Journal of Computer Science

Research Article

Enhancing the CIC IoT Dataset 2023 for Improved Attack Detection through GANs Augmentation and Federated Learning

Shahad Alahmari and Noura Aleisa

Information Technology Department, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia

Article history
Received: 15-08-2024
Revised: 31-10-2024
Accepted: 20-11-2024

Corresponding Author:
Noura Aleisa
Information Technology Department, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia
Email: n.aleisa@seu.edu.sa

Abstract: The escalating frequency and sophistication of cyber-attacks on Internet of Things (IoT) devices present a pressing challenge to cybersecurity. With IoT device connections projected to exceed 42 billion by 2025, the vulnerability of these devices to cyber-attacks has never been more evident. This paper investigates the integration of Machine Learning (ML) and data augmentation, specifically Generative Adversarial Networks (GAN) and Federated Learning (FL), as innovative measures to fortify IoT security. The study aims to balance the CIC IoT Dataset 2023 using GAN-generated synthetic data and to enhance ML model performance through FL, with eXtreme Gradient Boosting (XGBoost) as the FL framework's backbone. The utilization of GAN for data augmentation addresses the persistent challenge of data imbalances in datasets. The comparison between the FL and traditional approaches in IoT security analytics reveals distinct advantages of FL, particularly in data privacy, scalability, and handling imbalanced data. While FL consistently delivers high accuracy, precision, recall, and F1-scores, the traditional approach varies more, often requiring additional data balancing and model tuning.

Keywords: Internet of Things, Privacy Preserving Mode, Security, Federated Learning, Data Augmentation

Introduction

The Internet of Things (IoT) is rapidly expanding with an approximated 42 billion connected devices by 2025 (Statista, 2023). Current research highlights a 300% surge in IoT-related attacks, with over 10.54 million incidents recorded in December 2022 (Maloo and Nikolov, 2022; Statista, 2022). Machine Learning (ML), as a new emerging technology, is able to enhance cybersecurity threat detection by enabling more accurate and efficient analysis of large datasets. The IoT has been experiencing a staggering expansion, set to reach an estimated 42 billion connected devices by 2025, marking an increase from about 8.74 billion in 2020, as shown in Fig. 1 (Statista, 2023).

This rapid growth, averaging a Compound Annual Growth Rate (CAGR) of around 25%, underscores the escalating integration of IoT into everyday life and industrial applications. However, this growth is paralleled by an increasing vulnerability to cyber-attacks. IoT devices, often characterized by inadequate security measures, have become a preferred target for cybercriminals. As projected by Cybersecurity Ventures, the anticipated worldwide expense of cybercrime is expected to hit USD 9.5 trillion in 2024 Statista (2023). Additionally, the escalating expenses associated with cybercrime impacts are foreseen to extend to $10.5 trillion by the year 2025 (Cybersecurity Ventures, 2023). The surge in IoT-related cyber-attacks has been alarming. Studies reveal that there has been an over 300% expand in such attacks during the previous years (Maloo and Nikolov, 2022). The fact that in December 2022, there were over 10.54 million recorded.

IoT incidents serves as more evidence of this tendency (Statista, 2022). The number of cyber events increased by 600% in the first quarter of 2023, while HTTP Distributed Denial-of-Service (DDoS) attacks significantly increased by 15% over the same period (Cloudflare, 2024). These attacks are not only growing in number but also in sophistication, with attackers exploiting a variety of vulnerabilities in IoT ecosystems. The vulnerabilities of IoT devices contribute significantly to this risk. It is found that over 47% of IoT devices have at least one critical vulnerability, making them susceptible to attacks such as data breaches, unauthorized access, and DDoS attacks (Aslan *et al.*, 2023). Moreover, IoT devices contribute to about 30% of

all network-based attacks, showcasing the critical need for enhanced security measures in this domain (Aslan *et al.*, 2023). The economic repercussions of these security breaches are substantial. In 2024, the estimated global cost of cybercrime is expected to reach USD 9.5 trillion, somewhat less than the estimated growth rate (Cybersecurity Ventures, 2023). Data breaches cost $4.45 million on average globally in 2023, a 15% rise in only three years that underscores the mounting financial strain on businesses (IBM, 2024). This financial impact underscores the call for powerful and scalable security approaches to protect the growing IoT infrastructure. The landscape of IoT threats is both diverse and complex. CIC IoT 2023 dataset, a significant resource for researchers, catalogs 33 different types of attacks executed across 105 IoT devices. These are divided into seven groups: DDoS, DoS, Reconnaissance, Web-based, Brute Force, Spoofing, and Mirai attacks (Neto *et al.*, 2023). Such diversity reflects the multifaceted nature of threats, ranging from DDoS attacks, which accounted for approximately 40% of all IoT security incidents, to more sophisticated spoofing attacks.

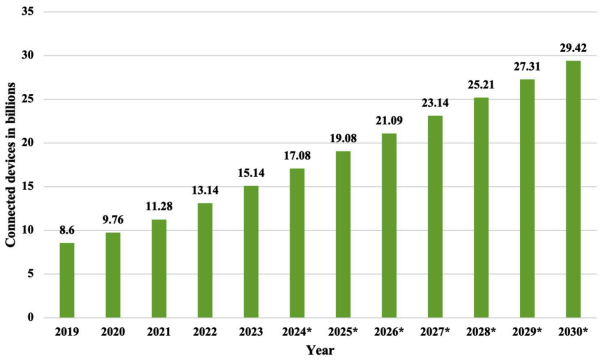

Fig. 1: IoT-connected devices globally between 2019 and 2023, with estimates ranging from 2022 to 2030 (calculated in billions)

In cybersecurity, ML is comparable to having a highly trained security guard watch over a building's access and departure points all the time. Although the guards may not be aware of every possible threat at first, as they watch how individuals arrive and leave, they begin to see trends and irregularities. Like ML algorithms that get better at detecting threats as they analyze more data, they get better at spotting suspicious activity over time. Numerous well-known ML algorithms have applications in cybersecurity. These include supervised algorithms like Random Forests, Decision Trees, and Support Vector Machines (SVMs), which categorize risks by using labeled training data (Doriguzzi-Corin and Siracusa, 2024). Deep Learning (DL) techniques, e.g., Recurrent Neural Networks (RNN) and Convolutional Neural Network (CNN) are excellent at analyzing complex data structures, such as malware code, while unsupervised algorithms like K means clustering and Hierarchical cluster in group data points to identify unusual patterns or outliers. Federated Learning (FL) is a decentralized ML approach that improves model performance while protecting data privacy (Gelenbe and Nakip, 2023). FL allows numerous devices or nodes to jointly train a global model while retaining their local data, eliminating the need to communicate sensitive data to a central server (Doriguzzi-Corin and Siracusa, 2024). It's comparable to a class of students preparing for an exam together; while each student maintains their notes and knowledge, the class as a whole increases their total understanding. FL with Differential Privacy (FLDP), Federated Averaging, and Federated Proximal are a few well-liked FL algorithms that are intended to provide security and resilience in the context of cybersecurity.

Data augmentation techniques are necessary for improving cybersecurity measures' efficacy since they diversify and strengthen the datasets used for training and evaluation. These methods entail modifying and growing the amount of data to increase the resilience of DL models and algorithms that is accessible. Generative Adversarial Network (GAN) is a popular data augmentation method in cybersecurity. GAN functions similarly to cyber-artists, with one network acting as the generator and another as the discriminator, producing fake data samples that replicate authentic cyber threats while trying to identify between the two. More complex and expanded data is produced by this adversarial process, which aids in the training of ML models to detect and neutralize a greater variety of cyber threats (Dunmore *et al.*, 2023). Other well-liked techniques that balance unbalanced datasets by either eliminating majority-class samples or duplicating minority-class samples include Random Oversampling and Under sampling. The diversity and complexity of cybersecurity data are further increased by methods like data obfuscation, noise injection, and feature shuffling, which makes it more difficult for hostile actors to take advantage of system weaknesses. This research addresses IoT security challenges through synthetic data augmentation and FL, which are used to balance the imbalanced dataset and preserve data privacy. To handle the class imbalance in the CIC IoT Dataset 2023, the GAN will be used to generate synthetic data for minority attack classes, which will reflect the characteristics of underrepresented attack types in the dataset. Further, it evaluates the effectiveness of FL in improving model performance for IoT security analytics while preserving data privacy, scalability, and handling imbalanced datasets by using a federated framework where considerable IoT devices train local models on their data subsets and collaboratively edit a global model. The effectiveness of FL will be evaluated by measuring the ML models' F1-score, recall, precision, and accuracy and comparing them to conventional centralized training methods. The study's innovative approach establishes a new standard for IoT security research by comparing the

performance of FL with traditional ML approaches to deliver insights into the benefits of using FL, specifically in data privacy, scalability, and balancing datasets. Finally, this study strives to set a new standard for IoT security by integrating synthetic data augmentation and FL, reporting a comprehensive methodology and results that future researchers can utilize to facilitate advancements in IoT security methods.

Literature Review

The literature review was conducted between 2021 and 2024, examining the domain of detecting IoT threats through the application of ML. Various noteworthy studies have been identified. These studies provide a range of topics and approaches geared towards enhancing the security and privacy of IoT networks. IoT has revolutionized multiple domains, but it has also introduced security vulnerabilities and privacy concerns. To address these challenges, researchers have explored the deploying of ML algorithms and Intrusion Detection Systems (IDS) to identify and mitigate potential threats within the IoT ecosystem. This review presents an outline of selected study efforts, each with its unique objectives, key themes, studied outcomes, and identified limitations in the context of IoT threat detection using ML.

