C. S. Iyengar, and S. M. Basha, “Intelligent interface for web system security enhancement,” *International Journal of Computational Learning & Intelligence*, vol. 2, no. 1, pp. 17–24, 2023.

[15] A. Z. Ablahd, “Using python to detect web application vulnerability,” *resmilitaris*, vol. 13, no. 2, pp. 1045–1058, 2023.

[16] M. Daka and D. E. Banda, “Strengthening web application security through technical measures,”

[17] J. Kaur, U. Garg, and G. Bathla, “Detection of cross-site scripting (xss) attacks using machine learning techniques: a review,” *Artificial Intelligence Review*, pp. 1–45, 2023.

[18] I. Putra, Y. Prayudi, and A. Luthfi, “Live forensics untuk mengenali karakteristik serangan file upload guna meningkatkan keamanan pada web server: Indonesia,” *JIIP-Jurnal Ilmiah Ilmu Pendidikan*, vol. 6, no. 6, pp. 4387–4394, 2023.

[19] A. Stasinopoulos, C. Ntantogian, and C. Xenakis, “Commix: Detecting and exploiting command injection flaws,” *Dept. Digit. Syst., Univ. Piraeus, Piraeus, Greece, White Paper*, 2015.

[20] R. S. Olson and J. H. Moore, “Tpot: A tree-based pipeline optimization tool for automating machine learning,” in *Workshop on automatic machine learning*, pp. 66–74, PMLR, 2016.

[21] V. A. Dev and M. R. Eden, “Gradient boosted decision trees for lithology classification,” in *Computer aided chemical engineering*, vol. 47, pp. 113–118, Elsevier, 2019.

[22] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, “Multi-layer perceptrons,” in *Computational intelligence: a methodological introduction*, pp. 53–124, Springer, 2022.

[23] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for knn classification,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, pp. 1–19, 2017.

[24] R. E. Al Mamlook, A. Nasayreh, H. Gharaibeh, and S. Shrestha, “Classification of cancer genome atlas glioblastoma multiform (tcga-gbm) using machine learning method,” in *2023 IEEE International Conference on Electro Information Technology (eIT)*, pp. 265–270, IEEE, 2023.

[25] J. Feng, H. Xu, S. Mannor, and S. Yan, “Robust logistic regression and classification,” *Advances in neural information processing systems*, vol. 27, 2014.

[26] A. S. Jaradat, R. E. Al Mamlook, N. Almakayeel, N. Alharbe, A. S. Almuflih, A. Nasayreh, H. Gharaibeh, M. Gharaibeh, A. Gharaibeh, and H. Bzizi, “Automated monkeypox skin lesion detection using deep learning and transfer learning techniques,” *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 4422, 2023.