Malicious Agricultural IoT Traffic Detection and Classification: A Comparative Study of ML Classifiers

Omar Bin Samin [1,2,*], Nasir Ahmed Abdulkhader Algeelani [2], Ammar Bathich [2], Ghulam Mujtaba Adil [1], Abdul Qadus [1], and Adnan Amin [1]

[1] Center of Excellence in Information Technology, Institute of Management Sciences, Peshawar, Pakistan;
Email: ghulam.mujtabadil001@gmail.com (G.M.A.), abdulqadus@imsciences.edu.pk (A.Q.),
adnan.amin@imsciences.edu.pk (A.A.)
[2] Faculty of Computer & Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia;
Email: nasir.ahmed@mediu.edu.my (N.A.A.A.), ammar.bathich@mediu.edu.my (A.B.)
*Correspondence: omar.samin@imsciences.edu.pk (O.B.S.)

***Abstract*—The number of internet-connected devices is rising, resulting in a global network of connected devices, referred to as Internet of Things (IoT). The technologically advanced agriculture industry employs IoT to monitor their environment and automate required functionality. IoT devices generate enormous amount of confidential and critical data, hence, securing the information is of significant importance. This research proposes integrating computationally intensive Machine Learning (ML) classifiers with resource-constrained IoT devices in order to safeguard the obtained data. This study analyses Naïve Bayes and Decision Tree for a cutting-edge Edge-IIoTset cybersecurity dataset encompassing 15 classes of IoT traffic derived from Soil Moisture, Temperature, Humidity, Water Level, and Water pH Sensors to enhance IoT data security. The experimental results of both the ML classifiers on given subsets of Edge-IIoTset presented Decision Tree as superior option, achieving accuracy of 72% and 73% for ML and DNN Edge-IIoTset respectively as compared to Naïve Bayes with accuracy of 47% and 45% respectively.**

***Keywords*—internet of things, anomaly detection, malicious activity classification, naïve bayes, decision tree**

I. Introduction

In today's technologically advanced era, the number of internet-connected objects is growing exponentially, resulting in a global network of connected items referred to as Internet of Things (IoT). IoT devices are able to monitor their surroundings, gather data, and transmit information to a remote system, which can then be utilized for data analysis and decision-making [1, 2]. It is anticipated that by 2025, between 60 to 75 billion IoT devices will be interconnected globally [3]. In an IoT network, things are interconnected via Bluetooth, Wireless Sensor Networks (WSN) or wireless mobile telecommunications technologies (3G, 4G, and 5G) in order to transmit collected data via the internet [4]. IoT has countless applications in several domains, such as agriculture, transportation, energy, healthcare, smart cities, and industry, etc. [5].

Agriculture is essential to the existence, growth, and advancement of mankind since it provides the vast majority of food [6]. The advanced technological agriculture industry strives to improve the quality and output of agricultural goods by utilizing IoT to sense environmental data to promptly meet crop requirements [7]. An agricultural IoT network generates a vast amount of data that often needs to be accessed remotely. The most of this generated data is confidential and valuable; therefore, its security is of utmost importance. However, because of their limited computational capabilities, they are unable to analyze massive volumes of data in a limited span of time and are more susceptible to attacks such as Distributed Denial of Services (DDoS), Man in the Middle (MITM) Attack, Backdoor Attack and Ransomware Attack etc. [8, 9]. Machine Learning (ML) is employed to boost computational capabilities and make IoT devices more resistant to intrusions and attacks. IoT along with ML is transforming the way of living and growing more rapidly than ever. IoT has emerged as the dominant technology with the fastest evolution rate, complemented by ML, which made IoT devices smart, intelligent, and automated.

ML focuses mostly on algorithm development that enables machines to independently learn from data and experiences, find patterns, and make predictions with minimum human intervention [10]. The two most prominent types of ML are Supervised Learning and Unsupervised Learning.

A. Supervised Learning

Machines are trained on labelled data and given the ability to predict outputs based on the provided training. The machine is thereby trained using the input and matching output [11]. Both the input and output values are

Manuscript received March 24, 2023; revised April 18, 2023; accepted May 10, 2023; published August 17, 2023.

known; however, the mapping function is unknown (see Fig. 1).

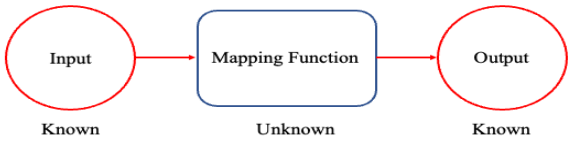

Figure 1. Supervised learning.

The expression that is formulated by ML to map input data to output is known as the "Mapping Function". The function shows how the elements are paired. Mathematically, for the input (X) and output (Y), ML algorithms seek the optimal mapping function (f) such that (see Eq. (1)):

$$Y = f\ (X) \tag{1}$$

Supervised learning can be utilized for classification and regression scenarios. Naïve Bayes, Nearest Neighbor, and Decision Trees etc. are few of notable supervised learning ML techniques.

B. Unsupervised Learning

Machines are trained on unlabeled data, and it aims to group the unsorted data based on the similarities, differences, and patterns [12]. Both the mapping function and output values are unknown; however, the input is known (see Fig. 2).

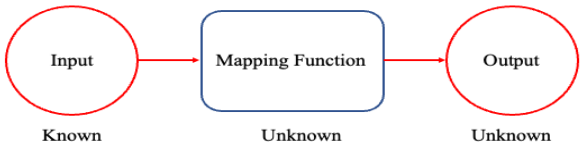

Figure 2. Unsupervised learning.

In such scenarios, ML algorithms map the function that finds similarity among different input data instances (samples) and group them based on the similarity index, which is the output for unsupervised learning. It can be utilized for clustering and association scenarios. Gaussian Mixture, Fuzzy C-Means, Hierarchical and Hidden Markov Model etc. are few of notable unsupervised learning ML techniques.