IoT Security Threats and Vulnerabilities

The threats and vulnerabilities in IoT security include security difficulties, risk, vulnerability, and cyberattacks. IoT devices are susceptible to attacks at each layer, and defensive configurations are needed to prevent these devices from being affected. Security and privacy considerations pose significant hurdles in IoT, requiring established methods to overcome security vulnerabilities (Alamareen *et al.*, 2023). Organizations implementing IoT need to address security issues and ensure the confidentiality of data through protocols like datagram transport layer security (Parmar and Sheth, 2022). The expanding nature of IoT networks makes them vulnerable to powerful cyberattacks, and authentication attacks, such as malware posing as a legitimate device, are common. Additionally, industrial IoT generates large amounts of data, and device manipulation attacks threaten the configuration and control of IoT devices (Haque and Tasmin, 2020). The rapid proliferation of IoT devices has led to the "shadow IoT" issue. Shadow IoT refers to the use of unauthorized or unmanaged IoT devices within an organization's network (Richa, 2021). These devices often lack proper security measures and can introduce vulnerabilities into the network, making it easier for attackers to breach it. Unauthorized access, eavesdropping, man-in-the-middle attacks, unauthorized control, DDoS attacks, insecure updates, weak passwords, inadequate authentication, lack of data encryption, physical security threats, and shadow IoT all pose risks to IoT systems.

ML in IoT Security

ML algorithms analyze massive cybersecurity datasets and IoT device profiles. ML includes a number of methods, such as reinforcement learning, supervised learning, and unsupervised learning. Unsupervised learning finds patterns without labels, whereas supervised learning uses labeled data to train models. Both paradigms contribute to IoT security by detecting anomalies and predicting potential attacks. They learn to identify potential threats, both known and unknown vulnerabilities. ML models can detect IoT vulnerabilities related to weak encryption settings and configure networks to block threats (Asharf *et al.*, 2020).

ML-based IDS can effectively detect anomalies and assaults in IoT networks. These systems continuously learn from network traffic patterns and adapt to new threats. By analyzing data from various sensors and devices, they identify suspicious behavior and raise alerts. ML extracts insights from raw data to protect IoT devices against cyberattacks intelligently. DL techniques, such as CNN and RNN, enhance security intelligence. These models learn complex patterns and contribute to robust threat detection (Khan, 2021).

While ML is powerful, it faces challenges in handling dynamic IoT environments. Traditional ML methods struggle with scalability, real-time processing, and resource constraints. Researchers are exploring hybrid approaches that combine ML with domain-specific knowledge (Chen *et al.*, 2023).

Mishra *et al.* (2022) conducted a comparative analysis of ML algorithms for intrusion detection in edge-enabled IoT networks. Their work aimed to categorize network traffic using conventional ML classification algorithms on the NSL-KDD dataset. While achieving a testing accuracy of 79% with a training time of 1.2 seconds for the Multilayer Perceptron (MLP) model, it was observed that MLP relied heavily on network configuration for intrusion detection. Aqeel *et al.* (2022) focused on intrusion detection in IoT using supervised ML, particularly by leveraging application and transport layer features. Their research involved proposing TCP, MQTT, and feature clusters based on flow within the UNSW-NB15 dataset. They studied high accuracies of 97.37 and 98.67% for binary and multiclass classification, respectively, but faced challenges like over-fitting, dimensionality issues, and increased training time.

Data Augmentation Techniques

Dunmore *et al.* (2023) delved into the security implications of hardware Trojans within NoC switches of multi/many-core processors. Their research revealed that such embedded Trojans could lead to traffic analysis attacks, potentially leaking sensitive data. Training on a balanced dataset significantly increased neural network performance, marking a 15% rise in accuracy. This

advancement implies a solid step toward fortifying defenses against hardware Trojan attacks, a pressing concern in the field of cybersecurity. In the discipline of IoT security, Habibi *et al.* (2023) utilized the CTGAN model to tackle imbalanced data, achieving a noteworthy accuracy of 98.93% with an MLP classifier. Their results attest to the efficacy of data augmentation techniques in overcoming limitations inherent in previous models, thereby bolstering the detection rates of IoT botnet attacks. Concurrently, Rust-Nguyen *et al.* (2023) focused on enhancing darknet traffic classification. By utilizing AC-GAN and SMOTE to correct class imbalances, they witnessed a slight boost in accuracy. Despite this progress, they recognized the lack of robustness in classifiers as a critical area for improvement, suggesting that resilience against adversarial attacks remains a challenging frontier.

FL in IoT

In a study conducted by Zhang *et al.* (2020) to explore how FL handles the privacy and efficiency challenges of the IoT landscape by promoting collaborative model training without centralizing sensitive data. It emphasises applications in healthcare, smart cities, and autonomous driving sectors while determining challenges such as network bandwidth issues, limited device resources, and lack of standardization. The authors underline the importance of innovative solutions to leverage FL's potential in IoT systems fully. A study by Priyanka Mary Mammen on the collaborative ML technique FL, permits devices to learn a shared model without sharing their data. Introduced by Google in 2016, FL is particularly beneficial in sensitive domains like healthcare and finance, where data privacy is paramount. Other significant challenges were highlighted, including communication overhead, system and data heterogeneity, and vulnerabilities to security threats like membership inference and data poisoning attacks. The author emphasizes the need for innovative solutions to address these challenges and enhance the effectiveness of FL across diverse applications (Mammen, 2021).

Critical Evaluation of Methodologies

Richa (2021) presented a method for IoT intrusion detection using network traffic profiling and ML. Their approach actively monitored networked devices for tampering attempts and suspicious transactions, achieving an overall accuracy of 98.35%. However, the study excluded low-powered IoT devices from consideration. Mishra *et al.* (2022) focused on cyber threat intelligence for IoT using ML. Their research provides various themes such as IoT network security, anomaly detection, DDoS attack detection, and intrusion detection systems. They achieved an accuracy of 97.21% with the Random Forest algorithm. Nonetheless, the time-consuming nature of anomaly detection in virtual network analysis was identified as a limitation. Saba *et al.* (2022) addressed the security challenges of smart cities by employing ML models for IoT system protection. They studied an impressive accuracy rate of 99.7% using a voting classifier on seven datasets from the TON-IoT telemetry dataset. A limitation noted was the absence of consideration for the full spectrum of IoT devices and vulnerabilities in a smart city environment. Almomani *et al.* (2023) evaluated the efficacy of different ML classifiers, including AdaBoost, Gradient Boosting, CatBoost, and XGBoost, in detecting reconnaissance attacks on computer networks using the UNSW-NB15 dataset. Their work demonstrated significant progress in identifying reconnaissance activities within IoT networks by achieving a True Positive Rate of 90.08% and an F1-Measure of 93.57%. Similarly, Almomani *et al.* (2024) focused on predicting Denial-of-Service (DoS) attacks in IoT environments, employing various ML classifiers such as SVM, Na¨ıve Bayes, Random Forest, Logistic Regression, and Decision Tree. Utilizing the UNSW-NB15 dataset, they achieved remarkable accuracy and precision rates, with Random Forest reaching 99.4% accuracy and 99.2% precision.

In another study, Otoom *et al.* (2023) proposed a DL-based solution for accurately detecting brute force attacks on IoT networks, utilizing the MQTT-IoT-IDS2020 dataset. Their model achieved a notable accuracy of 99.56%, emphasizing the efficacy of DL in addressing cybersecurity challenges in IoT environments. Contrastingly, Zhang *et al.* (2020) investigated adversarial attacks on ML-based security systems, introducing a black-box method for generating adversarial examples. While achieving a high accuracy of 99.74% in attacking DoS scenarios, the study highlighted limitations concerning accessibility to target classifier labels. Sharma *et al.* (2023) developed a DL model for detecting Mirai botnet attacks on IoT devices, achieving precision, recall, and F1-score rates exceeding 97%. Their work emphasized the need to address challenges related to streaming data in IoT environments. Addressing specific attack vectors, Khan (2021) proposed an algorithm for detecting spoofing attacks in IoT, demonstrating effectiveness in simulated environments. Similarly, Alamareen *et al.* (2023) introduced PhishCatcher, a client-side defense mechanism against web spoofing attacks, achieving high accuracy and precision rates in detecting spoofed web pages. Moreover, Eshmawi *et al.* (2024) developed a robust detection system for GPS spoofing attacks on small UAVs, achieving an accuracy of 99.74% using ML ensemble approach. Nookala Venu *et al.* (2022) focused on detecting Mirai botnet attacks in IoT using ML techniques, showcasing high accuracy and precision rates utilizing datasets such as CICIDS2017 and CTU-13. In exploring the landscape of ML applications for security, Richa (2021) presented an intriguing study on the integration of covert backdoor attacks during data

DOI: 10.3844/jcssp.2025.1688.1704

augmentation. Their work highlighted a dual-edged outcome; augmentation improved model accuracy by 5.2%, yet it simultaneously opened up vulnerabilities due to dependency on external libraries. This dependency poses a risk, making systems susceptible to attacks, thereby creating a gap that needs addressing in future studies. Subsequently, Tomislav *et al.* (2023) contributed to this discourse by using GANs for data augmentation in predicting at-risk students. They underscored the potential of balanced datasets to enhance ML performance.

