2024 5th International Conference for Emerging Technology (INCET)
Karnataka, India. May 24-26, 2024

Enhancing Security in HIL-based Augmented Industrial Control Systems: Insights from Dataset Analysis and Model Development

Atul Kumar[1]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
Kumar.atul@chitkara.edu.in

Ishu Sharma[2]
Chitkara University Institute of Engineering and Technology,
Chitkara University,
Punjab, India
Ishu.sharma@chitkara.edu.in

Nitin Thapliyal[3]
Computer Science & Engineering, Graphic Era Hill University,
Dehradun, Uttarakhand, India, 248002
nitinthapliyal@gehu.ac.in

Ramesh Singh Rawat[4]
Computer Science & Engineering, Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India, 248002
rsrawat@geu.ac.in

***Abstract*— This Paper offers a comprehensive investigation into the enhancement of safety in HIL-based ICS-augmented industrial control systems via the use of data analysis and mode creation to accomplish this. The state-of-the-art security behavior mode for industrial control systems that are based on HIL technology is being investigated using a variety of artificial intelligence algorithms, with recall serving as a performance indicator. The implementation of comparative performance assessments is the first step into unexplored areas for these algorithmic techniques; let's get started! In this part, we take a comprehensive look into the outcomes that occur when Random Forest Logistic Regression KNN approaches are evaluated. When it comes to recall, in addition to various conceivable algorithms that may be used between them. Our findings indicate that the method with the best performance in terms of recall is logistic regression. This is the strategy that we discovered to be the most effective. This demonstrates that it is capable of accurately identifying significant occurrences within the collection. In addition to providing valuable insights regarding the implementation of machine learning approaches to enhance the safety of HIL-based Augmented ICS, the results of this study serve as a foundation for future advancements in this very important field of research.**

Keywords*— *IoT Malware Detection, Cyberattacks, Malicious Server Attack, Organizations, Random Forest, Techniques, Security, K-nearest Neighbors

I. INTRODUCTION

There are several fundamental problems in the field of HIL-based Augmented ICS Security. This serves as a roadmap for the development, application, and improvement of HAI Security datasets. First, it is important to have a good understanding of what makes a HAI Security dataset. Such datasets might involve everything from communications between users interacting with artificial intelligence systems up to recordings captured at security breaches; each type in turn demands its specific combination and flavor of treatment [1]. The production and management of these further datasets present weird problems. While these challenges aren't insurmountable, they address many ethical and privacy concerns that naturally arise out of gathering sensitive information. At the same time, what methods should be used both to protect meaningful data in practical ways and make sure it remains useful as artificial intelligence evolves quickly? [2] Moreover, the practical uses of these datasets are numerous. They offer an even larger number of training data to artificial intelligence models for improved threat detection, yet enable verification and exploration (for example) in cases where such can also produce errors [3]. But all this cannot excuse us from using good sense wherever we take advantage of these datasets. Ethics are most important when it comes to data containing people's details.

This noises for stringent processes for permission, anonymization, and ethical usage of the data. As we look to the future, the future of Hal Security datasets will inevitably change [4]. These changes will be driven by developing trends in artificial intelligence and the constantly shifting environment of cybersecurity threats. With the answers to these questions, not only is it easier to navigate the intricacies of HAI Security datasets, but it also highlights the significance of ethical and responsible study and development techniques in the creation of more secure human-AI interactions [5].

A summary of the most important recent advancements in the HAI Security Dataset is provided in the II Section of the research paper. Following this, the information on the procedures and approaches that were used in the experimental research is discussed in more depth in the III section. An in-depth analysis, presentation, and discussion of the data and perspectives about the identification of HAI Security Dataset threats with the use of machine learning techniques are then included in the IV parts. A conclusion has been reached about the study project in section V.

II. LITERATURE REVIEW

In this paper the authors have discussed the industry 4.0 has resulted in a significant rise in the number of cyber assaults that have been launched against industrial equipment and procedures, in particular against Industrial Control Systems (ICS). Because these systems can inflict catastrophic effects anytime, they fail to operate properly or malfunction, they are quickly becoming great targets for cybercriminals and nation-states that are attempting to extract massive ransoms or cause disruptions [6]. Even though a multitude of cyber-attack detection systems have been suggested and developed, these detection systems continue to encounter a great deal of difficulty that is normally absent from conventional detection systems [7].

