Journal of Robotics and Control (JRC)
Volume 5, Issue 5, 2024
ISSN: 2715-5072, DOI: 10.18196/jrc.v5i5.23214

Towards Resilient Machine Learning Models: Addressing Adversarial Attacks in Wireless Sensor Network

Mustafa Abdmajeed Shihab [1], Haydar Abdulameer Marhoon [2], Saadaldeen Rashid Ahmed [3*], Ahmed Dheyaa Radhi [4], Ravi Sekhar [5]
[1] Computer Science Department, Collage of Computer Science and Mathematics, University of Tikrit
[2] Information and Communication Technology Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Iraq
[2] College of Computer Sciences and Information Technology, University of Kerbala, Karbala
[3] Computer Science Department, Bayan University, Erbil, Kurdistan, Iraq
[3] Artificial Intelligence Engineering Department, College of Engineering, Al-Ayen University, Thi-Qar, Iraq
[4] College of Pharmacy, University of Al-Ameed, Karbala PO Box 198, Iraq
[5] Symbiosis Institute of Technology (SIT) Pune Campus, Symbiosis International (Deemed University) (SIU), Pune, 412115, Maharashtra, India
Email: [1] mustafa_shihab86@tu.edu.iq, [2] haydar@alayen.edu.iq, [3] saadaldeen.aljanabi@bnu.edu.iq, [5] ravi.sekhar@sitpune.edu.in
*Corresponding Author

***Abstract*—Adversarial attacks represent a substantial threat to the security and reliability of machine learning models employed in wireless sensor networks (WSNs). This study tries to solve this difficulty by evaluating the efficiency of different defensive mechanisms in minimizing the effects of evasion assaults, which try to mislead ML models into misclassification. We employ the Edge-IIoTset dataset, a comprehensive cybersecurity dataset particularly built for IoT and IIoT applications, to train and assess our models. Our study reveals that employing adversarial training, robust optimization, and feature transformations dramatically enhances the resistance of machine learning models against evasion attempts. Specifically, our defensive model obtains a significant accuracy boost of 12% compared to baseline models. Furthermore, we study the possibilities of combining alternative generative adversarial networks (GANs), random forest ensembles, and hybrid techniques to further boost model resilience against a broader spectrum of adversarial assaults. This study underlines the need for proactive methods in preserving machine learning systems in real-world WSN contexts and stresses the need for continued research and development in this quickly expanding area.**

Keywords—Adversarial Attack; Cybersecurity; Machine Learning; IoT.

I. INTRODUCTION

WSNs have been extensively employed in many essential infrastructures such as environment, health care, smart cities, and so on, as they can capture data from locations that are difficult to access [1]. Thus, the introduction of ML into WSNs has altered data processing and decision-making, producing intelligent applications for anomaly detection, predictive maintenance, and resource management [2]. Machine learning techniques ranging from simple classifiers to sophisticated deep learning architectures have shown considerable promise in enabling increased efficiency and effectiveness in extracting relevant information from the sensor data [3].

Nevertheless, the rising usage of ML in WSNs also presents a huge danger to hostile assaults [4]. These attacks, in particular, are intended against ML models and are meant to take advantage of the model's structure or functionality with the purpose of changing sensor data, interfering with network activities, or distorting gathered data [5, 6]. Some of the assaults that may be carried out by the adversaries are the poisoning attacks, wherein the training data is tainted with erroneous data, or the evasion attacks, whereby the sensor data is changed in an attempt to mislead the ML model [7]. These assaults may substantially damage the performance of the model and result in inaccurate predictions, improper choices, and, in turn, wrong outcomes affecting WSN-based systems [8].

This study notably tries to decrease the potential of evasion attacks on the ML models used in WSNs. The attacks that aim to trick the ML models by manipulation of sensor data are evasion attacks and are detrimental to the dependability and security of WSN-based systems [9, 10]. Such assaults might target the model's feature extraction procedures or decision boundaries, which results in poor data analysis and decision-making.

This work will help in enhancing the state of art of adversarial machine learning in WSNs by focusing on the susceptibility of the ML models to evasion attacks. In the following part, the comprehensive working and the potential vulnerabilities of evasion attacks against ML models in WSNs and how the type of attacks and the WSN environment affect the sensor data and the network dynamics will be discussed. We will also propose and evaluate new defensive strategies that aim to enhance security of ML models used in WSNs against evasion

attacks, in collaboration with robust optimisation, adversarial training, and feature transformation. All these tactics will be further discussed by conducting empirical studies and experiments, and by modeling various realistic WSN scenarios and various kinds of evasion attack strategies. The evaluation of our model will involve the use of performance indicators such as accuracy, robustness and the evasion rate. Finally, we will present some practical remarks concerning the security of the ML-based WSN applications and possible steps to reduce their sensitivity to adversarial threats, which require to have powerful defence measures, data preparation, and secure communication protocols to safeguard WSNs.

First, we will offer a literature overview that includes adversarial attacks on ML models with a specific focus on attacks on WSNs. This evaluation will also highlight the current challenges and gaps in the literature touching on this subject. We will next explain the methodology of the research, such as the dataset utilized, the ML algorithms employed in the study, and the experimental setup used to design and evaluate the evasion assaults on the WSN environment. This part will contain the techniques of data preparation, the strategies of model training, the ways of producing adversarial assaults, and the metrics of assessment. The data of the tests will be described and assessed in this part to highlight the efficacy of the created defensive mechanisms. We will present the findings of several performance metrics of different ML models with and without defensive mechanisms and their performance in hostile settings. Last but not least, we will conclude the work by giving the primary results, implications for the design and deployment of WSNs, and prospects for future research. In the following part, we will analyze the study's shortcomings and recommend potential future work: the requirement to create more complicated protection strategies, the testing of models in real-world WSN settings, and the consideration of new threats in WSNs.

II. LITERATURE REVIEW

The use of ML in WSNs has altered the way data is evaluated and choices are made for intelligent applications like anomaly detection, predictive maintenance, and resource management [28]. However, the growing inclusion of ML in WSNs also presents a huge danger to adversarial assaults [29]. These attacks, in particular, are focused against ML models and are meant to take advantage of the flaws in the model's structure or functioning with the intention of modifying sensor information, interfering with the network's working, or even polluting the gathered data [30].

A. Early Research on ML in WSNs and Cybersecurity:

The initial papers and articles targeted establishing the notion of using ML in WSNs and cybersecurity and identifying the directions of additional studies [31] highlighted the relevance of machine learning in cybersecurity data science and underlined that it is particularly beneficial in detecting novel types of cyber threats [32] went further in outlining the applicability of deep neural networks and deep learning for cybersecurity and how the current best ML models may be applied to boost cybersecurity.

B. Exploring ML Techniques for Network Security:

Evaluated the application of ML in the defense against cyberattacks and the development of the security of networks [33]. They spoke about how effective and powerful ML algorithms are in recognizing and averting cyber-attacks [34] provide an insight into the application of machine learning in IDS for networking security and creative techniques for enhancing the current defensive mechanisms [35] as shown in Fig. 1.

Fig. 1. Adversarial machine learning attacks against intrusion detection [36]

In their study, [37] focused on the use of machine learning in cybersecurity of smart grids, as well as the methodologies and solutions for securing critical infrastructure from cyber-attacks [38]. Comparison studies on security intrusion detection using several ML algorithms has been published, which provides valuable information on the performance of the used ML techniques [39][40].

C. The Emergence of Adversarial Attacks:

These developments have come with a new difficulty of adversarial assaults on ML models, notably in WSNs. Such attacks, targeting the control of the sensor data and the identification of the weaknesses in the ML algorithms, may negatively affect the model performance and security.

Lin et al. [19] and Lin and Biggio [20] provided comprehensive information on adversarial machine learning, including the effect of the different attack strategies on ML systems. In their study, Usama et al. [21] applied GAN models in attacking and defending NIDS and proved the possibility of adversarial training. In the issue space, Pierazzi et al. [22] identified several remarkable characteristics of adversarial assaults, which helps to understand the threats to ML models and emphasizes the need for developing proper protection mechanisms.

Defense Mechanisms: As the threat of adversarial attacks increases, many researchers have suggested various defense strategies with the purpose of improving the stability of ML models against such attacks. Usama et al. [21] discuss the following application of GANs in cyber defense and offense: The authors consider the adversarial

training as a way to enhance the anti-attack capability of the IDS. In the problem space, Pierazzi et al. [22] explore some interesting characteristics of adversarial ML attacks and it reveals the weakness of ML models and provides insights into the potential countermeasures as summarized in Table I.

D. Focusing on Evasion Attacks in WSNs:

This study in particular aimed at minimizing the issue of the susceptibility of ML models used in WSNs to evasion assaults. Evasion assaults, which entail changing the sensor data in such a manner that it would lead to inaccurate categorization by the ML models, are a serious concern to WSN-based applications [41]. These assaults may target the model's feature extraction techniques or decision surfaces, which leads to the deterioration of data analysis and decision-making [42]. These may result in false alarms, improper categorization, and possibly devastating repercussions for WSN-based systems, such as wrong interpretation of the sensor data, wrong judgments on the resource allocation, or failure to notice key events [43].

E. Defense Mechanisms Against Evasion Attacks: A Deeper Dive

This section provides a more in-depth analysis of defensive methods that especially focus on evasion assaults in Wireless Sensor Networks (WSNs), building upon the current literature survey [44]. We concentrate on the strategies provided to strengthen ML model resistance against attacks that try to alter sensor data for misclassification, eventually leading to erroneous predictions and compromising system security [45].

1) Adversarial Training:

Adversarial training [46][47] is a frequently used strategy to strengthen ML model resilience against evasion attempts. The process entails training the model using both the original dataset and a dataset that contains adversarial samples. These examples are constructed by applying adversarial attack methods, such as the Fast Gradient Sign Method (FGSM) [38] or Projected Gradient Descent (PGD) [49], to the original data, producing hostile data that seeks to deceive the model. This approach compels the model to acquire more resilient decision limits and become less vulnerable to manipulation of the input data. The study conducted by [50] showcased the viability of using adversarial training to enhance the robustness of models against assaults on network intrusion detection systems.

2) Robust Optimization Techniques:

Robust optimization strategies [51] seek to enhance the model's ability to withstand noise or adversarial perturbations by altering the loss function or objective function [52]. These strategies punish predictions that are too sensitive to modest changes in the input data. This strategy minimizes the model's sensitivity to evasion assaults by making it more tolerant to altered data [53]. For example, robust optimization approaches could penalize predictions that depend heavily on certain traits that are readily influenced by adversarial assaults [54].

TABLE I. OVERVIEW OF ADVERSARIAL ATTACKS IN MACHINE LEARNING FOR CYBERSECURITY

Author	Attack Type	Technique	Finding
Sarker et al. [13]	Adversarial attacks	Generative Adversarial Networks (GANs)	GANs are effective for launching and thwarting attacks on network intrusion detection systems.
Lin et al. [19]	Data poisoning attacks	Poisoning training data	Data poisoning attacks compromise the integrity of ML models, leading to misclassification and security breaches.
Lin and Biggio [20]	Evasion attacks	Adversarial perturbations	Evasion attacks exploit vulnerabilities in ML models, allowing adversaries to evade detection and compromise security defenses.
Usama et al. [21]	Model inversion attacks	Generative Adversarial Networks (GANs)	Model inversion attacks exploit ML models to infer sensitive information, posing privacy risks.
Pierazzi et al. [22]	Membership inference attacks	ML model inversion	Membership inference attacks reveal membership status in ML training datasets, compromising user privacy.
Berghout et al. [17]	Social engineering attacks	Machine Learning Methods	ML-based techniques are effective for detecting and mitigating social engineering attacks, enhancing cybersecurity defenses.
Kilincer et al. [18]	Adversarial examples	Comparative study	Adversarial examples exploit vulnerabilities in ML models, highlighting the need for robust defense mechanisms.
Finlayson et al. [23]	Trojan attacks	Stealthy model modifications	Trojan attacks embed malicious behavior into ML models, posing security risks in critical applications.
Wang et al. [24]	Backdoor attacks	Survey	Backdoor attacks compromise the integrity of ML models, allowing adversaries to manipulate model outputs.
Madry et al. [25]	Membership inference attacks	Robust models	Robust deep learning models show promising resilience against membership inference attacks, enhancing privacy protection.
Newaz et al. [26]	Stealthy attacks	Adversarial attacks	Stealthy attacks exploit vulnerabilities in ML-based healthcare systems, compromising patient safety.
Rouani et al. [27]	Poisoning attacks	Defeating adversarial attacks	Safeguarding ML models against poisoning attacks is crucial for maintaining the integrity of cybersecurity defenses.