Nevertheless, they warned against the pitfalls of under sampling, which could inadvertently eliminate critical data, thus distorting the predictive accuracy of the models. Further, Strelcenia and Prakoonwit (2023) introduced K-CGAN, a new data augmentation model tailored for credit card fraud detection. Their model, adept at learning from genuine transactions, was shown to optimize the accuracy of fraud detection models by 7.3%. This significant enhancement suggests that K-CGAN could be a valuable tool in the ongoing fight against credit card fraud.

Justification for Chosen Methods

The review of existing literature highlights several key findings in the discipline of IoT security, demonstrating the effectiveness of ML models in identifying cyber threats. For instance, the uses of various ML classifiers have shown remarkable accuracy and precision in identifying attacks such as DoS, reconnaissance, and spoofing, as well as botnet activities. Studies have successfully utilized datasets like UNSW-NB15, MQTT-IoT-IDS2020, and CICIDS2017 to train models that achieve good performance metrics. However, these studies also reveal limitations, such as the challenge of overfitting, the time-consuming nature of anomaly detection, and the exclusion of low-powered IoT devices. Particularly, imbalanced datasets emerge as a pervasive challenge across the research, leading to models that might overlook less frequent but potentially more harmful cyber threats. This shows the critical need for innovative approaches to generate synthetic data and apply FL to enhance the robustness and privacy of IoT security solutions. Addressing these research gaps, our study proposes to leverage GAN for synthetic data augmentation and FL for distributed, privacy-preserving ML models, aiming to address the prevalent concern of class imbalance in the CIC IoT Dataset 2023. The existing literature indicates a nascent application of GANs in IoT security for balancing datasets and a significant underutilization of FL in this context. By integrating these methodologies, our research seeks to not only enhance the detection rates of infrequent attacks through a more balanced dataset but also to preserve data privacy in the process of collaborative learning among IoT devices. Furthermore, by establishing an enhanced dataset as a new benchmark for IoT security research, this study aims to fill the void in comprehensive approaches that combine data augmentation with advanced ML techniques in a unified framework. Hence, our research not only addresses the identified gaps but also contributes to setting a new standard for future IoT security analytics.

Materials and Methods

As the foundation of any scientific investigation, the research methodology section provides a methodical framework for conducting and evaluating research. This section defines the methodology, techniques, and processes utilized to accomplish the research goals presented in this investigation.

Proposed Methodology

The aim of this paper is to assess how well GANs and FL perform to improve security analytics for IoT settings while also methodically addressing the imbalance in the CIC IoT Dataset 2023. The steps involved are as follows (Figure 2):

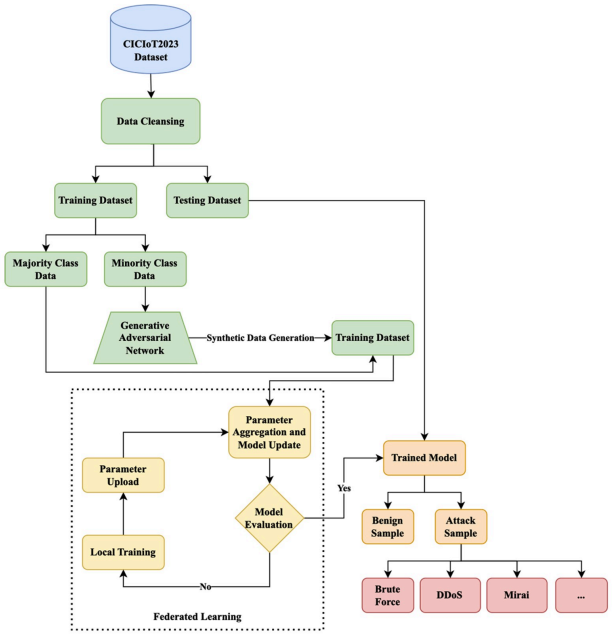

Fig. 2: Proposed methodology diagram

Data Preparation and Cleansing

Initial work will involve a detailed examination of the CIC IoT Dataset 2023 to identify any inconsistencies, missing values, or outliers that may affect the quality of the data. Data cleansing techniques, such as noise reduction, outlier removal, and missing value imputation, will be applied to guarantee the credibility of the dataset for the subsequent stages of the methodology.

Dataset Segregation

From the cleaned dataset, two distinct subsets will be produced: the training dataset and the testing dataset. The

training dataset will be used for the development and optimization of the ML models, while the testing dataset will be retained for the final evaluation of the model's performance.

Imbalance Identification

Within the training dataset, we will analyze to quantify the extent of class imbalance. This will involve statistical measures to determine the distribution of the different attack types and to identify minority and majority classes.

Generative Adversarial Network Implementation

For minority attack classes, GANs will be used to artificially create new data. Discriminator and generator, the two neural networks that make up a GAN, are trained simultaneously in a process known as competition. The Adam optimizer's beta1 parameter is set to 0.5, the learning rates are set to 0.0002, and the batch size is set to 64 for both networks. With early ending conditions if the generator reaches a loss plateau against the discriminator, the GAN is trained for a maximum of 500 epochs. The generator can now create data that is indistinguishable from real, while the discriminator can now distinguish between genuine and synthetic data. Until the generator generates data that cannot be distinguished from the real thing, this adversarial process is repeated.

Synthetic Data Generation and Integration

The GANs will generate synthetic data instances that adhere to the characteristics of the minority classes. This generated data will then be integrated with the original training data to generate a balanced dataset.

FL Setup

In the FL setup, multiple IoT devices will participate in the training process. Each device will perform local training on its subset of the balanced training dataset and compute updated model parameters. Learning rates, batch sizes, and iteration thresholds are optimized during training to prevent overfitting. As an example, the model iterates until the validation accuracy stabilizes for a batch size of 32 devices with a learning rate of 0.1. This decentralized method reduces the need for data centralization and protects privacy by guaranteeing that sensitive data stays on the device.

Parameter Aggregation and Model Update

To update the global model, the local model parameters from every device will be transferred to a central server and combined. The devices will thereafter receive the updated global model from the server for additional training. Until the model performance stops improving or satisfies predetermined criteria, this iterative process will continue. Because of its effectiveness and performance in unbalanced datasets, XGBoost is used as the basis model on every device. On a central server, the Federated Averaging (FedAvg) method is used to aggregate local model updates from devices. The FedAvg method uses the sample sizes of each device's training set to compute weighted averages.

Model Evaluation

Once training is accomplished, the model will be evaluated using a separate testing dataset. Evaluation metrics, including precision, recall, and F1-score, will be used to compare the model's performance before and after applying GAN-based data augmentation and FL training.

This detailed methodology will ensure that the research objectives are met and that the findings can enhance IoT security analytics. The methodology also includes plans for extensive documentation and the potential for reproducibility, which are critical for advancing research in this field.

Dataset Description

The CIC IoT Dataset 2023 is a comprehensive dataset designed for research and analysis in the field of IoT security. It consists of network traffic data encompassing various cyber-attack scenarios commonly encountered in IoT environments. The dataset contains 33 different attack labels, grouped into major classes, along with a multitude of features extracted from network flows Neto *et al.* (2023). 33 attacks were conducted in an IoT topology collected of 105 devices. The class distribution diagram is demonstrated in Figure 3.

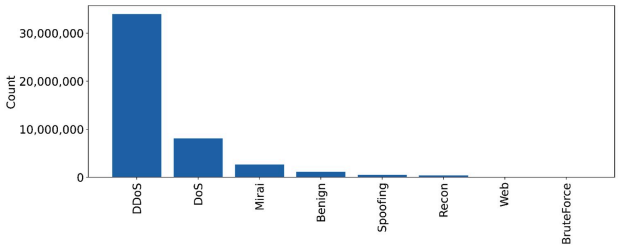

Fig. 3: The class distribution of the CICIoT2023 dataset (Neto *et al.*, 2023)

There are 33984560, 8090738, 2634124, 1098195, 486504, 354565, 24829, and 13064 instances of the DDoS, DoS, Mirai, Benign, Spoofing, Recon, Web, and Brute Force classes, respectively. From Figure 3, it is evident that the dataset is highly imbalanced. The majority of instances belong to the DDoS class, with 33984560 records, while other attack classes have comparatively lower occurrences, such as DoS, Mirai, Benign, Spoofing, Recon, Brute Force and Web Attacks. These classes are underrepresented compared to the DDoS class indicating a significant class imbalance.

Data Pre-Processing

The procedures for cleaning and preparing the data for additional analysis are described in this paragraph.

These procedures include data normalization, encoding categorical variables, addressing missing and duplicate values, and eliminating outliers (Zelaya, 2019).

Missing values in the dataset can hinder the analysis and modeling process. Therefore, it is essential to address them appropriately. All features in the dataset might have missing values identified. Duplicate values in the dataset can skew analysis results and model performance. Duplicate records are determined based on all feature values. Duplicate records will be removed, retaining unique instances in the dataset. The dataset integrity is assessed post-removal to ensure the preservation of essential information. Outliers can significantly affect the performance of ML models. Therefore, outlier removal techniques are employed to enhance model robustness. Outliers are detected using statistical methods such as z-score, Inter Quartile Range (IQR), or domain-specific knowledge. Outliers were removed from the dataset.