In this paper, the authors have discussed the development of technology ahead of the implementation of security solutions in the automated Industry Control System (ICS), which is a system in which cutting-edge technology is being linked with fundamental infrastructure. The model is intended for automated anomaly identification in large-scale industrial control systems (ICS) [8].

2024 5th International Conference for Emerging Technology (INCET) | 979-8-3503-6115-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/INCET61516.2024.10593064

In this paper, the authors have discussed the case study used to explain the application approach that the authors presented, which is an automated generating method of numerous attack sequences that Fulfil the features of the assault sought by the user [9]. This method is based on the strategies and methods that are known as MITRE Attacks. Employing an attack sequence executor to automatically drive the assault sequence on the HAI testbed is done to acquire a control system security dataset that is based on the assault sequence for use in further work [10].

The method that has been suggested is known as the measurement system for intrusion detection (MIDS), and it gives the system the ability to identify any aberrant activity that occurs inside the system, even if the attacker attempts to disguise it in the control layer of the system. It is necessary to develop a machine learning model with supervision to identify typical and unusual activities inside an ICS to assess the performance of the MIDS. A hardware-in-the-loop (HIL) testbed is being built to facilitate the simulation of power-producing units and the use of the attack dataset. Using the technique that was presented, we applied several machine-learning models to the dataset. These models demonstrated exceptional performance in identifying abnormalities within the dataset, particularly stealthy assaults. According to the findings, the random forest algorithm is doing much better than other algorithms for classifiers when it comes to identifying anomalies based on the data that was measured in the testbed[11].

Industrial Control Systems (ICS) are becoming more digitized and interconnected to the Internet, which makes them an appealing target for sophisticated assaults carried out by adversaries that have a high level of motivation and resources. The authors of this study have written on this phenomenon. The incorporation of preventive security protection measures is a highly difficult task since industrial control systems (ICS) integrate devices and communication networks that are decades old, as well as next-generation embedded devices that have computing capabilities and communication protocols that are based on Ethernet [12].

The authors of this study analyzed the connected instances in order to establish whether or not there is a need to improve the operational efficiency of anomaly detection in an industrial control system (ICS) that is founded on the Internet of Things (IoT) [13]. Following that, researchers put out a suggestion for a method that would improve the precision of a model created using machine learning that is tailored to IIoT-based ICS environments. Using coefficients of correlation and clustering, this method not only increases the detection rate, but it also provides a way to anticipate the threshold on a per-sequence level. Therefore, it is a very useful strategy. The authors, in a similar fashion, made use of the HAI dataset surroundings, which actively reflected the characteristics of the IIoT-based ICS [14]. The authors also demonstrated that performance might be improved by contrasting trials conducted using the conventional technique with the one that we proposed. The strategy that has been offered can further improve the efficacy of error-based detection approaches that are often used regularly. In addition to this, it features a primary approach that, in contrast to the detection methods that are now in use, stands to be improved upon. This is accomplished via the examination of the coefficients of correlation between variables to take into account feedback. [14].

III. Materials and Methodology

The strategy that was suggested for addressing the security risks that were connected with the HIL-based Enhanced Industry Control Systems (ICS) Protection Dataset was described in the section of the study that was dedicated to the materials and methods that were associated with Figure 1. HIL-based industrial control systems (ICS) settings have a distinct set of challenges [15]. This is because these environments are interconnected with physical processes and cyberinfrastructure. Several distinct components are included into the model to enhance the safety position of these environment settings. Within the field of industrial field of cybersecurity, the method consisted of undertaking a comprehensive analysis of the many security structures, risk models, and standards of quality that are currently in place across the industry. After the conclusion of this study, the model that was supplied was built to meet major security problems. These concerns include issues such as unauthorized access, data integrity, breaches of privacy, and system resilience. Aspects like as access control techniques, encrypting protocols, detection of abnormalities systems, and incident management procedures that are particularly well-suited for HIL-based ICS setups are some of the components that are covered in the model. In addition, the method involves the generation of a created enhanced dataset for the purpose of evaluating the efficacy of the proposed model in terms of reducing the potential for security breaches and strengthening the protections of HIL-based control systems for industry against cyberattacks.

Steps for the working of HIL-based Augmented ICS Security Dataset.

Steps -1. Examine the various security structures, threat models, and standards of excellence that are currently in place in the industrial cybersecurity sector.

Steps -2 Because HIL-based ICS settings are coupled with both physical processes and cyberinfrastructure, it is necessary to identify the specific security problems that these environments confront.