C. Semi-supervised Learning

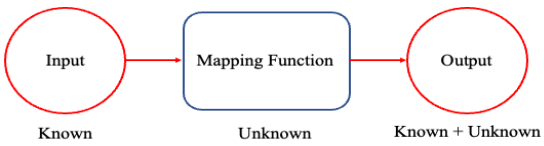

Figure 3. Semi-supervised learning.

Machines are trained on the combination of labeled and unlabeled data. It employs labeled data to learn how to classify new data points and unlabeled data to enhance its predictions. By combining both labeled and unlabeled data, semi-supervised learning improves the accuracy of the model and reduce the volume of labeled data required for training. In semi-supervised learning, the input is known and the mapping function is unknown, while the output is divided into two distinct groups, one of which is known and the other is unknown (see Fig. 3).

Semi-supervised learning can be utilized for classification, regression as well as for clustering and association scenarios.

Despite the fact that a substantial amount of research has been conducted in agricultural IoT security, however, the majority of the conducted research work is based on either a restricted amount of data or a relatively limited diversity of IoT attacks. The proposed research work focuses on finding the optimal ML technique between two renowned ML classifiers. i.e., Naïve Bayes and Decision Tree for classification and identification of normal and vast variety of malicious activities specifically in agricultural IoT domain using Edge-IIoTset. Thus, bridging the gap between computationally excessive ML techniques and resource constrained agricultural IoT devices to improve system's security as well as efficiency.

The organization details of the paper are as follows. Section II: Background Study discusses the related work linked with classification and identification of malicious IoT activities utilizing both Naïve Bayes and Decision Tree classifiers. Section III: Methodology discusses the implementation details in depth, and Section IV: Results & Discussion presents the important findings of this study. Finally, Section V: Conclusion concludes the paper by declaring the optimal classifier specifically for Edge-IIoTset.

II. Background Study

Due to the exponential growth of IoT devices, the significance of IoT security is increasing constantly. There has been a considerable amount of work done in IoT security and malicious activity's detection, some of which is discussed below. To strengthen the process of understanding, the background study is divided into the following two subsections:

A. Naïve Bayes

Naïve Bayes is a useful classification technique for both binary and multi-class classification. It is a method of categorization (classification) which is based on supervised learning and Bayes' theorem (see Eq. (2)), assuming all features stated in a dataset are independent, meaning one feature has no effect on other feature, hence, the name "naive". The probability of an occurrence is computed based on the frequency of values in historical data using the given formula [13]:

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)} \tag{2}$$

In Eq. (2), $P\ (A \mid B)$ is posterior probability of class given predictor, $P\ (B \mid A)$ is likelihood, $P(A)$ is prior probability of class and $P(B)$ is prior probability of predictor. Naïve Bayes is able to classify data by assigning a label to the records using conditional probability [14, 15].

It functions optimally when the dataset is small and rich in attributes. Moreover, it can manage both continuous and discrete data.

Foo *et al.* [16] presented an anomaly detection approach utilizing K-Means and Ada-boosted Naïve Bayes classifier along with UNSW-NB15 [17] and TON IoT telemetry datasets [18]. The approach utilizes unsupervised k-means clustering to initially group attacks into 7 categories, and Ada-Boosted Naïve Bayes algorithm assists the clusters to determine which group corresponds to a particular attack. The suggested work claims to achieve 90% to 100% scores (accuracy, precision, and recall).

Majeed *et al.* [19] proposed an approach for developing IoT assisted drones with an intelligent cybersecurity system that will aid in detecting network security threats. The proposed work utilized Naïve Bayes with KDD'99 dataset [20] for this purpose. The suggested approach achieved an accuracy of 96.3%. Mehmood *et al.* [21] suggested an approach to secure IoT environment from DDoS attacks. The study utilized Naïve Bayes along with NSL-KDD [22] for threat detection. Manimurugan [23] suggested a method to connect IoT devices to cloud and fog computing for anomaly detection. This approach utilized Naïve Bayes and Principal Component Analysis (PCA) for anomaly detection. The suggested work utilized UNSW-NB15 dataset [17] and achieved an accuracy of 92.48%.

B. Decision Tree

A decision tree is a supervised learning method used for both classification and regression analysis. The decision tree functions optimally with both categorical and numeric data. They are hierarchical data structures that divide input data space into several subspaces in order to predict target variables [24]. The decision tree consists of the root node, internal nodes, leaf nodes, and branches (see Fig. 4).

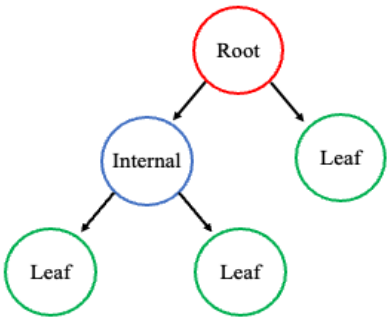

Figure 4. Basic structure of decision tree.

1) Root node

It marks the starting point of the tree and is the top node [25]. It may have one or more child nodes, but can never have sibling nodes and represents the entire dataset to be analyzed.

2) Internal node

It is also known as chance node, extends from the parent node and is linked by branch. Internal node branches will connect to other internal nodes or leaf nodes [26].

3) Leaf node

It is also known as terminal node or end node, marks the end point of the tree and cannot be further divided. It represents the final result [27].

Information gain and entropy are the most typical attribute selection measures. Information gain determines which specific feature most effectively separates the training dataset depending on the target classification, whereas entropy assesses the randomness of the dataset [25]. Entropy and information gain can be calculated as follows (see Eqs. (3) and (4)):

$$E\ (S) = \sum_{i=1}^{c} x_i - P_i\ log_2\ P_i \qquad (3)$$

In Eq. (3), E represents entropy and P_i represents the probability of samples S belonging to a class i.

$$Gain(S, A) = E(S) - \sum_{v\ \in\ values\ (A)} \frac{|S_v|}{|S|} E(S_v) \qquad (4)$$

In Eq. (4), v is the possible values for attribute A, S is the set of samples and S_v is a subset of samples where S=v.