Categorical variables in the dataset require transformation into numerical representations for compatibility with ML algorithms. For nominal categorical variables, one-hot encoding was performed to generate binary columns for every category. Data normalization is applied to scale numerical features within a consistent range, facilitating convergence and improving model performance. To preserve the relative relationships between data points, numerical features are scaled to a specific range, usually between 0 and 1 (Zelaya, 2019).

By systematically addressing missing values, duplicate values, outliers, encoding categorical variables, and normalizing data, the CIC IoT Dataset 2023 is prepared for effective analysis and modeling, ensuring the integrity and reliability of research findings and conclusions.

Synthetic Data Generation Using GAN

Synthetic data generation refers to creating artificial data samples that resemble real data instances but are generated algorithmically rather than being observed or collected from real-world sources. This technique is essential in various fields, including ML, where access to labeled data may be limited or expensive. Synthetic data generation allows researchers and practitioners to augment existing datasets, balance class distributions, and create diverse training data for building robust ML models. Furthermore, by creating data that anonymizes sensitive information while maintaining the statistical characteristics of the original dataset, synthetic data can help allay privacy concerns and facilitate safer data research and sharing.

There are a number of methods used to create synthetic data, and each has advantages and disadvantages.

A common technique is the use of GANs, which simultaneously train a discriminator and a generator neural network. It learns to produce synthetic data samples that are precise duplicates of authentic data. It gives the discriminator the ability to distinguish between authentic and synthetic data. A different method uses probabilistic models, like Variational Autoencoders (VAEs), which train a low-dimensional data representation and use the learned latent space distribution to sample new data. Additionally, to create synthetic data, conventional statistical procedures like bootstrapping and data augmentation methods like SMOTE are used (Fonseca and Bacao, 2023).

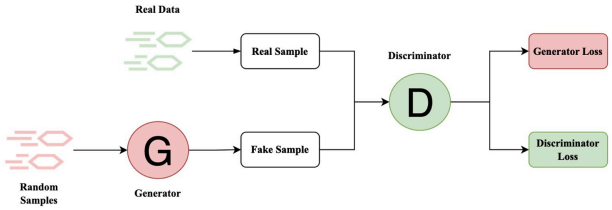

Fig. 4: Basic architecture of GAN Adapted from GAN Structure (Google Developers, 2022)

GANs stand out as a powerful and versatile tool for synthetic data generation in various domains, e.g., medical data, financial data, etc., offering significant advantages over traditional techniques.

Introduction to GAN

Generative Adversarial Networks, or GANs, have emerged as a revolutionary paradigm in ML, especially in the field of generative modeling. GANs, first presented by Ian Goodfellow and colleagues in 2014 (Dunmore *et al.*, 2023), have fundamentally changed how we tackle the problem of generating new data samples that closely resemble those in a given dataset. Generator networks and discriminator networks are two neural networks that can be competitively pitted against one another. This is the basic idea of GANs. This novel adversarial configuration facilitates a dynamic learning process in which the discriminator aims to discern between synthetic and genuine data, while the generator attempts to generate realistic data samples (Dunmore *et al.*, 2023). The development of new and realistic data examples is made possible by GANs' ability to produce data distributions that closely resemble the training data.

Working Principle of GAN

GANs operate by utilizing the new idea of adversarial training, in which two neural networks—the discriminator and the generator—participate in a competitive learning process. Learning a distribution of data that is similar to the training data distribution is the aim of a GAN. The generator network seeks to produce synthetic data samples that are identical to real data samples by using random noise as input. On the other

hand, the discriminator network's task is to differentiate between real data samples from the training dataset and fake data samples that the generator creates (Dunmore *et al.*, 2023). Figure 4 shows the basic architecture of GAN.

A GAN's training process alternates between two stages: the discriminator phase and the generator phase. Random noise is used as input by the generator network to create synthetic data samples during the generator phase. The discriminator is subsequently given these artificial samples with the goal of accurately classifying them as false. Following this, the discriminator learns by combining actual data samples from the training dataset with fake data samples that were produced by the generator during the discriminator stage. To improve its capacity to distinguish between real and phony samples, the discriminator adjusts its settings. At the same time, the discriminator provides feedback to the generator, which modifies its parameters to create increasingly realistic synthetic samples that are more difficult for the discriminator to distinguish from actual data. In a dynamic game of cat and mouse, the discriminator and generator networks compete with one another to outperform the other as training goes on. Both networks develop iteratively as a result of this adversarial learning process; the discriminator gets better at differentiating between real and fake data, while the generator eventually learns to create more realistic synthetic samples. In the end, competition reaches a point where the generator generates synthetic data that is nearly identical to real data and the discriminator is unable to distinguish between real and synthetic samples. At this point, the GAN has successfully learned the underlying data distribution, enabling it to generate novel and realistic data samples consistent with the training dataset.

Integration of Synthetic Data with the Original Dataset

Once synthetic data has been generated using techniques such as GANs, it is crucial to integrate this synthetic data seamlessly with the original dataset while ensuring the preservation of data integrity and maintaining representativeness across classes. The integration process involves several steps to effectively merge the synthetic data with the original dataset (Joshi *et al.*, 2024):

- **Concatenation:** The simplest approach to integrate synthetic data with the original dataset is by concatenating the synthetic samples with the existing data. This involves adding the synthetic data instances as additional rows or observations to the original dataset while preserving the feature structure and labels.
- **Balancing Class Distributions:** Synthetic data generation is often employed to address class imbalance issues in the original dataset. Therefore, during integration, special attention should be given to balancing the class distributions by strategically incorporating synthetic samples for minority classes. This helps ensure that each class is adequately represented in the integrated dataset, leading to more robust machine-learning models.
- **Randomization:** To prevent biases and maintain the randomness of the data, synthetic samples should be integrated with the original dataset in a randomized manner. This involves shuffling the combined dataset to ensure that the synthetic samples are distributed evenly across different sections of the dataset, preventing any systematic biases that may arise from the integration process.

Evaluation Metrics for Synthetic Data Quality

To ensure that the produced samples accurately reflect the underlying data distribution, it is crucial to evaluate the quality of synthetic data produced by GANs or other methods. The fidelity and usefulness of synthetic data in comparison to real data can be measured using a variety of evaluation metrics. The quality of synthetic data is frequently assessed using the evaluation measures listed below (Vujović, 2021):

- **Frechet Inception Distance (FID):** A widely used metric for comparing the distributions of synthetic and actual data is called FID. Using a pre-trained Inception network, it determines the Wasserstein-2 distance between the multivariate Gaussian distributions of feature embeddings taken from synthetic and real data.
- **Inception Score (IS):** IS assesses the variety and caliber of artificial images produced by GANs. It penalizes high-confidence predictions and calculates the entropy of class predictions based on synthetic images using a pre-trained classifier (such as the Inception network). Better performance is shown by higher IS metrics, which assess the quality and diversity of generated samples.

FL for Privacy-Preserving Classification

With its innovative method for training models on decentralized data sources while maintaining data security and privacy, FL has become a ground-breaking paradigm in the field of ML. In order to train their models, traditional ML models need centralized data aggregation, which gathers and stores data from multiple sources on a single server. This strategy, however, presents serious privacy issues because private user information could be vulnerable to hacks or illegal access. FL addresses these challenges by enabling model training directly on the local devices or servers where the data resides, without the need for data sharing.

Conceptual Framework of FL

FL presents a new paradigm for cooperative model training over dispersed data sources while preserving the confidentiality and privacy of data. According to FL's

DOI: 10.3844/jcssp.2025.1688.1704

conceptual framework, decentralized model training is the process by which several devices or edge servers work together to jointly build a global model without exchanging raw data. Figure 5 depicts the anatomy of a basic FL. This FL framework can be elucidated through several key components and processes (Jiang *et al.*, 2020):

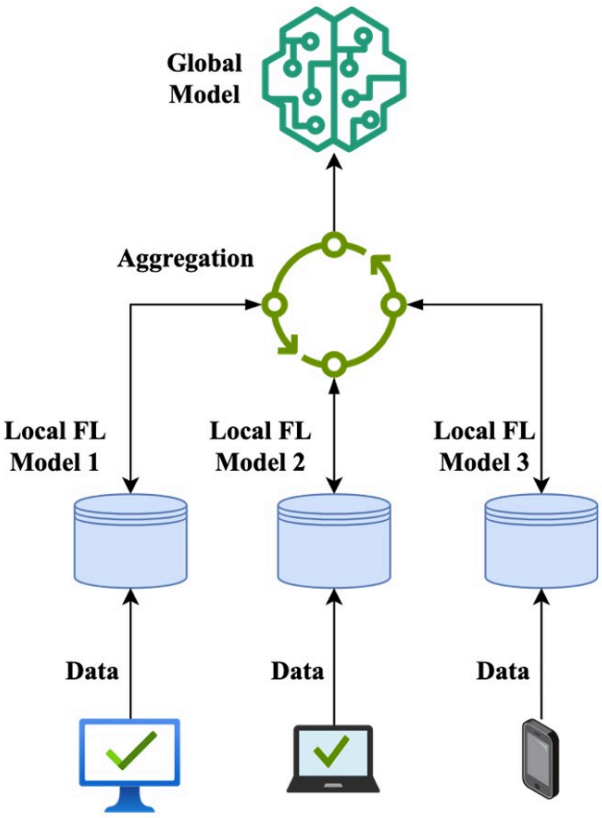

Fig. 5: Architecture of a simple FL approach Adapted from FL in smart city sensing: challenges and opportunities (Jiang *et al.*, 2020)