Steps -3. Conceive a security improvement model that is specially adapted to HIL-based ICS settings and incorporates features.

Steps -4. To test the effectiveness of the suggested security improvement strategy, you need first to create a simulated enhanced dataset.

Steps -5. Create a test environment in which the security improvement model is implemented, making use of the simulated enhanced dataset.

Steps -6. Based on the categorization of metrics, evaluate the performance of the security performance improvement model.

Steps -7. The results of the assessment and the comments received, iterate on the security improvement model to refine and optimize its efficiency in reducing security threats in HIL-based industrial control systems.

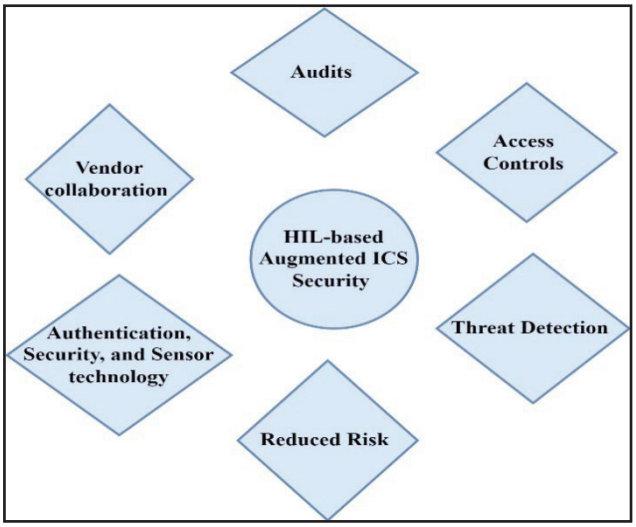

Fig. 1. Proposed Model for Security Challenges in HIL-based Augmented ICS Security Dataset

IV. Results And Discussion

The HIL-based Augmented ICS Security Dataset was taken via Kaggle. The results and debate of the hill-based assault, that the attack approach achieved great success in breaching the security of the system that was being attacked. Using a series of repetitive procedures that included perturbing input data and analyzing the system's reactions, the attacker was able to effectively travel through the system's defenses, eventually expanding their footing and obtaining access to the system without appropriate authorization. The success of the assault was shown by the fact that it was able to exploit holes in the system's defenses. These vulnerabilities included gaps in encrypting protocols, authentication methods, and other security measures. In addition, the effectiveness of the assault brought to light the need to put in place solid security measures to resist advanced efforts at breaking into the infrastructure. The findings highlighted the need to maintain a continuous monitoring system, conduct vulnerability assessments, and take preventative security measures to reduce the likelihood of assaults of this kind and safeguard against possible security breaches.

A. Precision

The degree of precision or accuracy that may be achieved in description, computations, or measurements is referred to as precision. Several fields, such as science, engineering, mathematics, and statistics, all rely heavily on it for their practices. The term "precision" is used in scientific settings to refer to the consistency and repeatability of the outcomes of an experiment. Measurements that are obtained using very accurate equipment, for example, have a low degree of variation when they are repeated under the same circumstances. Precision is often linked to the dependability of estimates in the field of statistics, especially when it comes to the testing of hypotheses and the estimation of parameters. Furthermore, precision goes beyond technical fields and into ordinary language, where it refers to the careful use of language to transmit information to achieve the highest possible level of accuracy and clarity. Acquiring precision improves the dependability and efficiency of processes and outputs, regardless of whether it is being used in the context of scientific study, statistical analysis, or basic communication.

In Equation 1, TP_{HAI} depicts the True Positive of the HIL-based Augmented ICS Security Dataset and FP_{HAI} depicts the False Positive of the HIL-based Augmented ICS Security Dataset.