Ferrag *et al.* [28] proposed an intrusion detection system (IDS) using decision tree along with CICIDS2017 dataset [29] and the BoT-IoT dataset [30]. The suggested work was able to classify the IoT traffic as "Normal" or "Attack" and achieved accuracy of 96.665% and 96.995% for both datasets respectively. Injadat *et al.* [33] presented ML framework utilizing decision tree to identify attacks on IoT devices. The suggested framework was able to classify IoT traffic into two classes (i.e., Normal and Attack). Using Bot-IoT dataset [30], the suggested framework's performance is assessed and achieved an overall accuracy of 99.99%. Leevy *et al.* [34] presented a classification approach for received IoT traffic. A minimal number of dataset features and a decision tree classifier are employed by this approach. To obtain the desired results, the work requires predictive models to have AUC and AUPRC mean scores greater than 0.99.

Douiba *et al.* [31] implemented an anomaly detection model using decision tree and gradient boosting along with NSL-KDD [22], BoT-IoT [30], and Edge-IIoTset [35] datasets. The suggested work was able to classify various IoT attacks and achieved an accuracy of 99.9%. Pohan *et al.* [32] developed an IoT security mechanism to detect injection attack using Catboost, Decision Tree, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) along with Edge-IIoTset [35] dataset. The suggested mechanism was able to detect only one IoT threat (i.e., injection attack) with the overall accuracy of 95.59% for Catboost, 93.47% for Decision tree, 91.83% for SVM and 93.46% for MLP.

The aforementioned literature on IoT security either deals with a small amount of data or a very limited variety of IoT attacks. Also, it is not appropriate to evaluate and rank the performance of various ML classifiers utilizing different datasets. The proposed work will address each of

these concerns. The contribution of this research is to identify the optimal ML classifier capable of detecting the most number of IoT attack classes based on different volumes of data from the same dataset specifically for agricultural applications.

III. Methodology

Presently, various ML techniques and classifiers for detecting and classifying IoT based normal and attack data traffic claim to be optimal. However, it is not possible to evaluate their performance and effectiveness in the various scenarios including diverse datasets with variable numbers of training/testing records and IoT attacks. In order to validate their claims, this proposed research seeks to establish a solid foundation for ML classifier selection for protecting IoT networks against attacks by comparing Naïve Bayes and Decision Tree on a single dataset specifically for agricultural applications. This research focuses on using ML based classifiers in IoT agriculture domain.

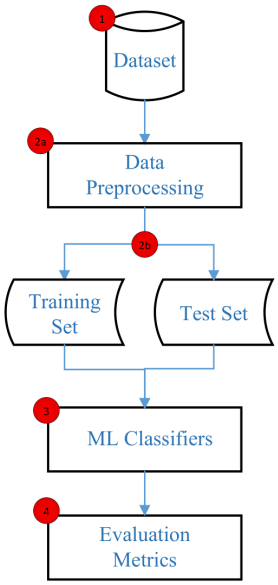

Figure 5. Intended methodology Pipeline.

The two ML classifiers (Naïve Bayes and Decision Tree) are arbitrarily selected specifically for the diverse and state-of-the-art dataset "Edge-IIoTset" available for public research. This study contributes to exploring the use of ML classifiers for efficiently detect and classify IoT traffic into “Normal” and “Attack” classes to take timely required actions. The findings of this study will provide a basis for declaring Naïve Bayes or Decision Tree as the optimal choice for classifying IoT traffic for agricultural applications based on a specific dataset. The Fig. 5 represents the proposed pipeline of an intended methodology.

A. Dataset

The selection of datasets is crucial for the detection and classification of IoT attacks. The state-of-the-art Edge-IIoTset [35] public IoT security dataset is selected. The dataset contains data regarding IoT traffic, having a total of 15 classes, including 14 classes of IoT attacks and 1 class of normal IoT data (see Table I). In the proposed study, the public dataset is utilized since public datasets allow researchers to compare their valuable contributions to those of others.

TABLE I. EDGE-IIoTset Details [35]

IoT Traffic	Class	Records
Normal	Normal	11,223,940
Attacks	Backdoor	24,862
	DDoS HTTP	229,022
	DDoS ICMP	2,914,354
	DDoS TCP	2,020,120
	DDoS UDP	3,201,626
	Fingerprinting	1,001
	MITM	1,229
	Password	1,053,385
	Port Scanning	22,564
	Ransomware	10,925
	SQL Injection	51,203
	Uploading	37,634
	Vulnerability Scanner	145,869
	XSS	15,915
	Backdoor	24,862
	Total	**20,952,648**

The dataset is derived from Soil Moisture, Temperature, Humidity, Water Level and Water pH Sensors etc., and can be used for agricultural IoT security applications. The following are the specifics of the IoT data traffic classes included in the Edge-IIoTset dataset:

(1) **Normal:** Legitimate data and requests.
(2) **Backdoor Attack:** Installs backdoors to take control of vulnerable IoT network components.
(3) **DDoS HTTP Attack:** Manipulates HTTP and post unwanted requests [36, 37].
(4) **DDoS ICMP Attack:** Overwhelms the target device with Internet Control Message Protocol (ICMP) echo requests (pings).
(5) **DDoS TCP Attack:** Overwhelms the target device with SYN requests to disable it to respond to new connection requests.
(6) **DDoS UDP Attack:** Overwhelms the target device with numerous User Datagram Protocol (UDP) packets to disable its processing and responding capabilities.
(7) **Fingerprinting Attack:** Analyzes IoT data packets to identify IoT device and server vulnerabilities.
(8) **MITM Attack:** Intercepts the communications between two IoT devices or IoT device and a server [38].
(9) **Password Attack:** Gains unauthorized access to an IoT device by cracking its password.
(10) **Port Scanning Attack:** Identifies the IoT network’s weak points or open doors.
(11) **Ransomware Attack:** Encrypts IoT data or systems to block or restrict access till the subject pays a ransom [39, 40].
(12) **SQL Injection Attack:** Reads/ inserts/ updates/ delete sensitive information from the database by injecting SQL query.