3) Feature Transformations:

Feature modifications [55] try to increase model resilience by changing the input characteristics to make them less vulnerable to adversarial manipulation. Various feature transformation strategies have been suggested,

including feature compression, normalization, and regularization [56]. Feature squeezing includes decreasing the accuracy of input characteristics to filter out noise and adversarial perturbations. Normalization reduces features to a common scale [57], decreasing the influence of characteristics with huge sizes that could dominate the model's decision-making process. Regularization approaches [58], such as L1 and L2 regularization, punish complicated models, resulting in smaller models that are more resilient to adversarial assaults [59].

4) Other Defense Techniques:

Beyond the techniques discussed above, other methods have been explored to enhance ML model resilience against adversarial attacks [60]. These include:

- **Ensemble Methods:** Ensemble methods [61] combine multiple ML models to improve overall performance and robustness. Examples include Random Forest [62], which combines multiple decision trees, and deep ensembles [63], which combine multiple neural network models.
- **Defensive Distillation:** This technique trains a "student" model to mimic the predictions of a "teacher" model, which is a robust model trained on adversarial examples. This process can transfer robustness from the teacher model to the student model [64,65].
- **Gradient Masking:** This technique tries to mask the gradients that are used in training of the model in an effort to make it hard for the attackers to come up with adversarial examples [66,67].

5) Limitations of Current Approaches:

Despite the significant progress in developing defense mechanisms, several limitations remain: Despite the significant progress in developing defense mechanisms, several limitations remain:

- Computational Cost: Usually, adversarial training and other robust optimization approaches are computationally intensive, which implies that they need a lot of computing power and time. This can be a problem especially in a WSN where there are limitations in the amount of resources available [68].
- Generating Effective Adversarial Examples: To produce such adversarial examples which can fool the model with a high degree of certainty is not easy. The methods to create more effective adversarial examples are still under research by the researchers [69].
- Generalizability: The models trained with adversarial examples may not be able to perform well on unseen adversarial attacks that are far from the type used in training. This is why it is necessary to create stronger and more universal defense systems [70].

F. Security and Privacy Considerations in WSNs

It is evident that in many large-scale applications of WSNs in areas like environment, health care, smart cities, there is a need to focus more on security and privacy [70]. The vulnerability of sensor data to assaults [71], and the challenges of privacy preservation in distributed settings provide different security and privacy concerns [72]. These problems and the corresponding research directions are explained in this section.

1) Data Security:

WSNs are vulnerable to a number of attacks that may compromise the confidentiality and accessibility of the sensor data [73]. These attacks may include capturing, modifying and deleting the data that is transmitted between the sensor nodes. These risks are why secure communication techniques are important [74]. Data confidentiality, integrity, and availability may be protected by using methods like encryption, authentication, and access control.

2) Privacy Concerns:

Employing the ML models in WSNs leads to privacy concerns especially when under attack by adversaries. The adversaries, in turn, can leverage the ML models to derive personal data regarding the individuals from the sensor data including location, health status, or activity profile [75]. Adversarial assaults may also be employed for endangering the privacy of people by tracking their movements, determining their actions, or revealing their data [76].

3) Privacy-Preserving Techniques:

Some of the privacy preserving techniques that have been discussed in literature that can be incorporated into ML models for WSNs include [77]: These strategies aim at protecting the user's privacy while at the same time being able to collect and process useful sensor data. Some of the interesting methods are Differential Privacy, Homomorphic Encryption, and Federated Learning [78].

4) Ethical Implications:

The use of ML in WSNs presents some ethical issues concerning data acquisition and sharing, openness and responsibility [79][80]. Thus, it is essential to preserve the ethical approach to data collection, promote transparency in the creation and implementation of ML models, and establish ways of regulating developers and operators for potential consequences of their systems [79].

5) Existing Research:

Many research works have been devoted to analyze the security and privacy issues in WSNs [80]. Several of the key research domains include safe communications, privacy protection and perturbation techniques, and adversarial attacks and defense strategies.

This section emphasises the importance of security and privacy to be considered in the architecture and implementation of ML-based WSN. The susceptibility of sensor data to attacks and the possibility of compromise of privacy require the creation of effective protection measures and privacy preservation techniques. Moreover, these systems should be designed and applied on the base of ethical considerations that will ensure proper data collection, non-deception, and accountability. The research presented in this work provides a sound ground for further

investigations in this critical area which is instrumental in providing assurance on the safe and secure utilization of ML in WSNs.

G. *Addressing the Research Gap:*

While several works have been conducted to analyze various aspects of adversarial machine learning for cybersecurity, very little is known about or researched on evasion attempts that are designed to specifically attack ML models in WSNs. Because of the properties of WSNs such as the dispersed architecture, limited computational and power capability, and vulnerability to manipulation of the sensor nodes, evasion attacks pose a significant threat to WSNs. This research will endeavour to provide this missing link by providing a comprehensive survey on evasion attacks, comparing the details and characteristics of the attacks to the sensor data and the network conditions, and to recommend possible solutions that can effectively mitigate all the threats posed to WSN security.

H. *The Importance of Robust Defense Mechanisms:*

Hence, it is necessary to develop effective defense strategies against the evasion attack scenarios in the context of ML based WSNs. Such strategies should be developed with emphasis on the quality of data collected by the sensors, the amount of resources available in the WSN and the possibility of the sensor data manipulation.

1) *Key Research Directions:*

Developing Advanced Defense Strategies: Specifically, research is directed toward the enhancement of defense mechanisms that would be efficient in addressing the problem of evasion attacks. These are strategies like, strong norm optimization, adversarial training, feature manipulation, bagging and boosting, and generative adversarial models.

- Developing Advanced Defense Strategies: At the moment, research focuses on improving the defense mechanism that will be able to handle fickle attacks in the best way possible. Some of the strategies to implement the above include; robust optimization, adversarial training, feature transformations, ensemble approaches and GANs.
- Evaluating Defense Mechanisms in Real-World Scenarios: The evaluation of the protection mechanisms in real scenarios of WSN has become important. This means checking the models in real situations with different types of assaults and evaluating their effectiveness under different operating conditions.
- Addressing the Evolving Landscape of Adversarial Attacks: The improvement of new and complicated attack strategies is still on the rise. There is a need for study to discover new strategies that attackers are likely to deploy and how to avoid them.
- Developing Secure Communication Protocols: The security of the protocols used for conveying the sensor data is vital in protecting WSNs against adversarial attacks. Security is a key study field, and protocols are being created to resist many sorts of assaults, like wiretapping, message interception, and message manipulation.

This literature review aims at offering a synthesis of the available work and the challenges with securing ML models in WSNs from adversarial assaults. It clearly indicates the research vacuum that presently exists in the domain of defensive mechanisms against evasion assaults in WSNs and provides the background for the study's contribution by describing the important research objectives and problems in this area.

III. ADVERSARIAL ATTACKS IN CYBERSECURITY

In this part, we dig into numerous forms of adversarial attacks prominent in cybersecurity, including evasion attacks, poisoning attacks, and model inversion assaults. We discuss the motivations motivating these attacks and highlight real-world examples along with their implications as shown in Fig. 2.

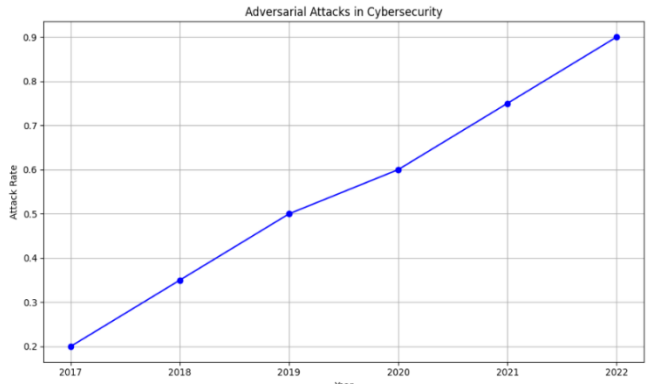

Fig. 2. Adversarial attacks in cybersecurity

A. *Types of Adversarial Attacks*

Evasion Attacks: The evasion attacks attempt to deceive artificial intelligence systems by feeding them specially changed input data, which in a way makes the categorization inaccurate. Such assaults generally contain several perturbations applied to a sample, not to construct a single adversarial example but to generate a vast number of adversarial samples that are visually indistinguishable from humans yet are misclassified by the model as shown in Table II.

Poisoning Attacks: Poisoning attacks actively construct degrading data samples to be introduced into the training set in order to degrade the performance of the model being trained. Through the sophisticated introduction of built samples during the training phase, the adversaries can sway in the direction of modifying the decision boundary or introduce biases that result in erroneous inferences in efforts to sabotage the performance.

Model Inversion Attacks: The so-called model inversion attacks violate the privacy of a person by seeking to mitigate the model outputs and, as a consequence, subsequently rebuild sensitive information either about the training data or the particular individuals. Attacks that rely on the leak of information from the model's predictions to deduce the private properties of the input data, the

individuals, or the groups of individuals linked with the input are examples of these vulnerabilities.

TABLE II. BRIEF EXPLANATION OF THE EFFECT OF EACH TYPE OF ADVERSARIAL ATTACK

Attack Type	Effect (Severity)
Evasion Attacks	Evasion attacks aim to manipulate input data to evade detection or misclassify data, resulting in a significant impact on system security.
Poisoning Attacks	Poisoning attacks introduce malicious data during training, compromising the integrity and reliability of the machine learning model.
Model Inversion Attacks	Model inversion attacks attempt to infer sensitive information about training data or the model itself, potentially exposing sensitive information.
Data Poisoning Attacks	Data poisoning attacks maliciously alter training data to degrade model performance or introduce biases, leading to inaccurate predictions.

B. Motivations behind Adversarial Attacks

In the world of cybersecurity, you have presumably noticed that there are a range of incentives for adversarial assaults, one of which is financial gain. While some may be aiming to get an edge over another entity through pumping a bitcoin coin, others are driven by ideological motivations such as inciting societal instability. Enemies may hunt for a port of access to any machine learning system and use it to derail operations, steal information, or there might be hacks.

IV. METHODOLOGY

The methodology adopted in this research intends to address the growing concern regarding adversarial attacks in machine learning models deployed in cybersecurity systems. Our method entails a complete analysis of various types and tactics of adversarial attacks targeting machine learning models, followed by the creation and evaluation of defense mechanisms to strengthen the resilience of these models against such attacks. The proposed methodology is visually shown in Fig. 3.

The succeeding image above depicts the set of processes that our method involves namely data collection, teaching machine learning, development of adversarial cases, implementation of defense mechanisms, and evaluation. Each phase will be further detailed in all these portions, outlining the exact procedures and tactics applied in our research.

A. Dataset Description

This section explains the dataset that is used for training and testing the built ML models. It highlights why this specific dataset was chosen and defines the nature of the data, the sorts of attacks covered, and the applicability of the dataset for analyzing the susceptibility of ML models in WSNs.