- **Local Training:** In FL, data is generated or stored locally on edge servers or individual devices, where model training takes place. Using local data and local computational resources, each local device trains a model separately. Data privacy is maintained by this local training method, which enables devices to learn from their own data without disclosing it to a central server or other devices.
- **Model Aggregation:** Following local training iterations, model updates are sent to a central server or aggregator for aggregation in the form of model parameters or gradients. The global model is updated by the aggregator, which receives model updates from several devices and aggregates them. Averaging and weighted aggregation are two examples of aggregation techniques that can be used to efficiently integrate model updates while maintaining the global model's quality.
- **Communication Protocol:** Communication between local devices and the central server is facilitated through secure and efficient communication protocols. These protocols ensure the transmission of model updates while preserving data privacy and security. Techniques, e.g., SMC and differential privacy may be employed to encrypt or obfuscate communication to protect sensitive data during transmission.
- **Iterative Optimization:** Model aggregation and local training are carried out repeatedly until convergence or a predetermined stopping criterion is satisfied in FL's usual iterative optimization method. Each iteration involves training local models with local data, then aggregating model updates to improve the global model. Through iterative optimization, the global model's performance is continuously improved while collaborative learning across dispersed data sources is made possible.
- **Evaluation and Validation:** Throughout the FL process, evaluation and validation mechanisms are employed to assess the performance and quality of the global model. Metrics, e.g., loss, accuracy, and convergence rate may be monitored to evaluate the effectiveness of FL in achieving the desired learning objectives. Additionally, techniques such as differential privacy analysis may be applied to validate the privacy guarantees provided by FL approaches.

The conceptual framework of FL encompasses decentralized model training, secure communication, iterative optimization, and evaluation mechanisms, enabling collaborative learning across distributed data sources while preserving data privacy and security. This framework lays the foundation for the practical implementation and deployment of FL in various real-world applications and domains.

eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting is referred to as XGBoost. Gradient boosting is implemented in an effective and scalable manner, an ML technique used for regression, classification, and ranking problems (Sagi and Rokach, 2021). XGBoost has gained popularity in ML competitions and industry applications due to its performance and speed. In contrast to conventional gradient boosting, XGBoost builds upon this base in several significant ways, including the following:

- **Gradient Boosting Framework:** The core idea behind XGBoost is gradient boosting. In a sequential approach, it builds a series of decision trees, each one attempting to correct the errors of its predecessors. By employing a gradient descent technique, the model reduces the loss when adding more models.

- **Regularization:** Model complexity is managed by XGBoost by the inclusion of a regularization term in its goal function. By doing so, overfitting is lessened, which is a common problem with standard gradient boosting methods.
- **System Optimization:** XGBoost has been designed to be highly efficient and scalable. It utilizes both hardware optimization (such as multi-threading on a single machine) and software optimization techniques (like cache awareness and block structure for out-of-core computation) to achieve high performance on large datasets.
- **Sparsity Awareness:** XGBoost can automatically handle missing data. This means that it can still make splits even if some data are missing, which can be particularly useful for sparse datasets.
- **Tree Pruning:** While traditional gradient boosting uses a depth-first approach for tree generation, XGBoost uses a depth-wise approach and prunes trees using a depth-first approach once a maximum depth is reached. This results in more optimized trees.

XGBoost is known for delivering high performance and speed compared to other implementations of gradient boosting. XGBoost supports classification, regression, ranking, and user-defined prediction problems. It can be used in conjunction with several programming languages, including Python, R, Java, and Scala. Its built-in regularization helps avoid overfitting, making it more effective on a wide range of datasets. XGBoost can automatically learn the best way to handle missing data. XGBoost offers a wide range of parameters that can be tuned for optimal performance, such as learning rate, depth of trees, and regularization terms (Sagi and Rokach, 2021).

Incorporating XGBoost into a FL framework leverages the strength of XGBoost in handling large-scale and highdimensional data with high efficiency and accuracy, while FL guarantees that the data perseveres on the local devices, preserving privacy and reducing the risk of data leakage.

Evaluation Metrics

To evaluate various elements of model performance, a range of evaluation measures are used in the assessment of ML models, including those trained using FL methods. The predicted accuracy, class-wise performance, and overall efficacy of the model are all revealed by these indicators. FL algorithms are frequently assessed using the assessment measures listed below (Vujović, 2021):

Accuracy: The model's overall correctness is measured by how accurate its predictions are. The equation to calculate accuracy is as follows:

$$Accuracy = \frac{TN+TP}{FN+FP+TP+TN} \quad (1)$$

Precision: The model's precision is determined by dividing all of its positive predictions by the percentage of real positive predictions. The following formula can be used to determine precision:

$$Precision = \frac{TP}{FP+TP} \quad (2)$$

Recall: The percentage of genuine positive predictions among all actual positive cases in the dataset is known as recall, which is also frequently called true positive rate or sensitivity. The following formula is used to determine recall:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score: The harmonic mean of precision and recall, or F1-score, provides a reasonable assessment of a model's performance. Here is the formula to determine the F1-score:

$$F1\text{-}Score = 2 \cdot \frac{Recall \cdot Precision}{Recall+Precision} \quad (4)$$

Metrics from GANs can also be relevant, particularly when evaluating synthetic data quality. Some common metrics for evaluating GAN-generated data include:

- FID: Evaluates how closely the distributions of synthetic and real data are similar.
- IS: Evaluates the variety and quality of artificial images produced by GANs.

Results and Discussion

The study's results and analysis present an in-depth understanding of how well the different approaches used in IoT security analytics performed. Expanding on this approach, the assessment of GAN-based synthetic data augmentation provides insightful information on how well the CIC IoT Dataset 2023 addresses class imbalance. To offer a greater understanding of the influence of synthetic data augmentation on model performance and overall security analytics, this section explores the particular metrics and improvements that are attained using this technique.

Evaluation of GAN-Based Synthetic Data Augmentation

It is crucial to assess the synthetic data produced by GAN in order to determine its quality and appropriateness for improving the CIC IoT Dataset 2023. The integrity and diversity of the produced synthetic data have been assessed using two important metrics: IS and FID Score.

The FID is a widely accepted metric for evaluating the similarity between two datasets in terms of their distribution of features. The real and synthetic datasets are more comparable when the FID score is lower. In our evaluation, the calculated FID score of 0.0371 suggests a remarkable closeness between the synthetic data and the real dataset, as shown in Figure 6. The obtained FID

DOI: 10.3844/jcssp.2025.1688.1704

score well below 0.1 indicates a high level of fidelity in the synthetic data generation process. According to this, the distribution of features in the synthetic data is quite similar to that seen in the actual CIC IoT Dataset 2023. A low FID score confirms that the synthetic data is realistic and can faithfully capture the features of minority attack classes in the dataset. The close alignment between the synthetic and real datasets ensures that the augmented dataset maintains its authenticity, thereby bolstering the credibility and applicability of the dataset in IoT security research scenarios.

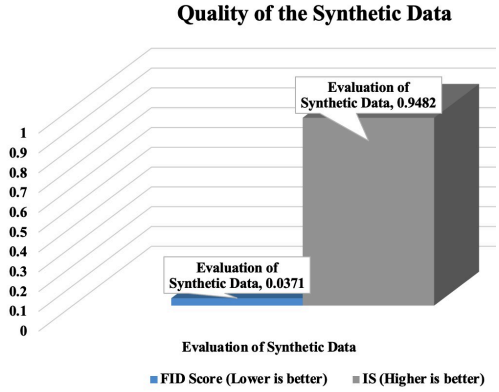

Fig. 6: Synthetic data evaluation based on IS and FID scores

Conversely, the IS serves as a tool for assessing the caliber and variety of data produced by GANs. By evaluating the predictability of labels, it gauges the synthetic data's diversity and quality. Better quality and diversity of the produced data are indicated by a higher IS. As seen in 6, the computed IS of 0.9482 indicates a good degree of quality and variety in the synthetic data. This suggests that the created data enriches the dataset by adding variability in addition to capturing the distribution of attributes. A high IS highlights the GAN's capacity to produce a variety of synthetic data examples, guaranteeing that the augmented dataset includes a broad spectrum of attack scenarios and variants. The high IS raises trust in the synthetic data's resilience and suggests that it may be used to train and assess ML models for IoT security analytics.

The evaluation of the GAN-based synthetic data augmentation using FID Score and IS demonstrates its effectiveness in producing high-quality, diverse, and realistic data. These metrics validate the suitability of the synthetic data for rectifying the class imbalance in the CIC IoT Dataset 2023, thereby bolstering the research objectives aimed at enhancing IoT security analytics.

Performance of FL without Synthetic Data

The evaluation of FL without synthetic data augmentation, conducted on 10, 15, and 20 client devices, provides insights into the effectiveness of this approach in training ML models for IoT security analytics using the original imbalanced CIC IoT Dataset 2023, as depicted in Figure 7.