(13) **Uploading Attack:** Uploads files that contain malware command in order to control the device.
(14) **Vulnerability Scanner Attack:** Searches and identifies IoT network security vulnerabilities.
(15) **Cross-site Scripting (XSS) Attack:** Sends a malware script to the user, allowing access to sensitive data [41].

ML-Edge-IIoTset and DNN-Edge-IIoTset are two subsets of Edge-IIoTset given by the dataset's developers (see Tables II and III).

TABLE II. ML-EDGE-IIOTSET DETAILS [35]

IoT Traffic	Class	Records
Normal	Normal	24,301
Attacks	Backdoor	10,195
	DDoS HTTP	10,561
	DDoS ICMP	14,090
	DDoS TCP	10,247
	DDoS UDP	14,498
	Fingerprinting	1,001
	MITM	1,214
	Password	9,989
	Port Scanning	10,071
	Ransomware	10,925
	SQL Injection	10,311
	Uploading	10,269
	Vulnerability Scanner	10,076
	XSS	10,052
	Backdoor	24,862
	Total	**157,800**

TABLE III. DNN-EDGE-IIOTSET DETAILS [35]

IoT Traffic	Class	Records
Normal	Normal	1,615,643
Attacks	Backdoor	24,862
	DDoS HTTP	49,911
	DDoS ICMP	116,436
	DDoS TCP	50,062
	DDoS UDP	121,568
	Fingerprinting	1,001
	MITM	1,214
	Password	50,153
	Port Scanning	22,564
	Ransomware	10,925
	SQL Injection	51,203
	Uploading	37,634
	Vulnerability Scanner	50,110
	XSS	15,915
	Backdoor	1,615,643
	Total	**2,219,201**

The developers of Edge-IIoTset suggested utilizing ML-Edge-IIoTset when employing ML techniques, while DNN-Edge-IIoTset when employing deep learning techniques. Each of these subsets contains all classes from the entire dataset, but varying quantities of data to accommodate the requirements of investigators.

B. Data Preprocessing

Edge-IIoTset contains a total of 63 attributes, 17 of which are of the "object" type containing null, zero values and mixed datatype. These object attributes may comprise integer, floating-point, or string data and are denoted as "Not a Number (NaN)"; however, they have no significance on the application of ML techniques. The datasets (ML-Edge-IIoTset and DNN-Edge-IIoTset) are preprocessed by removing all attributes having all zero, null and NaN values, leaving 46 attributes to be utilized for classification.

As depicted in Tables I–III the datasets are significantly unbalanced. Unbalanced class distribution in the training dataset is the cause of the imbalanced classification problem. These unbalanced datasets are transformed into balanced datasets using a data augmentation approach referred to as Synthetic Minority Oversampling Technique (SMOTE), in order to address the imbalance issue, yielding in balanced datasets (see Tables IV and V) with exact of 6.66% data for each class. To equalize class distribution, SMOTE generates synthetic samples for the minority class. This is accomplished by selecting instances at random from the minority class, locating their k-nearest neighbors (which are often members of the same class), and producing synthetic examples as a linear combination of the selected instance and its neighbors. This procedure is carried out multiple times until the appropriate level of class balance is attained.

TABLE IV. BALANCED ML- EDGE-IIOTSET DETAILS

IoT Traffic	Class	Records
Normal	Normal	24,301
Attacks	Backdoor	24,301
	DDoS HTTP	24,301
	DDoS ICMP	24,301
	DDoS TCP	24,301
	DDoS UDP	24,301
	Fingerprinting	24,301
	MITM	24,301
	Password	24,301
	Port Scanning	24,301
	Ransomware	24,301
	SQL Injection	24,301
	Uploading	24,301
	Vulnerability Scanner	24,301
	XSS	24,301
	Backdoor	24,301
	Total	**364,515**

TABLE V. BALANCED DNN-EDGE-IIOTSET DETAILS

IoT Traffic	Class	Records
Normal	Normal	1,615,643
Attacks	Backdoor	1,615,643
	DDoS HTTP	1,615,643
	DDoS ICMP	1,615,643
	DDoS TCP	1,615,643
	DDoS UDP	1,615,643
	Fingerprinting	1,615,643
	MITM	1,615,643
	Password	1,615,643
	Port Scanning	1,615,643
	Ransomware	1,615,643
	SQL Injection	1,615,643
	Uploading	1,615,643
	Vulnerability Scanner	1,615,643
	XSS	1,615,643
	Backdoor	1,615,643
	Total	**24,234,645**

Both the datasets are divided into training sets (comprising 70% of the datasets) and test sets (comprising 30% of the datasets).

C. ML Classifiers

Naïve Bayes and Decision Tree are selected for the proposed comparative study. The specifics of both the mentioned classifiers are already discussed in the Background Study.

D. Evaluation Metrics

Evaluation metrics are used to evaluate the performance of the statistical and ML models. To compare the performances of Naïve Bayes and Decision Tree, the following four evaluation metrics are used.

1) Accuracy

Accuracy is a measurement of the system's true performance in correctly detecting and rejecting objects. It is the ratio between correct observations and total observations (see Eq. (5)). It is commonly expressed in percentage [42].

$$Accuracy = \frac{T_P + T_n}{T_P + T_n + F_P + F_n} \quad (5)$$

2) Precision

The correctly recognized positive samples relative to the total number of predicted positive samples (both true and false) determines precision [43, 44]. It is the ratio between correct observations and total observations for a particular class (see Eq. (6)).

$$Precision = \frac{T_P}{T_P + F_P} \quad (6)$$

3) Recall

The recall is the proportion of accurately identified positive samples relative to the total positive samples (see Eq. (7)). The greater the recall, the greater. the number of positive samples identified [43].

$$Recall = \frac{T_P}{T_P + F_n} \quad (7)$$

4) F1-Score

The F1-Score is determined by calculating the harmonic mean of the precision and recall of a classifier to create a single statistic (see Eq. (8)). It is typically used to compare the outcomes of two distinct classifiers [42–44].

$$F1 - Score = \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

In Eqs. (5)–(8), TP is positive, and the predicted value is also positive, Tn is negative, and the predicted value is also negative. Similarly, FP is negative, and the predicted value is positive, FN is positive, and the predicted value is negative. A TP predicts the positive class of a model correctly, likewise, a Tn predicts the negative class of a model correctly.