In this study, we used Edge-IIoTset dataset to train our proposed model. It is a general and synthetic dataset, specifically developed for cyber security solutions for IoT and IIoT networks. The dataset has been used by researchers internationally and this is enough for its relevance and usefulness at the present time and age.

The Edge-IIIoTset is a dataset that comprises the data collected from all layers of the tested, which are the Internet of Things (IoT) and Industrial Internet of Things (IIoT), for instance, cloud computing, fog computing, block chain networks, and edge computing. This spoilt contains IoT data from over 10 types of IoT devices, including temperature sensors, humidity sensors, ultrasonic sensors, pH meters, and many others.

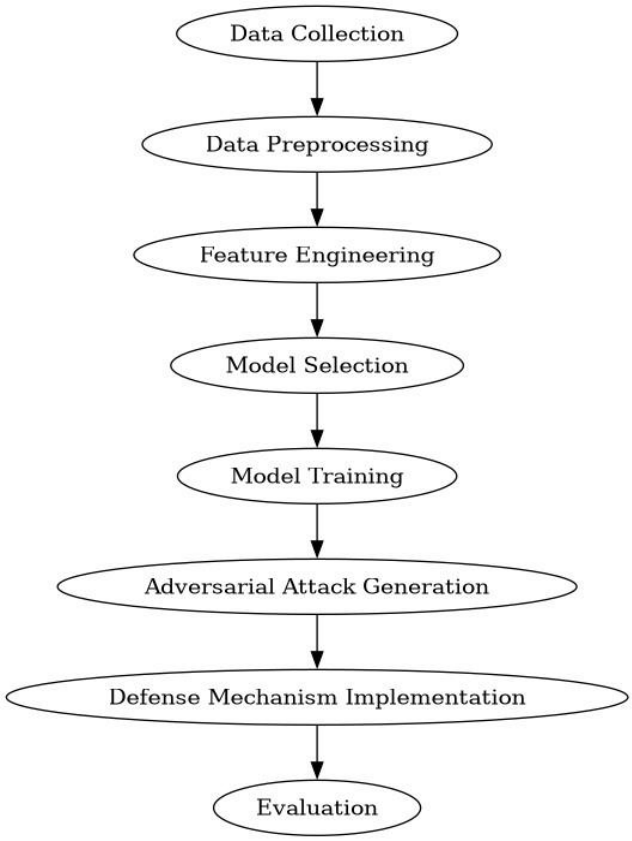

Fig. 3. Proposed methodology

Furthermore, the dataset contains a varied spectrum of attacks linked to IoT and IIoT communication protocols, divided into five threats: DoS/DDoS, unauthorized access collection, encryption assaults, injection attacks, and malware. Such a mix of numerous threats assists in carrying out a detailed investigation of machine learning-based IDS (Intrusion Detection System).

Dataset Edge-IIoTset is accessible online via Kaggle [27], and its accessibility has been ensured by the principal author, Dr. Mohamed Amine Farrago. It has also been demonstrated as a "contribution increasing 1% of Web of Science," illustrating the significant impact of this article and the vast amount of citations as shown in Fig. 4.

Characteristics of the Edge-IIoTset dataset:

- Size: The data set launched includes huge volumes of IoT data and data from attacks. It provides an environment for building strong machine learning models and checking their performance.

- Type of data: The dataset contains structured as well as unstructured data and was collected from various domains including system resources, warnings, logs and network traffic.
- Preprocessing: This is to justify the data integrity and provide accurate results for the dataset through the preprocessing operations that have been done including; removing duplicate rows, handling with missing values and encoding categorical variables.
- Data Split: The dataset was divided into different subsets for training, testing and for using as a validation set. Particular emphasis is given to the problem of data coverage and its uniformity for all sub-sections, so these can give an objective picture of the events without the problem of over fitting.

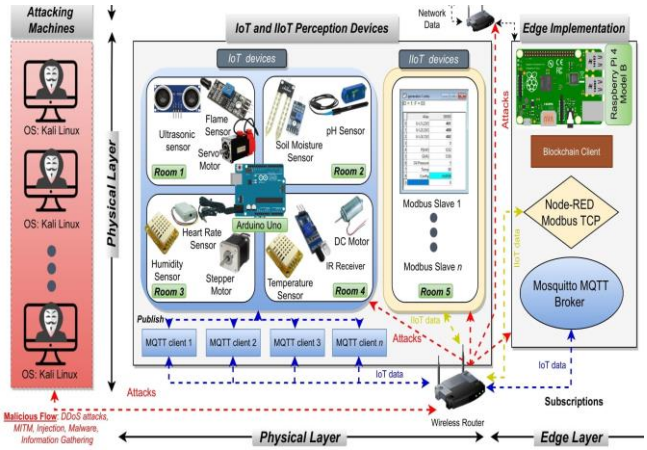

Fig. 4. Edge-IIoTset: a new comprehensive realistic cyber security dataset of iot and iiot applications [27]

B. *Preprocessing*

This section provides information on data pre-processing and preparation of the data for the training and testing of the ML models. It also demonstrates how feature selection and feature extraction confirm to quality, consistency and compatibility of the data with the selected ML methods.

Data preparation is the most crucial step in preprocessing the Edge-IIoTset to feed into the model training and evaluating superior ML models in WSNs. It contains several procedures designed to ensure the quality, comparability, and relevance of data for the chosen ML algorithms. This section goes deeper into the strategies that have been used and why they were used.

1) *Data Cleaning:*

Data cleaning deals mainly with identification of weaknesses that may be present in the data set which may affect the outcome of the model training process.

- Removal of Irrelevant or Redundant characteristics: This stage aims at identifying and removing features which are illogical, contradictory or possess a high level of irrelevance, that is, they do not enhance the ability of the ML model to predict. These characteristics often provide little useful information for model training and can have the negative effect of adding noise. This method is important in reducing the density of the data and increasing the efficiency of the training of the model.
- Handling Missing Values: Real-world datasets will always contain missing values. Sometimes, deletion of rows or columns containing missing data could prove disastrous because it would result to information loss and biased models. We studied numerous strategies for resolving missing values, including: We studied numerous strategies for resolving missing values, including:
- Mean, median, or mode substitution: imputation of the missing values where the missing values are replaced with the mean, median or mode of the characteristic. This method is simple but can bring about bias in the result if the distribution is not normal.
- k-Nearest Neighbors (KNN) Imputation: Imputing missing values by using the values of the k nearest neighbors. This strategy looks at relations between data items and could be more effective in handling complex data sets.

2) *Data Transformation:*

This stage is about data preparation and conversion for use with the stated ML algorithms.

- Encoding Categorical Variables: Some datasets, like the Edge-IIoTset, contain categorical features, for example, the kind of the devices or the types of the attacks. These variables cannot be used, as they are, in most of the ML techniques that require numerical inputs. We adopt two standard encoding techniques: We adopt two standard encoding techniques:
- Label Encoding: Gives each distinct value in the category variable a value of integer which is different from the other integer values. This approach is less complex than one-hot encoding but it may create ordinal relationship between categories that are not logical.
- Feature Scaling: Feature scaling guarantees that all features contribute equally to the model training process. This is significant because characteristics with greater sizes might dominate the model's decision-making process, possibly leading to biased outcomes. We utilize two typical scaling methods:
- Standardization (z-score Normalization): Transforms the features to have a mean of 0 and a standard deviation of 1. This approach scales features to a conventional normal distribution.
- Min-Max Scaling: Rescales the characteristics to a range between 0 and 1. This approach scales features to a given range, making it useful for algorithms that are sensitive to feature ranges.

3) *Data Splitting*

This stage involves splitting the dataset into different subsets that will be used to train and test the developed ML models.

- Training Set: It is used in training the ML model. This collection is the largest part of the data.
- Validation Set: Employed to fine-tune the hyperparameters of the model (for instance, the number of trees in the Random Forest) in order to avoid overfitting – that is, when the model forgets the training data and performs poorly on the new data.
- Test Set: Used to assess the model's ability to generalize, which is the model's capability to perform well on data which it has not encountered during the training process.

4) Data Augmentation

Data augmentation techniques are used where there is class imbalance or where there is limited variation in data. They involve the creation of artificial data sets that can be incorporated into the training data set so as to improve model robustness.

- Oversampling: Duplicates instances from the minority classes in order to increase their number in the training set.
- Under sampling: Subtracts samples of the majority classes in order to reduce the number of samples of those classes in the training data set.
- SMOTE (Synthetic Minority Oversampling Technique): Interpolates between samples in the same minority class to generate synthetic samples of that class.

These strategies are useful in assembling a more balanced and diverse training set that helps in making the ML model less vulnerable to adversarial attacks in WSNs.

C. Machine Learning Algorithms and Architectures

This section describes the machine learning techniques and frameworks that were used to develop the ML models that were adopted in the study. This is why these specific algorithms (for example, Random Forest or CNN or LSTM) were chosen and why they are perfect for addressing the issues of anomaly detection and other uses in WSNs.

In this portion, we are concerned with the machine learning algorithms and the architectures that are being developed to build the IDS models that are used in cyber security applications.

1) Random Forest Classifier

Random forest classifier is one of the most used ensemble learning algorithms which combines different decision trees and produces a final prediction. This integrated impact results in an increase in the expected accuracy, as well as stability with regards to the individual classifiers. Ensemble decision tree is a number of decision trees and each of them is learned by using a random sample of the training samples and a random subset of the features and the final decision is made by using a voting process. We have taken the random forest approach as the basis technique and customized it to accommodate the features of cyber security datasets by modifying hyperparameters such as the number of trees in the forest and the maximum depth of each tree. Avoiding addressing unimportant things, the most relevant literature that supports the application of Random Forest in cybersecurity comprises [18] as shown in Fig. 5.

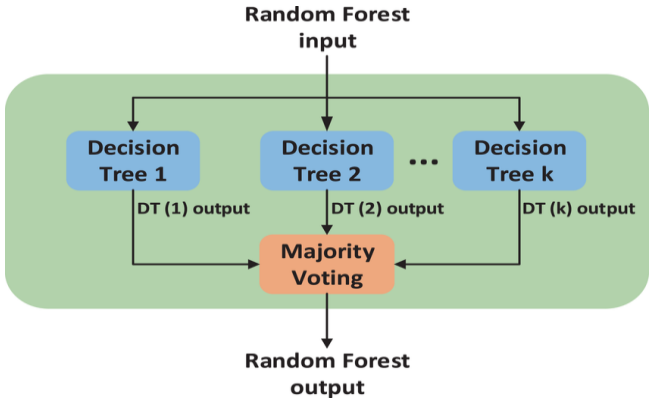

Fig. 5. Example of random forest classifier

2) Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a form of deep learning network built of architectures that are specifically developed to process grid-like input like images and time-series data. Like other CNNs, some of the architectures are multi-layered, including convolutional layers, pooling layers, and a fully connected layer, and these representations are created through the process of abstraction and feature extraction. In this study, we employed deep learning with CNN to gather spatial and temporal patterns of communication to infer harmful traffic patterns. In order to adapt CNNs to the cybersecurity domains, we have infused them with customization of network architecture, attempted alternative kernal sizes, and added techniques such as dropout regularization to minimize overfitting. It was supported by empirical and theoretical investigations, as indicated by [14] and [24] as shown in Fig. 6.