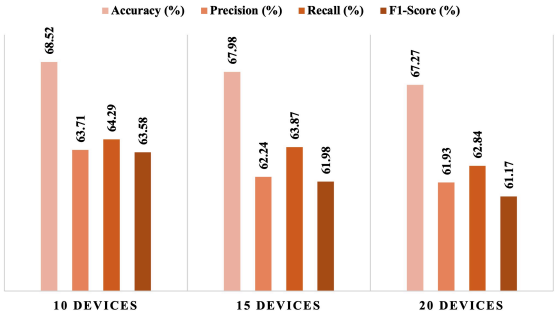

Fig. 7: Performance of FL without synthetic data

For 10, 15, and 20 client devices, respectively, the aggregated model's accuracies were 68.52, 67.98, and 67.27%, as seen in Figure 7. 63.71, 62.24, and 61.93% were the precisions of the aggregated model for 10, 15, and 20 client devices, respectively. For 10, 15, and 20 client devices, respectively, the aggregated model showed recalls of 64.29, 63.87, and 62.84%. F1-scores for 10, 15, and 20 client devices were 63.58, 61.98, and 61.17%, respectively, for the aggregated model.

The consistent performance across varying numbers of client devices suggests that FL is effective in training ML models for IoT security analytics using distributed data sources. Despite the imbalanced nature of the original dataset, FL demonstrates its potential to learn from decentralized data while preserving data privacy. The slight decrease in performance metrics as the number of client devices increases may indicate challenges associated with model aggregation and coordination across a larger number of decentralized devices. However, the relatively small fluctuations in performance metrics suggest that FL maintains robustness and scalability across different device configurations. The performance metrics of the aggregated model without synthetic data augmentation highlight the inherent challenges posed by a class imbalance in the original dataset. While FL shows promise in addressing data privacy concerns and leveraging distributed computing resources, its efficacy in handling imbalanced datasets may be limited without additional strategies such as synthetic data augmentation.

The evaluation of FL without synthetic data augmentation underscores its effectiveness in training ML models for IoT security analytics using decentralized data sources. However, the performance metrics also highlight the importance of addressing class imbalance in the dataset to further enhance model performance and robustness.

Performance of FL with Synthetic Data

The evaluation of FL with synthetic data augmentation was conducted on 10, 15, and 20 client devices, as shown in Figure 8. It provides crucial insights into the effectiveness of this approach in mitigating class imbalance and enhancing model performance for IoT security analytics using the augmented CIC IoT Dataset

2023. To evaluate the effectiveness of the FL framework with synthetic data augmentation, key performance measures including as F1-score, recall, precision, and accuracy have been evaluated. The aggregated model achieved exceptionally high accuracies of 95.82, 95.76, and 95.59% for 10, 15, and 20 client devices, respectively. The aggregated exhibited high precision scores of 94.44, 95.08, and 94.84% for 10, 15, and 20 client devices, respectively. The aggregated model demonstrated recall rates of 95.19, 95.65, and 95.26% for 10, 15, and 20 client devices, respectively. The aggregated model achieved high F1-scores of 95.02, 95.57, and 95.18% for 10, 15, and 20 client devices, respectively.

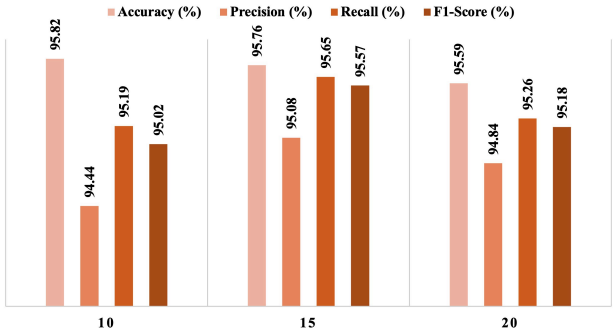

Fig. 8: Performance of FL with synthetic data

The substantial increase in all performance metrics compared to FL without synthetic data augmentation underscores the effectiveness of synthetic data in addressing class imbalance and enhancing model performance. The augmented dataset, enriched with synthetic data, provides a more representative and balanced training environment, leading to significantly improved model F1-score, recall, precision, and accuracy. The consistently high performance across different numbers of client devices highlights the robustness and scalability of the FL framework with synthetic data augmentation. Despite the decentralized nature of the training process involving multiple client devices, the aggregated model maintains high levels of accuracy and performance, indicating the reliability of the approach. The augmented dataset enables the FL framework to effectively capture the characteristics of minority attack classes, thereby improving the model's ability to detect and classify diverse cyber threats in IoT environments. The heightened performance metrics validate the efficacy of synthetic data augmentation in bolstering IoT security analytics and addressing the challenges posed by imbalanced datasets.

The evaluation of FL with synthetic data augmentation demonstrates its capability to significantly enhance model performance for IoT security analytics. The augmented dataset facilitates more accurate and robust model training, paving the way for more effective cyber threat detection and classification in IoT ecosystems.

Comparative Performance of With and Without Synthetic Data Using FL Approach

The comparison between the performance of FL with and without synthetic data augmentation provides critical insights into the efficacy of incorporating synthetic data to address class imbalance and enhance model performance for IoT security analytics. To assess the influence of synthetic data augmentation on model efficacy, key performance measures such as F1-score, recall, precision, and accuracy have been examined.

FL with synthetic data augmentation consistently outperforms FL without synthetic data across all configurations of client devices, as presented in Table 1. The augmented dataset, enriched with synthetic data, leads to a significant increase in accuracy, indicating a higher proportion of correctly classified instances. To increase the overall efficacy of IoT security analytics models, this development emphasizes how crucial it is to resolve class imbalance through synthetic data augmentation. Precision exhibits notable improvement with synthetic data augmentation. The higher precision values obtained with synthetic data augmentation indicate a reduced rate of false positives, signifying improved model reliability in identifying true attacks while minimizing false alarms. Recall shows consistent enhancement with synthetic data augmentation. The augmented dataset enables the model to capture a higher proportion of true positive instances, resulting in improved detection rates for minority attack classes and reducing the risk of overlooking critical security threats. The higher F1-scores obtained with synthetic data augmentation reflect a more balanced trade-off between precision and recall, indicating improved overall model effectiveness in handling imbalanced datasets. The consistent performance improvement across all performance metrics and configurations of client devices underscores the generalizability and robustness of synthetic data augmentation in enhancing FL for IoT security analytics. The augmented dataset facilitates more representative and balanced model training, leading to improved model generalization and robustness in detecting diverse cyber threats in IoT environments. This comparative analysis highlights the significant performance improvement achieved through synthetic data augmentation in FL for IoT security analytics, emphasizing its critical role in addressing class imbalance and enhancing model effectiveness.

Performance of ML Models with Traditional Approach

Using the original CIC IoT Dataset 2023, the performance of ML models using the traditional approach—both with and without synthetic data augmentation—offers important insights into how well different algorithms address class imbalance and improve model performance for IoT security analytics.

Decision Tree

With precision, recall, and F1-score values ranging from 63 to 64%, the Decision Tree model achieved an accuracy of 63.96%, as shown in Table 2. In terms of IoT security threat classification, this conventional method without artificial data performs mediocrely. With synthetic data augmentation, the Decision Tree model's performance increased dramatically, reaching precision, recall, and F1score values over 88% and an accuracy of 89.72%. This notable enhancement demonstrates how well synthetic data may reduce class imbalance and boost model performance.

Table 1: Comparative performance of with and without synthetic data using FL approach

Number of Devices	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
10	27.30	30.73	30.90	31.44
15	27.78	32.86	31.78	33.59
20	28.32	32.91	32.42	34.01

Naive Bayes

As shown in Table 2, the accuracy of the Naive Bayes model was 59.86%, while the precision, recall, and F1score values were about 59-60%. When compared to other algorithms, this model's performance without artificial data augmentation indicates a lesser efficacy in identifying IoT security issues. The Naive Bayes model's performance increased dramatically with the addition of synthetic data, reaching an accuracy of 93.32% with precision, recall, and F1-score values over 92%. This significant improvement emphasizes how important synthetic data is for boosting model efficacy, especially for methods like Naive Bayes that have built-in drawbacks.

Support Vector Machine (SVM)

With precision, recall, and F1-score values ranging from 61 to 62%, the SVM model achieved an accuracy of 61.69%, as shown in Table 2. Even if SVM performs mediocrely in the absence of synthetic data, it may be better. With the addition of synthetic data, the SVM model's performance significantly improved, attaining precision, recall, and F1-score values over 86% and an accuracy of 86.05%. Especially for algorithms like SVM, this notable improvement highlights how effective synthetic data is in correcting class imbalance and improving model performance.

XGBoost

According to Table 2, the XGBoost model showed an accuracy of 63.60%, with precision, recall, and F1-score values ranging from 63 to 64%. Without synthetic data, XGBoost's performance is mediocre, although it might be improved. The addition of synthetic data significantly improved the XGBoost model's performance, with accuracy of 94.62% and precision, recall, and F1-score values over 94%. This substantial improvement shows how much synthetic data may enhance model performance for powerful algorithms like XGBoost.

The benefit of synthetic data augmentation in reducing class imbalance and enhancing model efficacy for IoT security analytics is demonstrated by the performance of ML models using the conventional method. The substantial performance improvement across various algorithms underscores the importance of leveraging synthetic data to address inherent challenges in imbalanced datasets and enhance the reliability and robustness of IoT security analytics models.

Comparison between the FL Approach and Traditional Approach

The efficacy of each methodology in training ML models for IoT security analytics is revealed by contrasting the FL approach with the conventional approach. Table 3 compares key performance indicators including as F1-score, recall, precision, and accuracy to assess the relative advantages and disadvantages of each strategy.