IV. Results and Discussion

In this research, Edge-IIoTset (ML Edge-IIoTset and DNN Edge-IIoTset); an open access dataset of IoT data traffic (both authentic and malicious) has been used for training and testing purposes. The proposed research examined two approaches; Naïve Bayes and Decision Tree for the selected dataset. In order to predict and compare the performance of both classifiers, for both subsets of Edge-IIoTset (i.e., ML Edge-IIoTset and DNN Edge-IIoTset) evaluation metrics (accuracy, precision, recall and F1-score) are calculated.

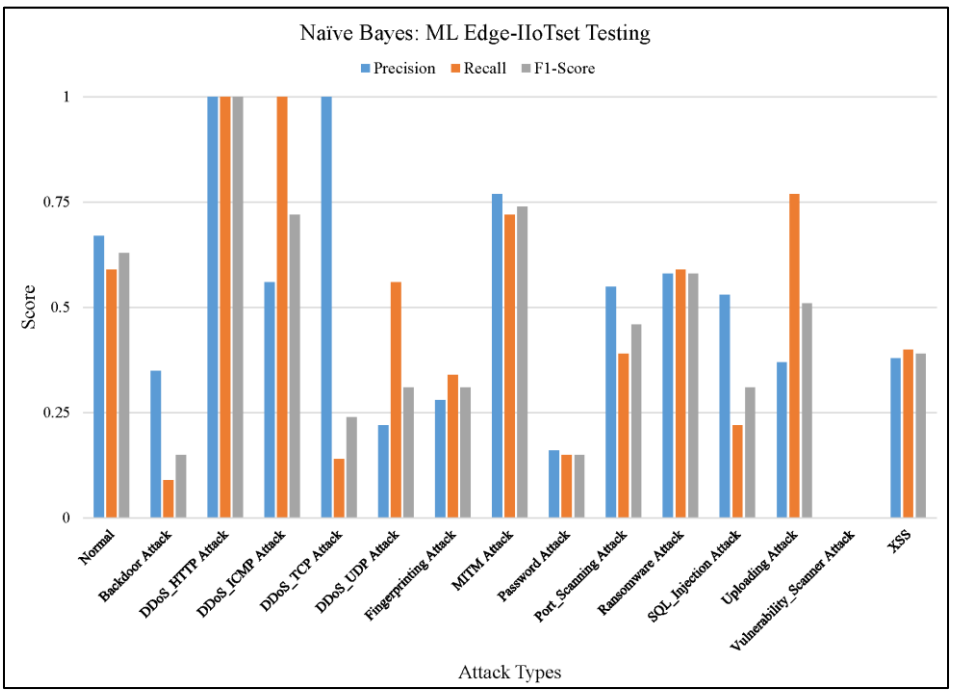

Figure 6. Naïve bayes: ML-Edge-IIoTset testing for 15 IoT traffic classes.

The research's findings and analyses shed light on how well Naive Bayes and Decision Tree classifiers work at identifying and categorizing harmful IoT data traffic. The results show without a doubt that the Decision Tree classifier performs better than the Naive Bayes classifier. This can be due to the Decision Tree's proficiency in handling the dataset's complicated linkages, which enhances generalization and predictability. Additionally, the open access Edge-IIoTset dataset has demonstrated to be a trustworthy and useful tool for researchers and industry professionals, enabling them to better comprehend and handle the difficulties brought on by harmful IoT data flow.

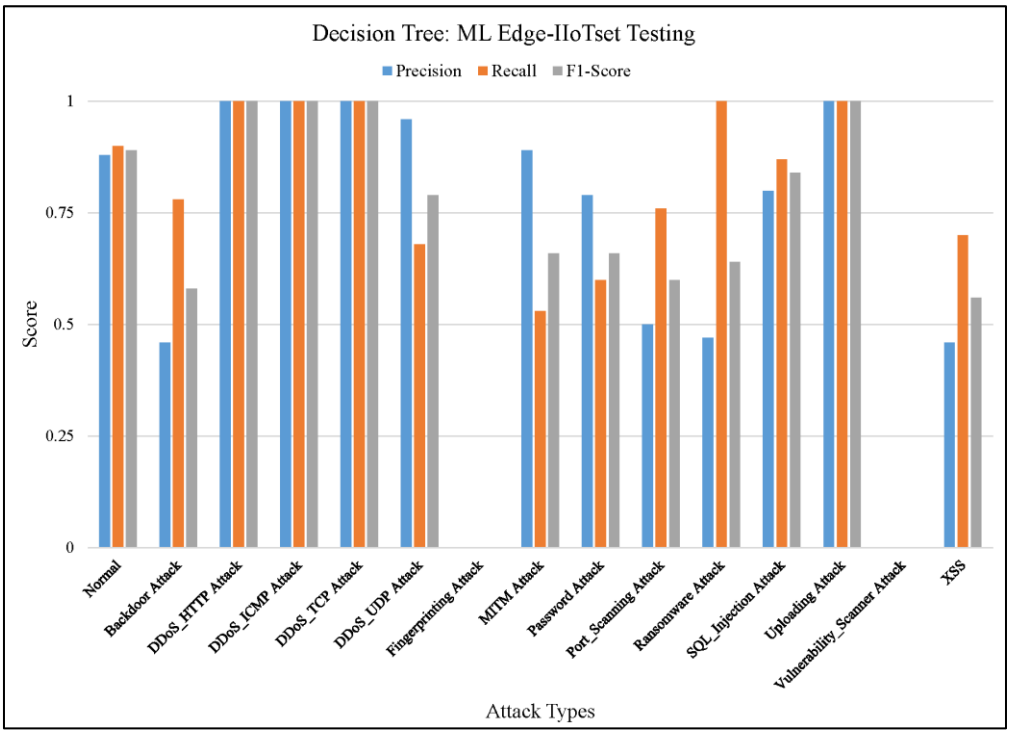

Figure 7. Decision tree: ML-Edge-IIoTset testing for 15 IoT traffic classes.