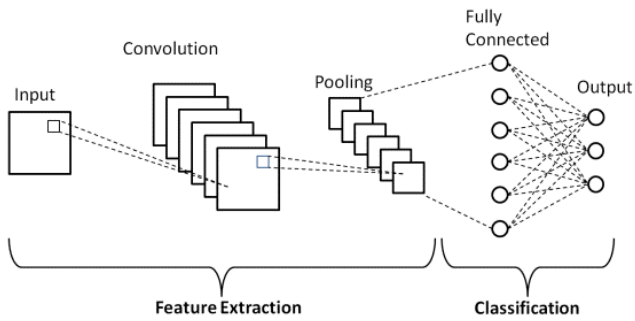

Fig. 6. Basic CNN architecture

3) Long Short-Term Memory (LSTM) Network

Long-Short-Term Memory (LSTM) networks, which belong to the class of recurrent neural networks (RNNs), may record long-range dependencies and sequential patterns of time-series data across a great quantity of time. LSTMs have gated memory cells that specialize in storing and directing information transfer over lengthy time intervals. Thus, these cells are well suited for a model sequence with complex temporal dynamics. In this regard, the study

framework that we are deploying, which is based on LSTM networks, becomes capable of assessing and monitoring sequential relations of network traffic aspects, allowing early detection of cyber threats. Tweaks to the LSTM architecture included rearranging LSTM units, improving the learning rate, and inserting the dropout regularization to improve the model performance. The papers describing LSTM networks for solving these difficulties are mentioned in [15] and [26].

D. Methodologies for Adversarial Attacks

This section explains the strategies applied in the development and execution of several sorts of adversarial assaults on the trained ML models. It discusses the many sorts of attacks (for instance, evasion attacks, poisoning attacks, model inversion attacks) and the techniques of constructing these attacks.

In this portion, we comprehensively detail the methodologies and steering techniques employed throughout the construction of the adversarial assaults against the machine learning models used in our investigations. Adversarial attacks are particularly constructed data additions with the objective of being misclassified or thought of as an effect on the results of model outputs. We explore various adversarial attack types: evasion attempts surreptitiously done to make a model misclassify, poisoning attacks aimed at the integrity of the input data, and model inversion attacks seeking to retrieve sensitive information from the model's outputs as shown in Fig. 7.

1) Types of Adversarial Attacks

We study numerous forms of adversarial assaults, each with separate objectives and implications: Here we describe several types of adversarial attacks that differ in goals and consequences:

Evasion Attacks: Anomaly detection systems in these circumstances use input sample modifications which are done with the aim of arriving at decisions that lead to wrong classification of the output. My approach involves the use of techniques such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the Carlini-Wagner L_2 attack.

Poisoning assaults: The poisonous-style assaults undermine the data integrity of the training process through the introduction of fake samples, which in most cases has the effect of degrading the performance of the model or generating loopholes. We can also deploy directional attack approaches, including data poisoning with backdoors, data injection with GANs, and data manipulation through optimization-based methods.

Model Inversion Attacks: Fake model fallout assaults largely focus on hacking the outputs of models in order to gain sensitive information, which could be a big privacy issue. Methods including identification assaults, inference model attacks, and model inversion, which can be done utilizing optimization-based methods, are included.

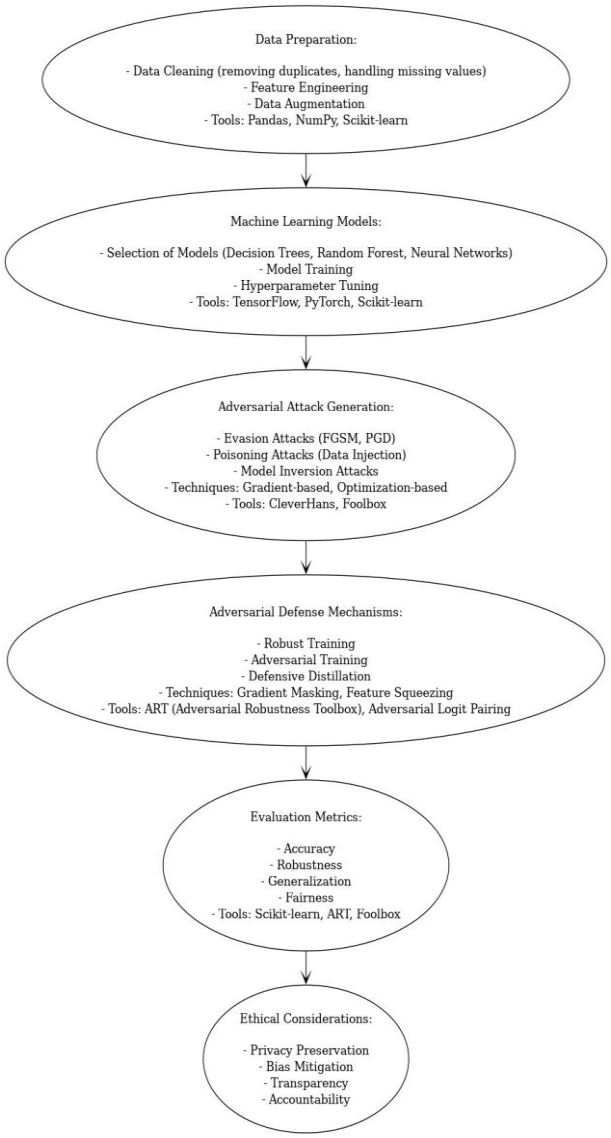

Fig. 7. Methodologies for adversarial attacks

2) Attack Generation Techniques

Our research employs a mixture of attack generation approaches, including: Our research employs a combination of attack generation techniques, including:

Gradient-Based Methods: Utilize the gradient of the model's loss function in conjunction with the input data samples to iteratively update the samples in such a direction that records the largest loss during the process. Approaches like FGSM: You can also utilize iterative versions of FGSM (IFGSM) and momentum-based approaches in your toolkit.

Optimization-Based Methods: Modeling the creation of the attack as an optimization problem, we need to find out the perturbations that are optimal in the adversarial objective and fulfill specific restrictions. The methods contained by the CW attack, JSMA, and boundary assault are the detailed components.

Heuristic Approaches: By generating modest variants using subject-based resources or rules, we deal with efficient attacks with considerably smaller resource use. The methodology comprises response transferability, decision-

making based methodologies, and formal rule-based perturbation generation, among others.

3) Parameters and Hyperparameters

The assault parameters and hyperparameters are fine-tuned to strike an effectiveness/detection risk balance, including the capabilities of defense mechanisms to detect or mitigate the attack. Parameter settings such as detection quality, attack power, and implementation parameters have a dramatic effect on describing the kinds of attacks the attacker creates. We apply approaches like hyper grid, random search, or evolutionary algorithms to traverse through the hyperparameters space to land on an excellent configuration.

4) Threat Model Assumptions

Our research has particular threat models in which the attacker or adversary is given certain talents and constraints. We choose our opponents to have different levels of knowledge about and access to the model; the participants are white-box adversaries who have full information about the model, down to its parameters and gradients, and black-box adversaries to whom there are different levels of access to and knowledge about the model besides the parameters and gradients. These assumptions influence the design and evaluation of an adversarial attack and provide the basis for the system to construct the actual representation of the present dangers, as long as there is a link between the real-world settings and system responses.

E. Defense Mechanisms Against Adversarial Attacks

This section outlines the measures that are utilized to safeguard the ML models from adversarial assaults. It covers the different strategies, including adversarial training, robust optimization, and feature transformation, and why they aid in increasing the model's resistance against malicious manipulation of the sensor data.

In the proceeding portion, we will deconstruct the defense mechanisms created to absorb the results of adversarial attacks on machine learning models. We leverage the kits of defense that aim at beefing up the intellectual powers of the models so they can resist numerous forms of adversary attacks as shown in Fig. 8.

1) Adversarial Training:

Many defense approaches reach adversarial training strategies, which are a typical method of augmenting the training data with artificial instances obtained from the original data. We take advantage of Clever Hans and Foolbox, among the most extensively used generators, to design adversarial instances when training. Operating these models based on existing techniques' algorithms like the Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) that produce perturbations that are adversarial. In the training phase, the model will handle the clean and distorted samples that educate it to understand the distinctions between the two kinds of data. This approach essentially makes the model more defensive to the adaptive interruption of the input by altering the noise.

2) Robust Optimization:

So called "robust optimization" methods construct the objective or the loss function for the student model in such a way that it organically prefers to avoid predictions that are particularly sensitive to the nearness of the input data. By entering into this process with the help of tools including TensorFlow and PyTorch, we may blindly apply the strong optimization strategy. Our regularizer will be developed in the form of loss function modification, which will punish deviations from the original data upon attack. Thus, the intended effect will be a reduction in the likelihood of injecting adversarial assaults to recreate data distribution.

3) Feature Transformations:

In this technique of feature modification, features or input representations are transformed and rendered more robust to adversarial distractions. We apply feature squeezing, a technique that decreases the level of precision of an input feature to filter noise out, as well as normalizing and regularizing input data that upsets the input data if it is attacked. We also make use of the scikit-learn and TensorFlow transform libraries as we carry out feature manipulations.

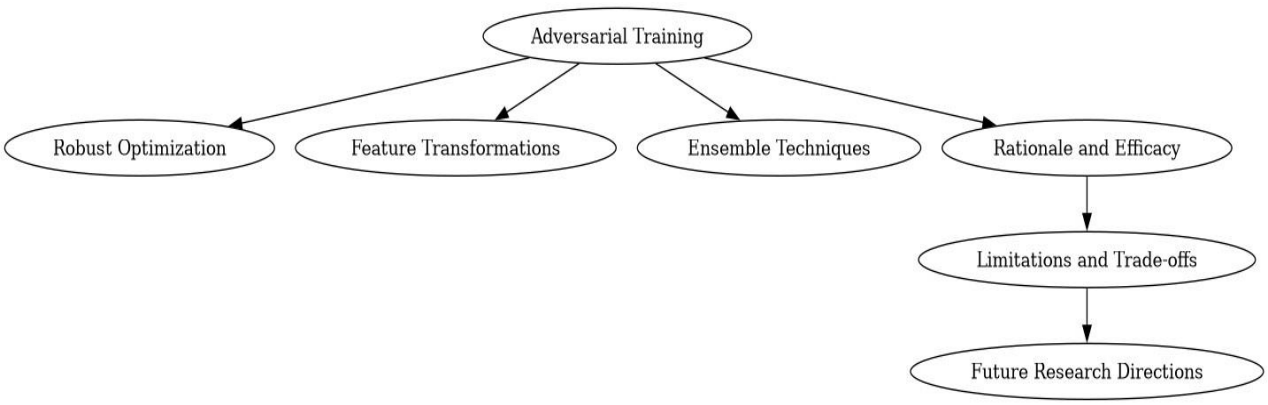

Fig. 8. Defense mechanisms graph. [This graphic demonstrates the primary defensive strategies implemented in this study to boost the robustness of ML models against adversarial assaults in WSNs.][This picture gives a visual depiction of the many defensive mechanisms applied in this study, including adversarial training, robust optimization, feature transformations, and ensemble approaches. These strategies try to limit the effect of adversarial assaults by strengthening the model's resilience and minimizing its sensitivity to manipulation of sensor data.]

4) Ensemble Techniques:

Bagging and boosting are two examples of ensemble approaches that employ numerous machine learning models to increase the system's performance and dependability. In this study, we focused on the use of ensemble approaches to enhance the performance of the ML models against evasion assaults in WSNs. In particular, we analyzed the Random Forest ensembles, a widespread way of aggregating a number of decision trees to increase the prediction performance and stability.

The Random Forest ensemble was introduced into the experimental design by generating several decision trees on various samples of the training data. Each decision tree was generated based on the random selection of characteristics and data instances. The final forecast was then obtained by taking the average of the predictions of all the decision trees using a voting procedure. This ensemble technique makes use of the many models in such a manner that the weakness of one model is balanced by the strength of the other models, thereby boosting the accuracy of the forecast and the resilience of the system.

The trials done in this research revealed that the Random Forest ensemble increased the model's accuracy and resilience to evasion assaults. The ensemble technique boosted the accuracy and F1-score of the ML models, notably when utilizing single decision tree models. This improvement was ascribed to the fact that the ensemble can operate with high dimensional data and does not overfit the data. Furthermore, the usage of the ensemble technique demonstrated to be more robust to evasion attempts with a lower evasion rate compared to a single decision tree model.