F1-Score, Recall, Precision, and Accuracy

When compared to the conventional method, the FL technique often yields greater values for F1-score, recall, precision, and accuracy. This is explained by FL's decentralized training methodology, which efficiently addresses class imbalance and protects privacy while enabling models to learn from a variety of data sources.

Data Privacy

The FL technique reduces the possibility of exposing private data to centralized servers by enabling model training to be carried out locally on client devices. In contrast, the traditional approach may compromise data privacy, especially when dealing with centralized data storage.

Scalability

FL approach exhibits high scalability as it leverages distributed computing resources across multiple client devices. This enables FL to handle large-scale datasets and accommodate increasing numbers of devices seamlessly. On the other hand, the traditional approach may face scalability challenges, particularly when dealing with large volumes of data and extensive computational requirements.

Imbalanced Data

FL approach addresses imbalanced data effectively by leveraging techniques such as synthetic data augmentation and distributed learning. This enables FL to create more balanced training datasets and improve model performance for minority classes. In contrast, the

traditional approach may require additional strategies to handle class imbalance, such Table 3. FL vs. traditional approach for IoT security analytics.

Table 2: Performance of ML models with traditional approach

Model	Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	Without Synthetic	63.96	63.89	63.73	62.73
	With Synthetic	89.72	89.95	89.14	88.14
Naive Bayes	Without Synthetic	59.86	59.74	59.07	59.29
	With Synthetic	93.32	93.24	93.09	92.73
Support Vector Machine	Without Synthetic	61.69	61.21	61.40	61.03
	With Synthetic	86.05	86.62	88.54	87.27
XGBoost	Without Synthetic	63.60	63.08	63.23	62.91
	With Synthetic	94.62	94.55	94.26	94.13

Table 3: FL vs traditional approach for IoT security analytics

Metric	FL	Traditional Approach
Accuracy (%)	Superior	Moderate to Good
Precision (%)	Consistently High	Varied (Moderate to High)
Recall (%)	Outstanding	Moderate to High
F1-Score (%)	Excellent	Moderate to Good
Data Privacy	Ensured	May Compromise
Scalability	Strong	Limited to Moderate
Imbalanced Data	Effectively Managed	Requires Additional Handling
Model Generalization	Robust	May Require Manual Adjustments

Model Generalization

FL approach enhances model generalization by training on diverse data from multiple devices. This improves the model's ability to adapt to unseen data and several IoT environments. The traditional approach may require finetuning to achieve similar levels of model generalization.

While both FL and the traditional approach have their respective strengths and limitations, FL emerges as a promising methodology for IoT security analytics, offering high performance, data privacy preservation, scalability, and effective handling of imbalanced data. However, in the end, the particular needs and limitations of the IoT security application will determine which of the two strategies is best.

Comparative Result with Previously Published Works

The comparison of different machine learning models for detecting IoT cyberattacks using the CIC 2023 IoT dataset in Table 4 reveals distinct performance outcomes while revealing critical limitations in existing studies. Previous studies such as Jony and Arnob (2024a); Neto *et al.* (2023); Jony and Arnob (2024b) have shown that traditional methods, including Logistic Regression, KNN, Decision Trees, and Random Forests, achieve up to 99.16% accuracy. However, challenges such as overfitting, the time-consuming nature of anomaly detection, and Imbalanced datasets remain, causing models to overlook less frequent but potentially harmful cyber threats. In contrast, advanced techniques like Generative Adversarial Networks (GANs) and Federated Learning (FL) significantly enhance performance, achieving accuracy rates of 95.3% in our study, improving infrequent attack detection rates and preserving data privacy through collaborative learning. By establishing an enhanced dataset as a new benchmark for IoT security research, we fill the void in comprehensive approaches combining data augmentation with advanced ML techniques, setting a new standard for future IoT security analytics.

Table 4: Comparative Result with other Works

Authors	Dataset	Methods	Purpose of Study	Evaluation Matrices	Limitation
Jony and Arnob (2024a)	CIC IoT Dataset 2023	LSTM	Detecting cyber attacks in IoT using the CIC IoT 2023 dataset	(Accuracy 0.9875), (F1 Score 0.9859), (Recall Score 0.9875), (Precision Score 0.9866)	Needs improvement in interpretability and scalability
Neto *et al.* (2023)	CIC IoT Dataset 2023	Logistic Regression, Perceptron, Adaboost, Random Forest, and Deep Neural Network	Propose a novel IoT attack dataset for security analytics	LR (Accuracy 0.8023), Perceptron (Accuracy 0.8196), Adaboost (Accuracy 0.6079), RF (Accuracy 0.9916), DNN (Accuracy 0.9861)	Limited to specific types of attacks; requires further optimization
Jony and Arnob (2024b)	CIC IoT Dataset 2023	Logistic Regression, KNN, Decision Tree, Random Forest	Evaluate ML algorithms for detecting IoT cyberattacks	RF (Accuracy 0.9916), KNN (Accuracy 0.9380), DT (Accuracy 0.9919), LR (Accuracy 0.8275)	Limited to specific attack types; LR performed the least effectively
This study	Enhanced CIC IoT Dataset 2023	Decision Tree, Naive Bayes, Support Vector Machine, XGBoost	Enhance IoT security through GANs and Federated Learning	DT (Accuracy 89.72), NB (Accuracy 93.32), SVM (Accuracy 86.05), XGBoost (Accuracy 94.62)	Computationally intensive; risks during aggregation

DOI: 10.3844/jcssp.2025.1688.1704

Conclusion

The advent of the IoT has revolutionized numerous industries, offering unprecedented connectivity and convenience: However, with this connectivity comes the inevitable challenge of ensuring robust security measures to safeguard IoT ecosystems against evolving cyber threats In this paper, it has been investigated the effectiveness of different approaches, specifically traditional ML and FL, in enhancing IoT security analytics using the CIC IoT Dataset 2023 Our investigation commenced with a comparative analysis of performance between traditional ML approaches and FL Traditional ML models, including Decision Tree, Na¨ıve Bayes, SVM, and XGBoost, were initially trained using the original dataset without synthetic data augmentation

These models exhibited moderate performance, achieving accuracy scores ranging from 59.86 to 63.96%.

However, their effectiveness was significantly enhanced when synthetic data augmentation was applied, with accuracy scores ranging from 86.05% to 95.82%. FL outperformed traditional ML approaches across all metrics, Presenting greater F1-score, recall, precision, and accuracy values. The FL framework leveraged the collaborative learning capabilities of multiple IoT devices while preserving data privacy and effectively addressing class imbalance through techniques like synthetic data augmentation. The substantial improvements observed in model performance with synthetic data augmentation underscore the critical role of addressing class imbalance in IoT security analytics. Synthetic data generation enabled the models to learn from more representative datasets, thereby enhancing their ability to detect and classify diverse cyber threats. Moreover, FL exhibited superior scalability and data privacy preservation compared to traditional ML approaches. By decentralizing model training and leveraging local data sources, FL mitigated the risk of exposing sensitive information to centralized servers while accommodating large-scale IoT environments seamlessly. These enhancements not only contribute to more accurate threat detection but also lay the foundation for scalable and privacy-preserving security analytics in IoT ecosystems.

The outcomes of this paper have major ramifications for IoT security practitioners and researchers. Firstly, the adoption of FL holds promises for enhancing the effectiveness and scalability of security analytics in IoT environments. By harnessing the collective intelligence of distributed devices, FL enables more robust threat detection and classification while safeguarding data privacy. Secondly, the integration of synthetic data augmentation techniques addresses the inherent challenges posed by imbalanced datasets, enabling ML models to better capture the nuances of minority attack classes. This, in turn, facilitates more comprehensive and accurate security analytics, reducing the risk of overlooking critical security threats. While this paper provides insightful findings into the efficacy of FL and synthetic data augmentation for IoT security analytics, several avenues for future research exist. Firstly, exploring advanced FL techniques such as differential privacy and secure aggregation could further enhance data privacy while maintaining model performance. Additionally, investigating novel synthetic data generation methods tailored specifically for IoT security datasets could lead to more effective model training and threat detection. Furthermore, longitudinal studies evaluating the long-term effectiveness and scalability of FL in real-world IoT deployments are warranted to validate its practical applicability.

This study underscores the critical importance of adopting innovative approaches like FL and synthetic data augmentation to enhance IoT security analytics. By leveraging distributed learning and addressing the class imbalance, FL offers a scalable, privacy-preserving framework for robust threat detection in IoT ecosystems. The significant improvements observed in model performance highlight the potential of these approaches to fortify cyber defense capabilities and mitigate evolving dangers in the rapidly evolving area of IoT. As we continue to explore the intersection of ML and IoT security, embracing innovative methodologies will be paramount in safeguarding the integrity and robustness of IoT systems countering evolving cyber threats. As the domain of IoT security continues to evolve, multiple promising research directions and development can further enhance the effectiveness of security analytics and mitigate emerging cyber threats. Integrating multi-modal data sources, such as sensor data, network traffic, and device logs, can enrich the training datasets and improve the robustness of security analytics models. Future work could explore methodologies for effectively integrating and leveraging diverse data sources to enhance threat detection abilities. Also, we will explore additional data balancing techniques, including SMOTE, ADASYN, and hybrid approaches to provide a comprehensive view of their comparative effectiveness in IoT security contexts.