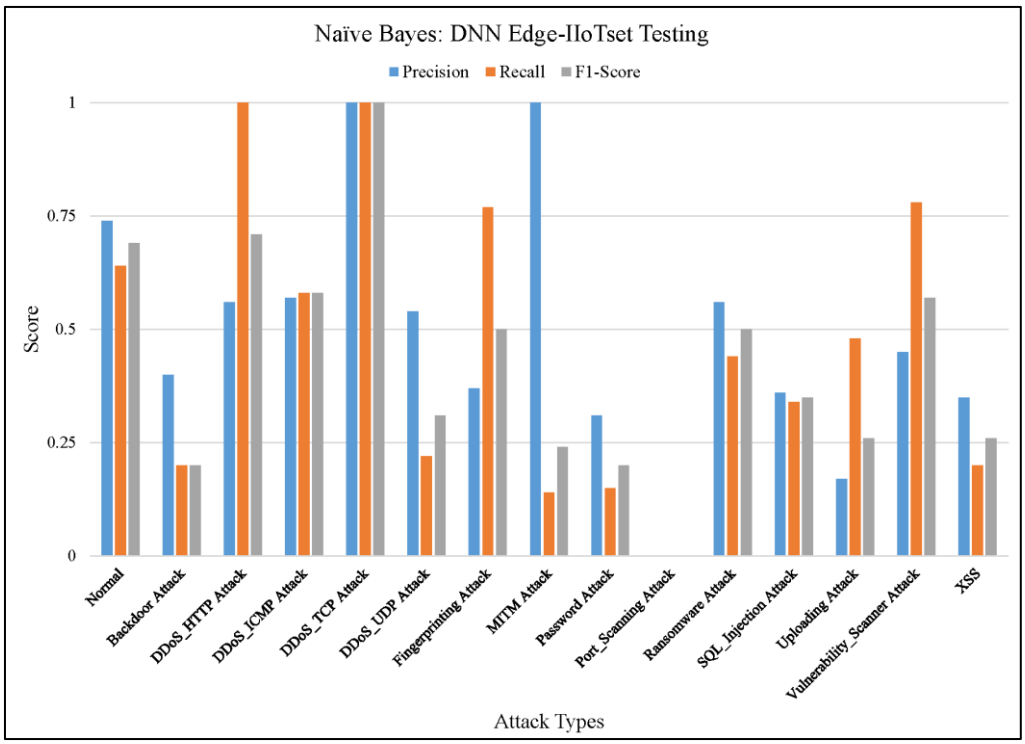

Figure 8. Naïve bayes: DNN-edge-IIoTset testing for 15 IoT traffic classes.

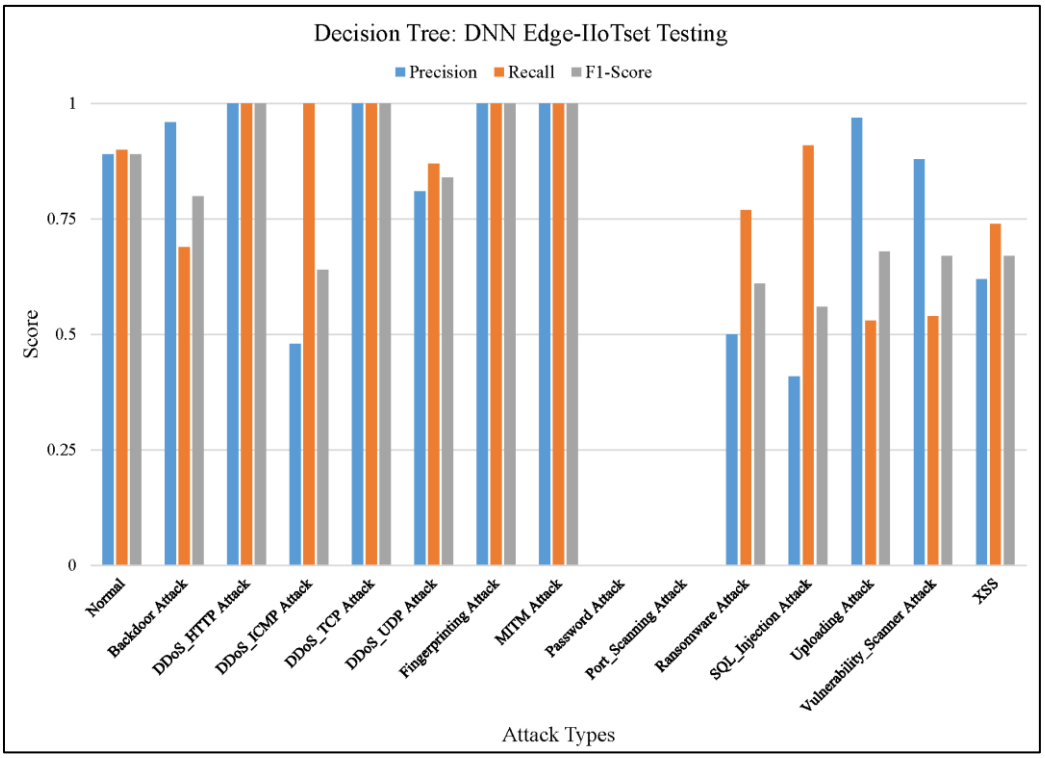

Figure 9. Decision tree: DNN-edge-IIoTset testing for 15 IoT traffic classes.

Figs. 6–9 represents precision, recall and F1-score of Naïve Bayes classifier and Decision Tree (at a maximum depth of 7), calculated for ML Edge-IIoTset, while Figs. 8 and 9 represents precision, recall and F1-score of Naïve Bayes classifier and Decision Tree (at a maximum depth of 7), calculated for DNN Edge-IIoTset respectively.