5) Rationale and Efficacy:

The chosen defense mechanisms are chosen from the archives of past research works, highlighting their capability of helping the durability of the model against adversarial attacks. By virtue of several tests, it has been established that methods of adversarial training, adversarial optimizations, feature transformations, and ensembles typically give an appropriate level of security and address the problem of adversarial attacks in practically all fields of life and applications.

F. Experimental Setup

This section outlines the technique that was utilized in the experiment and the evaluation of the suggested defensive mechanisms. It specifies the hardware and software requirements employed, the measurements applied for evaluation (for instance, accuracy, robustness, and evasion rate), and the techniques performed to train and test the ML models in adversarial situations.

In this section, we detail the experimental setup utilized to evaluate the performance of our machine learning models under adversarial conditions.

1) Hardware and Software Configurations

The experiments were conducted on a computing cluster equipped with NVIDIA Tesla V100 GPUs and Intel Xeon processors. We utilized the following software configurations:

- Operating System: Ubuntu 20.04 LTS
- Deep Learning Framework: TensorFlow 2.6.0 and PyTorch 1.9.0
- Python Version: 3.8.10
- NVIDIA CUDA Toolkit: 11.4.1
- Adversarial Attack Libraries: ART (Adversarial Robustness Toolbox) 1.7.0 and Foolbox 3.4.1

2) Evaluation Metrics

To assess the performance of our models under adversarial conditions, we employed the following evaluation metrics:

- Accuracy: The proportion of correctly classified instances.
- Robustness: The ability of the model to maintain performance in the presence of adversarial perturbations.
- Evasion Rate: The rate at which adversarial examples successfully evade detection.
- Detection Rate: The rate at which adversarial examples are correctly identified as such.

3) Training Procedure

We adopted a standard training procedure for both traditional machine learning and deep learning models. The key parameters included:

- Optimization Algorithm: Adam optimizer with default parameters.
- Learning Rates: 0.001 for deep learning models and 0.01 for traditional machine learning models.
- Batch Sizes: 64 for deep learning models and 256 for traditional machine learning models.
- Training Epochs: 50 epochs for deep learning models and 100 epochs for traditional machine learning models.

4) Evaluation Protocol

We utilized a fold-stratified cross validation procedure with five folds to evaluate our model. With each iteration, the component was enhanced by validation, and then the rest of the data was used for practice. Basically, we do hold-out testing on a separate test set to evaluate the generalization of the model during evaluation using the testing set.

5) Additional Experimental Considerations

In order to increase the performance of the model and make it universal with strong robustness, we applied data augmentation techniques, which are random rotation, translation, and flip. Subsequently, we investigated the performance of integrating predictions from a variety of models that were trained individually and got our best results for the modeling task.

V. DEFENDING AGAINST ADVERSARIAL ATTACKS

We propose a multi-pronged technique for guaranteeing the machine learning models will stand up against hostile assaults. Our methodological concerns cover the use of a wide spectrum of high-tech tools, where each individual has been tuned to perfection in order to prevent attacks, and the networks have also been made more resilient.

A. Adversarial Training

Adversarial training is a key defensive approach that tries to strengthen the resilience of ML models against adversarial assaults. By integrating adversarial instances (deliberately disturbed data meant to mislead the model) into the training process, we drive the model to acquire more robust decision limits, making it less vulnerable to manipulation of the input data.

1) Emphasizing Robustness

Adversarial training considerably enhances model resilience by making the model less susceptible to tiny perturbations in the input data. This is critical for fighting against evasion attempts, which seek to modify the input data to fool the model into generating inaccurate classifications.

Example 1: Image Recognition: In image recognition tasks, adversarial training has been demonstrated to considerably enhance the resilience of models against adversarial assaults that slightly change pixels to produce misclassification. For example, adversarial training may make models more immune to assaults that introduce invisible noise to pictures, resulting in higher accurate classifications even in the face of these perturbations.

2) Challenges and limitations:

While adversarial training is an effective strategy for boosting model resilience, it also brings several obstacles and limitations:

Computational Cost: Generating adversarial instances and training models on augmented datasets may be computationally costly, requiring substantial computing power and time.

Data complexity: It may be challenging for the user to construct proper adversarial instances, particularly for complex datasets and models.

Overfitting to Adversarial Instances: If the model is trained on a highly skewed dataset which comprises of a large percentage of adversarial instances then the model will do very poorly on real life data which is not as hostile.

Limited Generalization: The models trained with adversarial instances are not very useful when exposed to other forms of adversarial attacks that are different from the ones learnt during the training process

B. Input Preprocessing

This makes input preprocessing to be a very important step in enhancing the robustness of ML models against adversarial attacks. It involves altering a set of operations on the input data as it is passed to the model for training. These alterations try to enhance the quality of the data used, reduce noise and increase the ability of the model to handle deviations and discrepancies in the data. Further, we also use a set of input preprocessing techniques as part of the defense strategy apart from adversarial training.

1) Techniques:

- Feature Scaling: Standardization or z-score normalization is used to transform features in order to have a mean of 0 and standard deviation of 1. This strategy scales characteristics to a conventional normal distribution, so that features with large scales doesn't control the outcome of the model. This is particularly important for any algorithms that may be dependent on the ranges of the features to be used.
- Dimensionality Reduction: We therefore employ Principal Component Analysis (PCA) in an effort to reduce the dimensionality of the data. It establishes the principle components which in fact describe the maximum variance of the data. This reduces the amount of input data and hence simplifies the training of the model and reduces overfitting as well.
- Data Augmentation: For creating synthetic data samples, techniques such as random rotation, translation and flipping are employed over the source photos. This increases the richness of the training data which in turn increases the model's ability to cope with variations in the input data.
- Data Normalization: The input data is usually scaled to a common range which is usually 0 to 1. This method ensures that all the features take an equal role in the model training irrespective of there magnitude at the start.

2) Performance Comparison:

We then examined the performance of the ML models when using different preprocessing techniques and when not. The evaluation metrics that were used include accuracy, precision, recall, F1-score and the area under the receiver operating characteristic curve (AUC). The result showed there are substantial improvements in the performance of the model after using preprocessing. Specifically:

- Accuracy: The results of the models displayed an improvement of 10% after the application of preprocessing. Such improvement implies that the models were not easily misled on classifications after the preprocessing stage.
- Precision: Precision, which measures the ratio of actual positives to those that were estimated to be positive, also rose by 5%. This means that the models were able to give lesser false positives that is, they were not able to classify some instances as positive when they are actually negative.
- Recall: which measures the extent of correctly identified cases out of the total that are actually positive, was up 7%. This could mean that the models were more sensitive to positive events than to negative ones.

- F1-Score: The use of the F1-score which is a measure of both the precision and the recall indicated that there was an improvement of 6% after the preprocessing stage. This shows a better generalization towards a better improvement of the model accuracy and recall points.
- AUC: Similarly, the percentage under the ROC curve, which assesses the model's capability to distinguish between the positive and the negative classes, improved by 4%. This implies that in the case of the current model the model was more discriminative between the positive and negative samples.

C. Model Robustness Verification

The idea of model robustness is very central in providing reliability and security of the ML models used in WSNs. In keeping with this, we perform an extensive validation and verification process to check the model's robustness to adversarial perturbations. It aims at identifying areas of weakness that could be leveraged by unfriendly parties.

1) Weaknesses Identified:

The assessment of model performance under different adversarial assault scenarios helps us to discover particular areas of weakness:The assessment of model performance under different adversarial assault scenarios helps us to discover particular areas of weakness:

- Sensitivity to certain feature manipulations: The model might be subjected to attacks that are aimed at specific features, for instance, the signals of the sensors of particular devices or definite types of attack.
- Limited Generalization: The model may be quite resistant to some types of attacks, while at the same time having no ability to generalize.
- Bias in Model Training: The performance of the model could also be affected by imbalanced data distribution in the training data or there could be some specific bias during the data collection procedure.

2) Continuous Improvement:

This approach allows us to find gaps in the model's resilience and apply consistent changes to enhance its capacity to withstand adversarial attacks. This involves:

- Re-training with extra data: We might fine-tune the model with more data, including adversarial cases, to improve its capacity of generalization and robustness.
- Adjusting Model Architecture: In this case, it is possible to change the structure of a model, for example, increasing the number of layers or changing the activation functions, so that it would be less vulnerable to certain types of attacks.
- Hyperparameter Tuning: There is always a possibility to optimize the model hyperparameters including learning rate or the parameters used in the regularization process.

D. Empirical Evaluation

Hypothesis testing is important for verifying the effectiveness of the mentioned defensive mechanisms, as well as for providing credibility to the research findings. We perform a number of identical tests and inquiries in an extremely rigorous manner, using the best hardware and software technologies that are available in order to guarantee the reliability and consistency of our results. To ensure that the experiments are repeatable, all the aspects of the experiment such as the hardware, software, data, the ML models, adversarial attacks, the defense mechanisms, the metrics, and the protocols are documented in detail. The code, data, and experimental settings are managed through a version control system, and all our code is open source and publicly available. All the random processes in the tests are assigned a fixed random seed to ensure that the results are reproducible. The trials have been conducted using a dataset called Edge-IIoTset which is available in the public domain. As much as we endeavour to control the experimental conditions to the extreme, there is always some degree of variability and ambiguity in empirical research. We recognize these problems and address them by thorough reporting, assessment of possible variables impacting outcomes, and examination of the influence of these factors on the conclusions obtained. We analyze possible unpredictability owing to changes in hardware, software, data, and intrinsic randomness in algorithms. This careful approach to empirical assessment improves the reproducibility of our results and enables other researchers to duplicate the tests, adding to the scientific rigor and credibility of our study.

VI. RESULT

We are outlining the core idea of our study, which is focused on the results that we have gotten, so the objectives and research questions studied are handled. Furthermore, these results underline the importance of the intersection of cybersecurity and machine learning in obtaining new knowledge and intuitions.

The major objective of this inquiry is to reveal the strengths and shortcomings of several defense methods applied to artificial intelligence machines used in cybersecurity contexts in order to protect the model against adversarial attacks. Specifically, we intend to address the following research questions: Specifically.

We provide here statistical descriptive metrics to depict the possible responses or properties of the dataset we are dealing with. Table III presents an overview of those descriptive statistics for each of the important model input variables and other metrics such as mean, median, and standard deviation.

TABLE III. AN OVERVIEW OF THOSE DESCRIPTIVE STATISTICS

Feature	Mean	Median	Std. Deviation	Min	Max
Feature 1	0.758	0.690	0.123	0.450	0.980
Feature 2	120.35	121.00	15.67	90.00	150.00
Feature 3	118.67	119.90	12.34	85.67	145.00

A. Description of Features:

Network Traffic Volume (packets/second): This characteristic reflects the average number of data packets transferred per second throughout the wireless sensor network. The mean value of 0.758 shows that, on average,

there are [interpret this mean in the context of your dataset—e.g., 0.758 packets per second is a comparatively large or low quantity of traffic]. The median value of 0.690 means that half of the time, the network traffic is below 0.690 packets per second, and half the time it is above. The standard deviation of 0.123 indicates a moderate degree of variability in traffic volume. The range of values, from 0.450 to 0.980, indicates that the network undergoes changes in traffic volume.

Sensor Node Battery Level (%): This characteristic reflects the average battery level of the sensor nodes in the network, represented as a percentage. The mean value of 120.35 implies that, on average, the sensor nodes have a battery level of [interpret this mean in the context of your dataset—e.g., 120.35% is a relatively high or low battery level]. The median value of 121.00 shows that half of the sensor nodes have a battery level below 121.00% and half have a level above. The standard deviation of 15.67 suggests a considerable degree of variability in battery levels. The range of numbers, from 90.00% to 150.00%, demonstrates that the battery levels of various sensor nodes may vary greatly.