Limitations

This study has faced limitations regarding deploying GAN and FL. Their deployment can be computationally intensive, demanding much processing power and memory. Specifically, with large datasets like the CIC IoT Dataset 2023, training with GANs may increase operational costs and training times. Moreover, FL's communication cost can delay training, affecting real-time application capabilities, specifically in limited bandwidth environments.

Another limitation the study encountered was regarding data privacy. While FL improves data privacy by using local devices to store sensitive data, some risks remain. During the aggregation, the sensitive data from shared model parameters may be inferred by adversarial

users. Additionally, there is a risk of re-identification if the data is used to train GANs without proper anonymization. In this study, the efficacy of GANs and the quality and representativeness of the training data determine how well ML models perform. Using synthetic data in the CIC IoT Dataset 2023 may not accurately reflect real-world scenarios because of Biases or missing classes in the dataset, which could be reflected in the synthetic data.

Lastly, model interpretability is challenged due to the complexity of deploying GANs, restricting trust in model outputs. Understanding the alignment of generated synthetic data with real-world data can be challenging, particularly in applications where interpretability is critical.

Acknowledgment

We would like to thank the publisher for their support in publishing this research article and the university for providing the resources necessary to conduct our study. We also appreciate the editorial team's efforts in reviewing and editing our work.

Funding Information

This research received no external funding.

Author's Contributions

Shahad Alahmari: Conducted the core implementation, experimentation, and analysis of the proposed methodology.

Noura Aleisa: Provided overall supervision, technical guidance, and critical revisions throughout the research.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all authors have read and approved the manuscript, and there are no ethical issues involved.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Alamareen, A., Al-Mashagbeh, M. H., & Abuasal, S. (2023). Cyber Security & IoT Vulnerabilities Threats Intruders and Attacks Research Review. *Journal of Namibian Studies: History Politics Culture*, *33*(2), 330-345. https://doi.org/10.59670/jns.v33i.726

Almomani, O., Almaiah, M. A., MADI, M., Alsaaidah, A., Almomani, M. A., & Smadi, S. (2023). Reconnaissance attack detection via boosting machine learning classifiers. *AIP Conference Proceedings*, 060002. https://doi.org/10.1063/5.0174730

Almomani, O., AlSaaidah, A., Shareha, A., Alzaqebah, A., & Almomani, M. (2024). Performance evaluation of machine learning classifiers for predicting denial-ofservice attack in internet of things. *International Journal of Advanced Computer Science & Applications*, *15*(1).

Aqeel, M., Ali, F., Iqbal, M. W., Rana, T. A., Arif, M., & Auwul, Md. R. (2022). A Review of Security and Privacy Concerns in the Internet of Things (IoT). *Journal of Sensors*, *2022*, 1-20. https://doi.org/10.1155/2022/5724168

Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W., & Wahab, A. (2020). A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions. *Electronics*, *9*(7), 1177. https://doi.org/10.3390/electronics9071177

Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. *Electronics*, *12*(6), 1333. https://doi.org/10.3390/electronics12061333

Chen, W., Milosevic, Z., Rabhi, F. A., & Berry, A. (2023). Real-Time Analytics: Concepts, Architectures, and ML/AI Considerations. *IEEE Access*, *11*, 71634-71657. https://doi.org/10.1109/access.2023.3295694

Cloudflare (2024). DDoS Threat Report for 2023 Q2. *Cloudflare Blog*. https://blog.cloudflare.com/ddos-threat-report-2023-q2/

Cybersecurity Ventures (2023). 2023 Official Cybercrime Report. *eSentire*. https://esentire.com/resources/library/2023-official-cybercrime-report

Doriguzzi-Corin, R., & Siracusa, D. (2024). FLAD: Adaptive Federated Learning for DDoS attack detection. *Computers & Security*, *137*, 103597. https://doi.org/10.1016/j.cose.2023.103597

Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access*, *11*, 76071-76094. https://doi.org/10.1109/access.2023.3296707

Eshmawi, A. A., Umer, M., Ashraf, I., & Park, Y. (2024). Enhanced Machine Learning Ensemble Approach for Securing Small Unmanned Aerial Vehicles From GPS Spoofing Attacks. *IEEE Access*, *12*, 27344-27355. https://doi.org/10.1109/access.2024.3359700

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, *10*(1), 115. https://doi.org/10.1186/s40537-023-00792-7

Gelenbe, E., & Nakip, M. (2023). IoT Network Cybersecurity Assessment With the Associated Random Neural Network. *IEEE Access*, *11*, 85501-85512. https://doi.org/10.1109/access.2023.3297977

Google Developers (2022). Overview of GAN Structure. https://developers.google.com/machine-learning/gan/gan_structure

Habibi, O., Chemmakha, M., & Lazaar, M. (2023). Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. *Engineering Applications of Artificial Intelligence*, *118*, 105669. https://doi.org/10.1016/j.engappai.2022.105669

Haque, A., & Tasmin, S. (2020). Security threats and research challenges of iot-a review. *ArXiv:2101.03022*.

IBM (2024). Cost of a Data Breach. *IBM*. https://ibm.com/reports/data-breach

Statista (2022). Annual Number of IoT Malware Attacks Worldwide 2018 to 2022. *Statista*. https://statista.com/statistics/1377569/worldwide-annual-internet-of-things-attacks/

Statista (2023). IoT Connections Worldwide 2022-2033. *Statista*. https://statista.com/statistics/1183457/iot-connected-devices-worldwide/

Jiang, J. C., Kantarci, B., Oktug, S., & Soyata, T. (2020). Federated Learning in Smart City Sensing: Challenges and Opportunities. *Sensors*, *20*(21), 6230. https://doi.org/10.3390/s20216230

Jony, A. I., & Arnob, A. K. B. American International University-Bangladesh (AIUB), Dhaka, 1229, Bangladesh, & Arnob, A. K. B. (2024). Securing the Internet of Things: Evaluating Machine Learning Algorithms for Detecting IoT Cyberattacks Using CIC-IoT2023 Dataset. *International Journal of Information Technology and Computer Science*, *16*(4), 56-65. https://doi.org/10.5815/ijitcs.2024.04.04

Jony, A. I., & Arnob, A. K. B. (2024). A long short-term memory based approach for detecting cyber attacks in IoT using CIC-IoT2023 dataset. *Journal of Edge Computing*, *3*(1), 28-42. https://doi.org/10.55056/jec.648

Joshi, I., Grimmer, M., Rathgeb, C., Busch, C., Bremond, F., & Dantcheva, A. (2024). Synthetic Data in Human Analysis: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(7), 4957-4976. https://doi.org/10.1109/tpami.2024.3362821

Khan, M. A. (2021). HCRNNIDS: Hybrid Convolutional Recurrent Neural Network-Based Network Intrusion Detection System. *Processes*, *9*(5), 834. https://doi.org/10.3390/pr9050834

Mammen, P. (2021). Federated learning: Opportunities and challenges. *ArXiv:2101.05428*.

Mishra, S., Albarakati, A., & Sharma, S. K. (2022). Cyber threat intelligence for iot using machine learning. *Processes*, *10*(12), 2673. https://doi.org/10.3390/pr10122673

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors*, *23*(13), 5941. https://doi.org/10.3390/s23135941

Nookala Venu, D., Kumar, A., & Rao, M. (2022). Botnet attacks detection in internet of things using machine learning. *NeuroQuantology*, *20*(4), 743-754.

Otoom, A. F., Eleisah, W., & Abdallah, E. E. (2023). Deep Learning for Accurate Detection of Brute Force attacks on IoT Networks. *Procedia Computer Science*, *220*, 291-298. https://doi.org/10.1016/j.procs.2023.03.038

Parmar, P., & Sheth, R. (2022). Emerging Security Threats and Challenges in IoT. *Internet of Things and Cyber Physical Systems*, 51-69. https://doi.org/10.1201/9781003283003-3

Richa, E. (2021). IoT: Security Issues and Challenges. *Information and Communication Technology for Intelligent Systems*, 87-96.

Rust-Nguyen, N., Sharma, S., & Stamp, M. (2023). Darknet traffic classification and adversarial attacks using machine learning. *Computers & Security*, *127*, 103098. https://doi.org/10.1016/j.cose.2023.103098

Saba, T., Khan, A. R., Sadad, T., & Hong, S. (2022). Securing the IoT System of Smart City against Cyber Threats Using Deep Learning. *Discrete Dynamics in Nature and Society*, *2022*(1), 1241122. https://doi.org/10.1155/2022/1241122

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522-542. https://doi.org/10.1016/j.ins.2021.05.055

Sharma, A., Mansotra, V., & Singh, K. (2023). Detection of mirai botnet attacks on iot devices using deep learning. *Journal of Scientific Research and Technology*, 174-187.

Strelcenia, E., & Prakoonwit, S. (2023). Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. *AI*, *4*(1), 172-198. https://doi.org/10.3390/ai4010008

Tomislav, V., Hrvoje, L., Marija, D., Goran, M., & Robert, R. (2023). Data augmentation with gan to improve the prediction of at-risk students in a virtual learning environment. *International Conference on Artificial Intelligence in Education 2023*, 260-265.

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599-606.

Zelaya, C. V. G. (2019). Towards Explaining the Effects of Data Preprocessing on Machine Learning. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2086-2090. https://doi.org/10.1109/icde.2019.00245

Zhang, S., Xie, X., & Xu, Y. (2020). A brute-force black box method to attack machine learning-based systems in cybersecurity. *IEEE Access*, *8*, 128250-128263.