TABLE VI. EVALUATION METRICS

Classifier	Dataset	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	ML-Edge-IIoTset	47%	0.49	0.46	0.43
Decision Tree	ML-Edge-IIoTset	72%	0.68	0.71	0.68
Naïve Bayes	DNN-Edge-IIoTset	45%	0.46	0.44	0.40
Decision Tree	DNN-Edge-IIoTset	73%	0.70	0.72	0.69

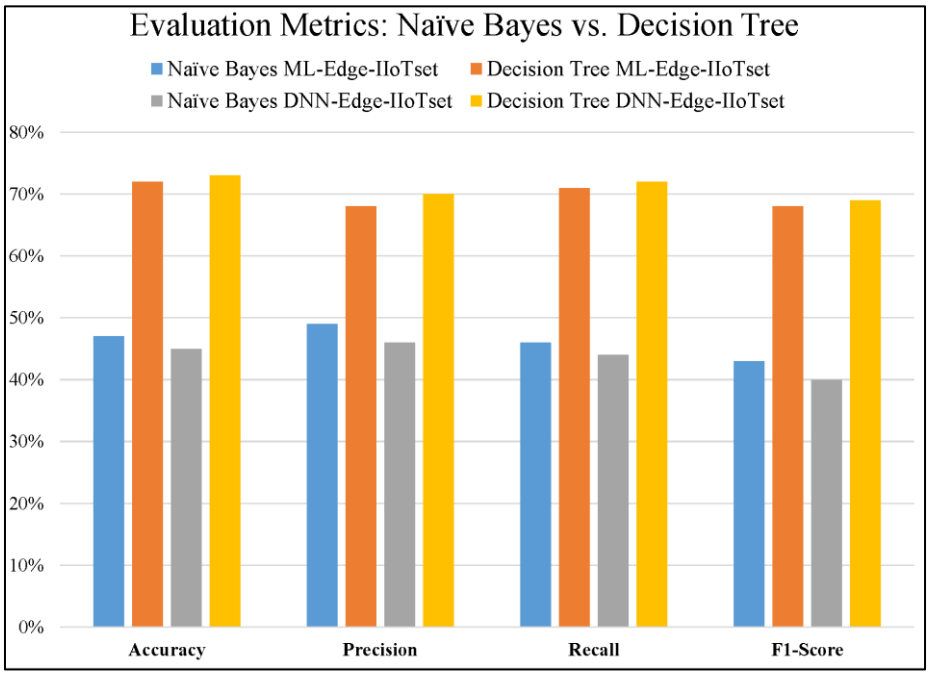

Figure 10. Evaluation metrics: Naïve bayes vs. decision tree.

Table VI lists the attained average evaluation metrics for the specified subsets of the Edge-IIoTset including both ML classifiers for detecting and classifying malicious IoT data traffic and clearly demonstrates that the Decision Tree outperform Naïve Bayes (see Fig. 10).

V. Conclusion

The number of IoT devices is rapidly increasing. IoT devices gather and communicate massive amounts of data, which are then analyzed using ML techniques for classification or prediction purposes to make IoT devices more secure and impervious to harmful assaults by enabling them to make decisions autonomously. In this study, two ML classifiers are examined for two subsets of Edge-IIoTset in which the volume of records differs between the subsets. ML Edge-IIoTset consists of fewer records than DNN Edge-IIoTset. ML takes more time to process when dealing with larger datasets. The results clearly shows that the Naïve Bayes performs better for ML-Edge-IIoTset as compared to DNN-Edge-IIoTset and Decision Tree performs better for DNN-Edge-IIoTset as compared to ML-Edge-IIoTset.

The assumption of independent predictors in Naive Bayes may have contributed to its less effective performance, which was less encouraging. The main finding of this study is that the amount of data directly influences how well ML classifier's function, with Decision Trees performing better as the dataset size increases and Naive Bayes performing worse. Overall, Decision Tree outperforms Naïve Bayes.

The limited availability of cellular networks and the internet in rural areas, which are required for the IoT devices to function and communicate, is the primary limitation of the proposed research when implementing the proposed study in a real-world scenario. In the future, a framework will be devised to address cellular network limitations in order to enable IoT use in rural areas.

Conflict of Interest

The authors declare no conflict of interest

Authors Contributions

Omar Bin Samin conceived the idea, designed and performed the experiments along with computation work. He also analyzed the data and drafted the manuscript. Nasir Ahmed Abdulkhader Algeelani provided domain knowledge and helped in analyzing results. He also reviewed drafts of the paper and approved the final draft. Ghulam Mujtaba Adil, Abdul Qadus and Adnan Amin helped in analyzing results and reviewed drafts of the paper. All authors had approved the final version.

References

[1] K. Sekaran, M. N. Meqdad, P. Kumar, S. Rajan, and S. Kadry, "Smart agriculture management system using internet of things," *Telkomnika*, vol. 18, no. 3, pp. 1275–1284, 2020.

[2] S. Pallavi and V. A. Narayanan, "An overview of practical attacks on BLE based IoT devices and their security," in *Proc. 2019 5th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2019, pp. 694–698.

[3] K. O. M. Salih, T. A. Rashid, D. Radovanovic, and N. Bacanin, "A comprehensive survey on the internet of things with the industrial marketplace," *Sensors*, vol. 22, no. 3, 730, 2022.

[4] A. Khanna and S. Kaur, "Evolution of Internet of Things (IoT) and its significant impact in the field of precision agriculture," *Computers and Electronics in Agriculture*, vol. 157, pp. 218–231, 2019.

[5] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of IoT devices by using mobile computing: A systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[6] K. Timmis and J. L. Ramos, "The soil crisis: The need to treat as a global health problem and the pivotal role of microbes in prophylaxis and therapy," *Microb Biotechnol*, vol. 14, no. 3, pp. 769–797, 2021, doi: 10.1111/1751-7915.13771

[7] X. Yang, L. Shu, J. Chen, M. A. Ferrag, J. Wu, E. Nurellari, and K. Huang, "A survey on smart agriculture: Development modes, technologies, and security and privacy challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 2, pp. 273–302, 2021.