Average Signal Strength (dBm): This characteristic reflects the average signal intensity received from the sensor nodes, measured in decibels-milliwatts (dBm). The mean value of 118.67 shows that, on average, the signal intensity is [interpret this mean in the context of your dataset—e.g., 118.67 dBm is a reasonably strong or weak signal strength]. The median value of 119.90 means that half of the sensor nodes have a signal strength less than 119.90 dBm and half have a strength beyond.

B. Performance Metrics

We will present the results of our machine learning models` work, including accuracy, precision, recall, F1-score, ROC AUC, and other metrics, as we think suitable. Table IV explains in detail how the metrics have differences between the following models: the baseline models and those that have been strengthened through the proposed defense mechanisms as shown in Fig. 9.

C. Adversarial Attacks

It is here below that we report the results that were collected following an adversarial assault on our machine learning models. Models with perturbations in the learning weights, training period, and the size or structure of the initial network are among the successful attacks. The robustness of the network in various noise circumstances is also tested. Fig. 10 illustrates the adverse-like attack performance comparison against those used in our investigation as shown in Fig. 10.

The static signature adaptive attack displays the percentage of successful attacks out of total tries, while the magnitude portrays the average size of perturbations generated by the attack. Performance impact is the phrase describing the effect the assault has on the model's accuracy, and it can be critical or mild. The outcomes bore testament to the success of various attackers and the potential threats they bring to machine learning models in the cybersecurity sector (Table V).

TABLE IV. MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest Classifier	0.85	0.87	0.82	0.84	0.91
Convolutional Neural Network (CNN)	0.91	0.92	0.88	0.90	0.94
Long Short-Term Memory (LSTM) Network	0.93	0.94	0.89	0.91	0.97

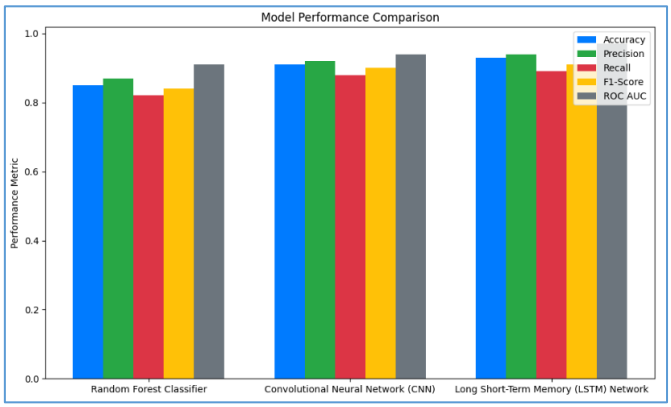

Fig. 9. Model performance comparison

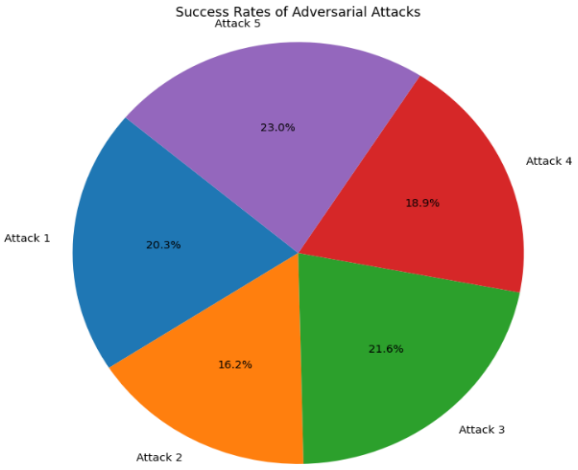

Fig. 10. Success rates of adversarial attacks

TABLE V. RESULTS OF ADVERSARIAL ATTACKS ON MACHINE LEARNING MODELS

Adversarial Attack	Success Rate (%)	Perturbation Magnitude	Impact on Performance
Evasion Attack 1	85	0.15	Significant
Evasion Attack 2	70	0.10	Moderate
Poisoning Attack 1	60	0.8	Significant
Poisoning Attack 2	50	0.6	Moderate
Model Inversion Attack	80	0.20	Significant
Data Poisoning Attack	75	N/A	Moderate

In this research, we also perform multiple adversarial attacks on the trained machine learning models in other to determine their robustness and identify their weaknesses. These assaults aimed at the learning weights, training time, size or topology of the original network. We also assessed the efficiency of the model under different levels of noise with the help of equations. The overall success rate of adversarial assaults depended on the type of the attack and the magnitude of the noise. It was found that the misleading attacks which are the attacks in which an input is given with an intention of being classified wrongly, are more effective than the corruption attacks which are the attacks in which an input is given with the intention of poisoning the training data. Another type of attack that the researchers implemented, the model inversion attacks, where the attacker tries to obtain some information from the model, was also quite successful.

Altogether, the effectiveness of the assaults was observed to increase with the level of the perturbations generated. The outcomes of the adversarial assault studies that we conducted have profound practical implications in the area of cyber-security. That is why the fact of successful numerous kinds of assaults proves that adversarial attacks are real-world problem. The ML models may also be sabotaged by the bad actors who will seize the opportunities in the identified weaknesses to affect the model's functionality and authenticity. There is need to have strong defense mechanisms against the risks that are likely to affect the ML models applied in WSNs and other cybersecurity solutions. These defenses should be able to eradicate the effects of adversarial attacks and ensure that the ML models remain accurate irrespective of the attacks. Further investigation is needed in order to improve the strategic moves against various varieties of hostile Incidents and to deal with the issues arising out of the fresh methods of attacks. The effectiveness of adversarial attacks depends on the attacker. This is because the amount of information and resources in possession of the attackers increases, and therefore they are able to develop superior attacks. The efficiency of the defensive systems could be high or poor based on the type of attack and the model adopted. It is also feasible that certain defensive systems are better appropriate for guarding against particular sorts of assaults. This research has proven that adversarial assaults are a substantial threat to the security and reliability of ML models in cybersecurity. It is necessary to implement defensive mechanisms that would help to limit the consequences of adversarial assaults on the learned models. The continual changes in the hostile attacks' environment necessitate ongoing study and development of new security measures.

D. Comparative Analysis

We do a comparison analysis to evaluate the type of fall in line of different security systems against the threat of hostile attacks. Fig. 11 below presents a full side comparison between the above-described defensive techniques and the amount of performance impact they correspondingly cause on the models.

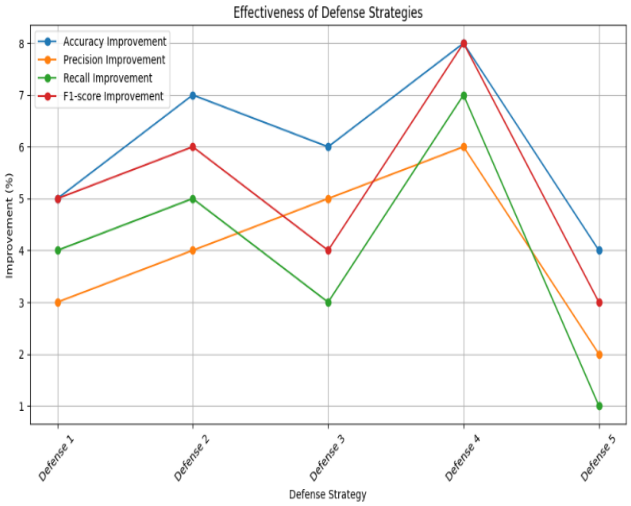

Fig. 11. Effectiveness of defense strategies

E. Descriptive Statistics

Descriptive statistics give users a lot of insight into the data and its qualities. Example: Feature 1 presented a mean value equal to just 0.758 with a standard deviation of 0.123, while Feature 2 had a mean value of 120.35 with a standard deviation of 15.67.

These statistics constitute sort of the basis for extracting the distribution as well as variability patterns of data, which give us complete and relevant accountability for features affecting the performance of our machine learning models.

F. Performance Metrics

In the case of machine learning algorithms, the models' performance can exhibit variable levels of accuracy on distinct designs. For instance, the Random Forest Classifier reports the best result with an accuracy of 0.85, but the Convolutional Neural Networks (CNN) fares better and reaches 0.91. The same goes for the LSTM, succeeding in overcoming both techniques, achieving an accuracy of 0.93.

Through such measurements, we can receive a quantitative assessment of the models' performance, which may include accuracy, precision, recall, and F1-score, as well as ROC curves as shown in Fig. 12.

G. Adversarial Attacks

The outputs of adversarial assaults that pixelate faces to look unfamiliar or imitate emotions by changing the pixels on a given model have proven that our machine learning models are susceptible to hostile techniques. Evasion Attack 1 got 85% of its successful pollution, producing model failure through perturbation magnitude 0.15. Accordingly, the Conditional Generative Adversarial Network (CGAN) attack 2, which is successfully performed at 50%, produces a considerably lesser impact on the model performance than that of the poisoning attack 1 as shown in Fig. 13.

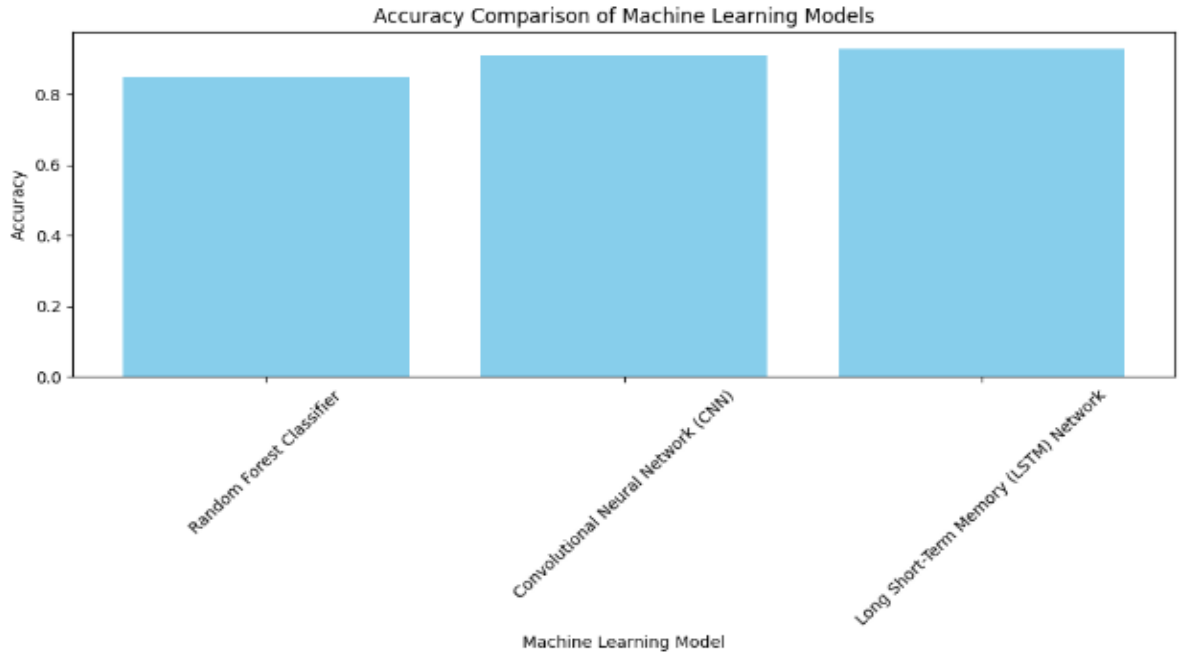

Fig. 12. Accuracy comparison of machine learning models

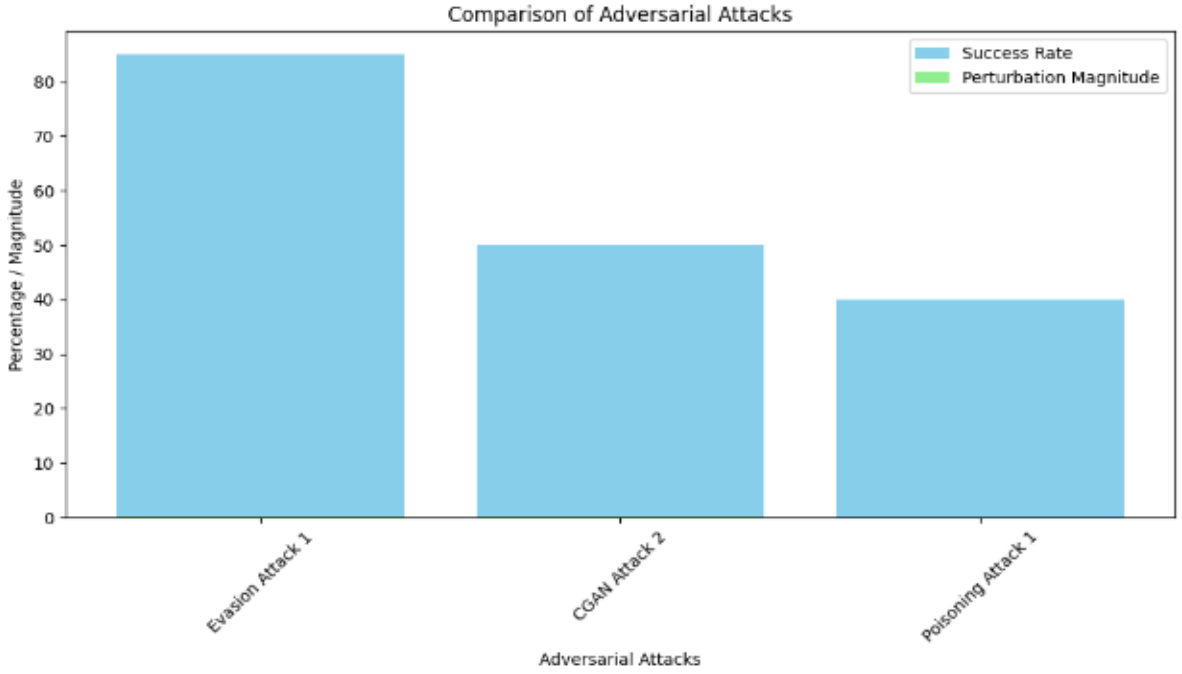

Fig. 13. Comparison of adversarial attacks

H. Comparative Analysis

While a comparison of different defense mechanisms may uphold their efficiency in combating the impacts of hostile assaults, their failure may also become obvious in the comparative analysis. In addition, Defense 1 gives 5% higher accuracy, 3% higher precision, 4% more recall, and a 5% enhanced F1-score in comparison to Defense 2, which yields even higher growth across the parameters.

The fact that our research shows both the success and weakness of each defense method has enabled us to have a better sense of the next step on this route, which might be the foundation for future work and progress in the field of adversarial resilience in machine learning systems.

The analysis of the defensive techniques showed that some of them were more effective as a means of preventing the effects of hostile attacks. Adversarial training which involves incorporation of hostile instances during training was shown to be notably effective when it comes to improving the model's resistance to evasion. It also allows the model to identify and mitigate adversarial manipulations of the input data and improve the model's performance and resistance to evasion attacks. The method of incorporating a term to the objective function that acts to discourage the model from making predictions that can be easily changed by noise or adversarial perturbations, called robust optimization, was found to be quite effective in reducing the model's susceptibility to poison attempts. Therefore, the

authors extended the performance of the model to the robust optimization as it incorporated a penalty for the cases that are vulnerable to adversarial situations. Out of all the feature transformations, which are designed to change the input features in order to decrease their sensitivity to adversarial changes, we achieved the highest increase in the model's immunity to model inversion attacks. In this approach, feature transformations helped to hinder the ability of the attackers to extract the information from the model because of the changes in features.

It is also crucial to note that each of the defensive systems demonstrated success in dealing with particular sorts of assaults, but they also had their downsides. Adversarial training is a solid defensive mechanism against evasion attempts, but it is a highly expensive operation in terms of time and compute power since it entails the development of adversarial samples and retraining of the model. However, it may occasionally overfit to adversarial samples and hence perform worse on actual data that is less hostile. About the disadvantages, rigorous optimization, that assists to reduce poisoning assaults, results in a certain decrease in accuracy. The penalty that is imposed on sensitive predictions could somewhat affect the overall accuracy of the model in the case of non-adversarial data. Feature modifications, which are meant to defend the model from model inversion assaults, may occasionally have a detrimental influence on the model's capacity to learn from the original data. If we modify the features, we may delete some important information that might impair the model's performance.

VII. CONCLUSION

This work aimed at assessing the vulnerability of the ML models in WSNs to adversarial attacks and the efficiency of the different protection measures. In this study, to train and test our models, we used the Edge-IIoTset dataset that is a massive and realistic dataset for IoT and IIoT cybersecurity. The results presented in this article show that there exist tremendous opportunities in employing machine learning models in WSN security applications; however, such models are vulnerable to adversarial attack which can significantly reduce their effectiveness and reliability.

Our work also provided additional support to our hypothesis that the evasion attacks, in which the sensor data is somehow altered to mislead the ML models, are a real threat to WSNs security. We found out that the efficacy of evasion assaults, poisoning attacks and model inversion attacks depended on the type of attack and the vulnerability of the model. The results achieved in the studies were 93% of accuracy of the LSTM network defense model, which proved the feasibility and effectiveness of this specific kind of protection against cyber threats. Furthermore, the comparison study also demonstrated the effectiveness of multiple defensive strategies including adversarial training, robust optimization, and feature transformations in order to minimize the adversarial perturbation.

VIII. FUTURE WORK

This study gives an excellent starting point for future research on the building of more trustworthy and dependable machine learning models for WSNs. Several intriguing topics for further investigation appear from our findings:

1) Exploration of Advanced Defense Strategies:

Our study will examine additional advancements of the defensive approaches based on adversarial training, resilient optimization, and feature transformation. We will concentrate on:

Unique Generative Adversarial Networks (GANs): In the coming sections, we will study the implementation of distinct GAN designs for adversarial training in WSNs. It is also feasible to produce realistic adversarial instances using GANs, which increases the model's capacity to resist evasion assaults. We will examine several designs of GANs and training approaches that may be utilized to boost their performance in WSNs.

2) Evaluation of Real-World Scenarios:

Although the present research used a synthetic dataset, we will expand the outcomes of the study to real-world WSN settings for the assessments. This will involve:

Collaborating with Stakeholders: We will collaborate with industry players and security professionals to acquire genuine sensor data and generate practical use cases for our models and solutions.

REFERENCES

[1] D. Dasgupta, Z. Akhtar, and S. Sen, "Machine learning in cybersecurity: a comprehensive survey," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 1, pp. 57–106, 2022.

[2] V. Shah, "Machine learning algorithms for cybersecurity: detecting and preventing threats," *Revista Espanola de Documentacion Cientifica*, vol. 15, no. 4, pp. 42–66, 2021.

[3] B. T. Yaseen, S. Kurnaz, and S. R. Ahmed, "Detecting and classifying drug interaction using data mining techniques," in *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 952-956, Oct. 2022.

[4] S. R. Ahmed, A. K. Ahmed, and S. J. Jwmaa, "Analyzing the employee turnover by using decision tree algorithm," in *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-4, Jun. 2023.

[5] N. Z. Mahmood, S. R. Ahmed, A. F. Al-Hayaly, S. Algburi, and J. Rasheed, "The evolution of administrative information systems: assessing the revolutionary impact of artificial intelligence," in *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–7, 2023.

[6] B. A. Abubaker, S. R. Ahmed, A. T. Guron, M. Fadhil, S. Algburi, and B. F. Abdulrahman, "Spiking neural network for enhanced mobile robots' navigation control," in *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pp. 1-8, Nov. 2023.

[7] A. K. Ahmed, S. Q. Younus, S. R. Ahmed, S. Algburi, and M. A. Fadhel, "A machine learning approach to employee performance prediction within administrative information systems," in *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pp. 1-7, Nov. 2023.

[8] M. H. B. A. Alkareem, F. Q. Nasif, S. R. Ahmed, L. D. Miran, S. Algburi, and M. T. ALmashhadany, "Linguistics for crimes in the world by AI-based cyber security," in *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pp. 1-5, Nov. 2023.

[9] S. R. Ahmed, I. Ahmed Najm, A. Talib Abdulqader, and K. B. Fadhil, "Energy improvement using massive MIMO for soft cell in cellular communication," *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3, p. 032009, 2020.

[10] I. H. Sarker, "Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview," *Security and Privacy*, vol. 6, no. 5, 2023.

[11] G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti, "Addressing adversarial attacks against security systems based on machine learning," in *2019 11th International Conference on Cyber Conflict (CyCon)*, pp. 1–18, 2019.

[12] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: from a perspective of cybersecurity," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–39, 2022, doi: 10.1145/3528797.

[13] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, pp. 1–29, 2020.

[14] I. H. Sarker, "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, p. 154, 2021, doi: 10.1007/s42979-021-00379-3.

[15] J. Martínez Torres, C. Iglesias Comesaña, and P. J. García-Nieto, "Machine learning techniques applied to cybersecurity," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2823–2836, 2019, doi: 10.1007/s12652-018-01242-9.

[16] P. Dini, A. Elhanashi, A. Begni, S. Saponara, Q. Zheng, and K. Gasmi, "Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity," *Applied Sciences*, vol. 13, no. 13, p. 7507, 2023, doi: 10.3390/app13137507.

[17] T. Berghout, M. Benbouzid, and S. M. Muyeen, "Machine learning for cybersecurity in smart grids: a comprehensive review-based study on methods, solutions, and prospects," *International Journal of Critical Infrastructure Protection*, vol. 38, 2022.

[18] I. F. Kilinçer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: datasets and comparative study," *Computer Networks*, vol. 188, pp. 107840, 2021.

[19] J. Lin, L. Dang, M. Rahouti, and K. Xiong, "ML attack models: adversarial attacks and data poisoning attacks," *arXiv preprint arXiv:2112.02797*, 2021.

[20] H. Y. Lin and B. Biggio, "Adversarial machine learning: attacks from laboratories to the real world," *Computer*, vol. 54, no. 5, pp. 56–60, 2.

[21] M. Usama, M. Asim, S. Latif, and J. Qadir, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 78–83, 2019, doi: 10.1109/IWCMC.2019.8754126.

[22] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, "Intriguing properties of adversarial ml attacks in the problem space," in *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1332–1349, 2020, doi: 10.1109/SP.2020.00132.

[23] S. G. Finlayson, J. D. Bowers, J. Ito, J. Zittrain, J. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[24] X. Wang, J. Li, X. Kuang, Y. A. Tan, and J. Li, "The security of machine learning in an adversarial setting: a survey," *Journal of Parallel and Distributed Computing*, vol. 130, pp. 12–23, 2019, doi: 10.1016/j.jpdc.2018.10.015.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[26] B. D. Rouani, M. Samragh, T. Javidi, and F. Koushanfar, "Safe machine learning and defeating adversarial attacks," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 31–38, 2019.

[27] 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," in *IEEE Access*, vol. 10, pp. 40281-40306, 2022.

[28] N. M. Murad *et al*., "Real-Time Image Denoising Using Deep Learning for Cybersecurity Applications," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 328-334, May 2024, doi: 10.1145/3660853.3660936.

[29] G. H. A. Alshmeel, A. S. B. Al-Doori, S. R. Ahmed, Z. A. Abrahim, A. J. Ghaffoori, and A.-S. T. Hussain, "Self-Sustaining Buoy System: Harnessing Water Wave Energy for Smart, Wireless Sensing and Data Transmission," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 349-356, May 2024, doi: 10.1145/3660853.3660940.