[8] V. Tomer and S. Sharma, "Detecting IoT attacks using an ensemble machine learning model," *Future Internet*, vol. 14, no. 4, 102, 2022.

[9] R. Ramadan, "Internet of Things (IoT) security vulnerabilities: A review," *PLOMS AI*, vol. 2, no. 1, 2022.

[10] S. Ray, "A quick review of machine learning algorithms," in *Proc. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 35–39.

[11] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: Machine intelligence approach for drug discovery," *Molecular Diversity*, vol. 25, no. 3, pp. 1315–1360, 2021.

[12] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[13] K. Rrmoku, B. Selimi, and L. Ahmedi, "Application of trust in recommender systems—Utilizing naive bayes classifier," *Computation*, vol. 10, no. 1, 6, 2022.

[14] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and naive bayes," *The Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021.

[15] J. K. Sethi and M. Mittal, "Efficient weighted naive bayes classifiers to predict air quality index," *Earth Science Informatics*, vol. 15, no. 1, pp. 541–552, 2022.

[16] L. Best, E. Foo, and H. Tian, "Utilising k-means clustering and naive bayes for IoT anomaly detection: A hybrid approach," pp. 177–214, 2022.

[17] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. 2015 Military Communications and Information Systems Conference (MilCIS)*, IEEE, 2015, pp. 1–6.

[18] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020.

[19] R. Majeed, N. A. Abdullah, and M. F. Mushtaq, "IoT-based cyber-security of drones using the naïve bayes algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[20] S. Hettich and S. Bay. KDD'99 network intrusion detection data set. UCI Machine Learning Repository. [Online]. Available: http://archive. ics. uci. edu/ml/

[21] A. Mehmood, M. Mukherjee, S. H. Ahmed, H. Song, and K. M. Malik, "NBC-MAIDS: Naïve Bayesian classification technique in multi-agent system-enriched IDS for securing IoT against DDoS attacks," *The Journal of Supercomputing*, vol. 74, no. 10, pp. 5156–5170, 2018.

[22] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[23] S. Manimurugan, "IoT-fog-cloud model for anomaly detection using improved naive bayes and principal component analysis," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2021.

[24] C. Jin, F. Li, S. Ma, and Y. Wang, "Sampling scheme-based classification rule mining method using decision tree in big data environment," *Knowledge-Based Systems*, vol. 244, 108522, 2022.
[25] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable Artificial Intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, 6634811, 2021.
[26] S. Abbas, R. Hodhod, and M. El-Sheikh, "Retrieval of behavior trees using map-and-reduce technique," *Egyptian Informatics Journal*, vol. 23, no. 1, pp. 55–64, 2022.
[27] C. S. Lee, P. Y. S. Cheang, and M. Moslehpour, "Predictive analytics in business analytics: Decision tree," *Advances in Decision Sciences*, vol. 26, no. 1, pp. 1–29, 2022.
[28] M. A. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, and H. Janicke, "RDTIDS: Rules and decision tree-based intrusion detection system for internet-of-things networks," *Future Internet*, vol. 12, no. 3, 44, 2020.
[29] Canadian Institute for Cybersecurity. (2017). Intrusion Detection Evaluation Dataset (CIC-IDS2017). [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html
[30] K. Nickolaos, M. Nour, E. Sitnikova, and T. Benjamin, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.
[31] M. Douiba, S. Benkirane, A. Guezzaz, and M. Azrour, "An improved anomaly detection model for iot security using decision tree and gradient boosting," *The Journal of Supercomputing*, vol. 79, no. 3, pp. 3392–3411, 2023.
[32] M. M. Pohan and B. Soewito, "Injection attack detection on internet of things device with machine learning method," *Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informatika)*, vol. 8, no. 1, pp. 204–212, 2023.
[33] M. Injadat, A. Moubayed, and A. Shami, "Detecting botnet attacks in IoT environments: an optimized machine learning approach," in *Proc. 2020 32nd International Conference on Microelectronics (ICM)*, IEEE, 2020, pp. 1–4.
[34] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar, and J. M. Peterson, "An easy-to-classify approach for the BoT-IoT dataset," in *Proc. 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, IEEE, 2021, pp. 172–179.
[35] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.
[36] D. N. M. R. Varre and J. Bayana, "A secured botnet prevention mechanism for HTTP flooding based DDoS attack," in *Proc. 2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, 2022, pp. 1–5.
[37] A. Agarwal, R. Singh, and M. Khari, "Detection of DDoS attack using ids mechanism: A review," in *Proc. 2022 1st International Conference on Informatics (ICI)*, IEEE, 2022, pp. 36–46.
[38] D. Coles, M. Peterson, S. Park, and M. Yun, "RokuControl-conducting MITM attacks on Roku," in *Proc. 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2022, pp. 440–444.
[39] P. Hack and Z. Wu. (2021). "We wait, because we know you". Inside the Ransomware negotiation economics. NCC Group. [Online]. Available: https://securitydelta.nl/media/com_hsd/report/460/document/-We-wait-because-we-know-you-Inside-the-Ransomware-negotiation-economics.pdf
[40] N. Sharma and R. Shanker, "Analysis of ransomware attack and their countermeasures: A review," in *Proc. 2022 International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, 2022, pp. 1877–1883.
[41] S. J. Weamie, "Cross-site scripting attacks and defensive techniques: A comprehensive survey," *International Journal of Communications, Network and System Sciences*, vol. 15, no. 8, pp. 126–148, 2022.
[42] D. Chicco and G. Jurman, "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.
[43] H. Dalianis, "Evaluation metrics and evaluation," *Clinical Text Mining*, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6
[44] J. O. Palacio-Niño and F. Berzal, "Evaluation metrics for unsupervised learning algorithms," arXiv preprint, arXiv:1905.05667, 2019, doi: 10.48550/arXiv.1905.05667