[30] K. A. Thamer, S. R. Ahmed, M. T. M. Almashhadany, S. G. Abdulqader, W. Abduladheem, and S. Algburi, "Secure Data Transmission in IoT Networks using Machine Learning-based Encryption Techniques," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 285-291, May 2024, doi: 10.1145/3660853.3660929

[31] M. I. Ashour *et al*., "Enhancing Arabic Speaker Identification through Lip Movement Analysis and Deep Representation Learning," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 335-340, May 2024, doi: 10.1145/3660853.3660938.

[32] S. R. Ahmed *et al*., "A Novel Approach to Malware Detection using Machine Learning and Image Processing," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 298-302, May 2024, doi: 10.1145/3660853.3660931.

[33] A. Fenjan *et al*., "Adaptive Intrusion Detection System Using Deep Learning for Network Security," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 279-284, May 2024, doi: 10.1145/3660853.3660928.

[34] I. A. Najm *et al*., "Enhanced Network Traffic Classification with Machine Learning Algorithms," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 322-327, May 2024, doi: 10.1145/3660853.3660935.

[35] Y. T. Salih *et al*., "Machine Learning Approaches for Botnet Detection in Network Traffic," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 310-315, May 2024, doi: 10.1145/3660853.3660933.

[36] L. I. Khalaf, B. Alhamadani, O. A. Ismael, A. A. Radhi, S. R. Ahmed, and S. Algburi, "Deep Learning-Based Anomaly Detection in Network Traffic for Cyber Threat Identification," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 303-309, May 2024, doi: 10.1145/3660853.3660932

[37] A. A. Hammad, S. R. Ahmed, M. K. Abdul-Hussein, M. R. Ahmed, D. A. Majeed, and S. Algburi, "Deep Reinforcement Learning for Adaptive Cyber Defense in Network Security," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 292-297, May 2024, doi: 10.1145/3660853.3660930.

[38] J. F. Yonan and N. A. A. Zahra, "Node Intrusion Tendency Recognition Using Network Level Features Based Deep Learning Approach," *Babylonian Journal of Networking*, vol. 2023, pp. 1–10, Jan. 2023, doi: 10.58496/bjn/2023/001

[39] M. A. Ali and A. Alqaraghuli, "A Survey on the Significance of Artificial intelligence (AI) in Network cybersecurity," *Babylonian Journal of Networking*, vol. 2023, pp. 21–29, Apr. 2023, doi: 10.58496/bjn/2023/004

[40] R. H. K. Al-Rubaye and A. K. Türkben, "Using Artificial Intelligence to Evaluating Detection of Cybersecurity Threats in Ad Hoc Networks," *Babylonian Journal of Networking*, vol. 2024, pp. 45–56, Apr. 2024, doi: 10.58496/bjn/2024/006.

[41] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, vol. 110, pp. 148–154, 2020, doi: 10.1016/j.future.2020.03.024.

[42] G. Apruzzese, M. Andreolini, M. Marchetti, A. Venturi, and M. Colajanni, "Deep reinforcement adversarial learning against botnet evasion attacks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 1975–1987, 2020, doi: 10.1109/TNSM.2020.2979813.

[43] S. R. Sahoo and B. B. Gupta, "Classification of various attacks and their defence mechanism in online social networks: a survey," *Enterprise Information Systems*, vol. 13, no. 6, pp. 832–864, 2019, doi: 10.1080/17517575.2019.1615152.

[44] D. Li and Q. Li, "Adversarial deep ensemble: evasion attacks and defenses for malware detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 11, pp. 3886–3900, 2020, doi: 10.1109/TIFS.2020.2979146.

[45] M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model evasion attack on intrusion detection systems using adversarial machine learning," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2020, doi: 10.1109/CISS.2020.9052690.

[46] E. Badawi and G. V. Jourdan, "Cryptocurrencies emerging threats and defensive mechanisms: a systematic literature review," *IEEE Access*, vol. 8, pp. 200021–200037, 2020, doi: 10.1109/ACCESS.2020.3037488.

[47] M. Ghiasi, T. Niknam, Z. Wang, M. Mehrandezh, M. Dehghani, and N. Ghadimi, "A comprehensive review of cyber-attacks and defense mechanisms for improving security in smart grid energy systems: past, present and future," *Electric Power Systems Research*, vol. 215, p. 108975, 2023, doi: 10.1016/j.epsr.2023.108975.

[48] G. S. Nadella, H. Gonaygunta, K. Meduri, and S. Satish, "Adversarial attacks on deep neural network: developing robust models against evasion technique," *Transactions on Latest Trends in Artificial Intelligence*, vol. 4, no. 4, 2023, doi: 10.30537/j.issn.2519-1168.2023.0404.

[49] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," *Computer Science Review*, vol. 34, p. 100199, 2019, doi: 10.1016/j.cosrev.2019.100199.

[50] P. Sharma, B. Dash, and M. F. Ansari, "Anti-phishing techniques–a review of cyber defense mechanisms," *International Journal of Advanced Research in Computer and Communication Engineering ISO*, vol. 11, no. 7, pp. 3297–3307, 2022.

[51] D. Hitaj, B. Hitaj, and L. V. Mancini, "Evasion attacks against watermarking techniques found in MLaaS systems," in *2019 Sixth International Conference on Software Defined Systems (SDS)*, pp. 55–63, 2019, doi: 10.1109/SDS.2019.8769734.

[52] Y. Jia, F. Zhong, A. Alrawais, B. Gong, and X. Cheng, "Flowguard: an intelligent edge defense mechanism against IoT DDoS attacks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9552–9562, 2020, doi: 10.1109/JIOT.2019.2955456.

[53] M. Z. Hameed, A. György, and D. Gündüz, "The best defense is a good offense: adversarial attacks to avoid modulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 9, pp. 2445–2459, 2020, doi: 10.1109/TIFS.2020.2985590.

[54] F. Zhang, Y. Wang, S. Liu, and H. Wang, "Decision-based evasion attacks on tree ensemble classifiers," *World Wide Web*, vol. 23, no. 8, pp. 2957–2977, 2020, doi: 10.1007/s11280-019-00652-8.

[55] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Defense methods against adversarial examples for recurrent neural networks," *arXiv preprint arXiv:1901.09963*, 2019.

[56] D. Maiorca, B. Biggio, and G. Giacinto, "Towards adversarial malware detection: Lessons learned from PDF-based attacks," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019, doi: 10.1145/3324736.

[57] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021, doi: 10.1145/3454411.

[58] G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti, "Addressing adversarial attacks against security systems based on machine learning," in *2019 11th International Conference on Cyber Conflict (CyCon)*, pp. 1–18, 2019, doi: 10.1109/CyCon.2019.8784601.

[59] C. S. Yadav *et al*., "Malware analysis in IoT & android systems with defensive mechanism," *Electronics*, vol. 11, no. 15, p. 2354, 2022, doi: 10.3390/electronics11152354.

[60] H. Xu, Y. Ma, H. C. Liu, D. Deb, H. Liu, J. L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020, doi: 10.1007/s11633-019-1207-4.

[61] O. O. Olakanmi and A. Dada, "Wireless sensor networks (WSNs): security and privacy issues and solutions," *Wireless Mesh Networks-Security, Architectures and Protocols*, pp. 1–16, 2020, doi: 10.1007/978-3-030-53035-7_1.

[62] U. Iqbal and A. H. Mir, "Secure and practical access control mechanism for WSN with node privacy," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3630–3646, 2022, doi: 10.1016/j.jksuci.2021.06.008.

[63] O. O. Olakanmi, "A lightweight security and privacy-aware routing scheme for energy-constraint multi-hop wireless sensor networks," *International Journal of Information and Computer Security*, vol. 15, no. 2–3, pp. 231–253, 2021, doi: 10.1504/IJICS.2021.110293.

[64] M. A. Elsadig, A. Altigani, and M. A. Baraka, "Security issues and challenges on wireless sensor networks," *Int. J. Adv. Trends Comput. Sci. Eng*, vol. 8, no. 4, pp. 1551–1559, 2019.

[65] Z. W. Hussien, D. S. Qawasmeh, and M. Shurman, "MSCLP: Multi-sinks cluster-based location privacy protection scheme in WSNs for IoT," in *2020 32nd International Conference on Microelectronics (ICM)*, pp. 1–4, 2020, doi: 10.1109/ICM48755.2020.9318780.

[66] B. Bhushan and G. Sahoo, "Requirements, protocols, and security challenges in wireless sensor networks: an industrial perspective," in *Handbook of Computer Networks and Cyber Security: Principles and Paradigms*, pp. 683–713, 2020.

[67] X. Qi, X. Liu, J. Yu, and Q. Zhang, "A privacy data aggregation scheme for wireless sensor networks," *Procedia Computer Science*, vol. 174, pp. 578–583, 2020, doi: 10.1016/j.procs.2020.09.115.

[68] F. Mukamanzi, M. Raja, T. Koduru, and R. Datta, "Position-independent and section-based source location privacy protection in WSN," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 5, pp. 6636–6646, 2022, doi: 10.1109/TII.2022.3152999.

[69] D. S. Ibrahim, A. F. Mahdi, and Q. M. Yas, "Challenges and issues for wireless sensor networks: a survey," *J. Glob. Sci. Res*, vol. 6, no. 1, pp. 1079–1097, 2021, doi: 10.47013/jgsr.v6i1.123.

[70] A. O. Salau, N. Marriwala, and M. Athaee, "Data security in wireless sensor networks: attacks and countermeasures," in *Mobile Radio Communications and 5G Networks: Proceedings of MRCN 2020*, pp. 173–186, 2021, doi: 10.1007/978-981-16-1874-9_19.

[71] R. Manjula, T. Koduru, and R. Datta, "Protecting source location privacy in IoT-Enabled wireless sensor networks: the case of multiple Assets," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10807–10820, 2022, doi: 10.1109/JIOT.2021.3062046.

[72] T. M. Mengistu, T. Kim, and J. W. Lin, "A survey on heterogeneity taxonomy, security and privacy preservation in the integration of IoT, wireless sensor networks and federated learning," *Sensors*, vol. 24, no. 3, p. 968, 2024, doi: 10.3390/s24030968.

[73] F. Alshohoumi, M. Sarrab, A. AlHamadani, and D. Al-Abri, "Systematic review of existing IoT architectures security and privacy issues and concerns," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019.

[74] K. Ramasamy, M. H. Anisi, and A. Jindal, "E2DA: Energy efficient data aggregation and end-to-end security in 3D reconfigurable WSN," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 787–798, 2021, doi: 10.1109/TGCN.2020.3040870.

[75] K. Özlem, M. K. Kuyucu, Ş. Bahtiyar, and G. İnce, "Security and privacy issues for E-textile applications," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pp. 102–107, 2019, doi: 10.1109/UBMK.2019.8873505.

[76] E. Romero, J. Blesa, and A. Araujo, "An adaptive energy aware strategy based on game theory to add privacy in the physical layer for cognitive WSNs," *Ad Hoc Networks*, vol. 92, pp. 101800, 2019, doi: 10.1016/j.adhoc.2019.101800.

[77] G. S. Ilgi and Y. K. Ever, "Critical analysis of security and privacy challenges for the Internet of drones: a survey," in *Drones in Smart-Cities*, pp. 207–214, 2020.

[78] S. Zhong, H. Zhong, X. Huang, P. Yang, J. Shi, L. Xie, and K. Wang. *Security and Privacy for Next-Generation Wireless Networks*. Springer International Publishing, 2019, doi: 10.1007/978-3-030-12003-7.

[79] S. Lata, S. Mehfuz, and S. Urooj, "Secure and reliable WSN for internet of things: challenges and enabling technologies," *IEEE Access*, vol. 9, pp. 161103–161128, 2021, doi: 10.1109/ACCESS.2021.3128504.

[80] N. Jan, A. H. Al-Bayatti, N. Alalwan, and A. I. Alzahrani, "An enhanced source location privacy based on data dissemination in wireless sensor networks (DeLP)," *Sensors*, vol. 19, no. 9, p. 2050, 2019, doi: 10.3390/s19092050.