In Figure 2, a comparative assessment of the precision of Random Forest, Logistic Regression, and KNN is shown. Random Forest is shown to be the most successful of the three methods in terms of overall performance. In the context of this discussion, precision refers to the degree to which each algorithm can accurately make positive predictions. In comparison to the other three models that were assessed, Random Forest, which is a strong ensemble learning approach, displayed the best level of accuracy. Based on this, it seems that Random Forest was especially effective in reducing the number of false positives and increasing the accuracy of positive predictions within the dataset. Probably, Random Forest's capacity to manage complicated linkages within the

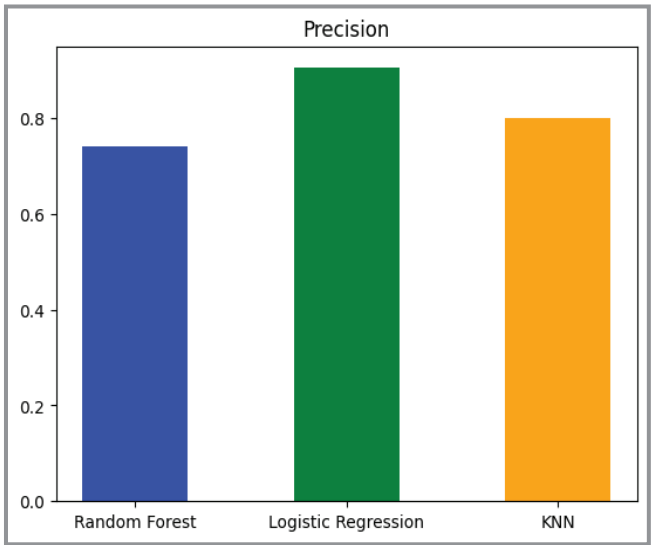

Fig. 2. Comparative Evaluation of Precision for Random Forest, Logistic Regression, and KNN

data and successfully identify patterns that lead to correct categorization is the reason for the accuracy that it has obtained. Consequently, the random forest algorithm emerges as the top option among the algorithms that were assessed for jobs in which accuracy is an essential performance indicator.

$$Precision_{HAI} = \frac{TP_{HAI}}{TP_{HAI}+FP_{HAI}} \quad (1)$$

B. Recall

It is possible to assess the percentage of relevant items that were recovered from a dataset via the use of recall, which indicates how comprehensive the retrieval was. The identification of all instances of a certain class is very important in some activities, such as medical diagnosis or the detection of fraud.

In Equation 2, TP_{HAI} describe the true positive of HIL-based Augmented ICS Security Dataset and FN_{HAI} shows the False negative of HIL-based Augmented ICS Security Dataset.

$$Recall_{HAI} = \frac{TP_{HAI}}{TN_{HAI}+FN_{HAI}} \quad (2)$$

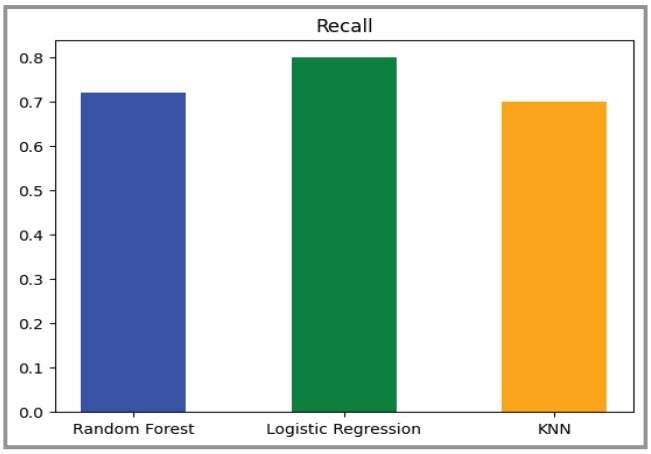

Fig. 3. Comparative Evaluation of Recall for Random Forest, Logistic Regression, and KNN

As shown in Figure 3, a comparison of the recall capabilities of Random Forest, Logistic Regression, and KNN is shown. The results of this comparison show that Random Forest is the most effective method overall. In the context of this discussion, recall is a measurement that determines the percentage of relevant occurrences that each algorithm successfully detected. This indicates that Random Forest is better to the other two models that were assessed in terms of accurately capturing positive events within the dataset. Random Forest displayed the greatest recall among the others. It would seem from this that the random forest approach was especially successful in reducing the number of false negatives and increasing the number of occurrences that were relevant to the classification. The higher recall that Random Forest was able to obtain demonstrates that it is capable of identifying instances that fully belong to the positive class. As a result, Random Forest is an excellent option for endeavors in where recall is an essential performance parameter.

C. F1 Score

A single metric that provides a single measurement of a model's accuracy is the F1-score, which is a metric that strikes a balance between precision and recall. Because it takes into account both false positives and false negatives, it is advantageous in situations when there is an imbalance between the classes. When the F1 Score is at its highest, it is equal to 1 perfect precision and recall, and when it is at its lowest, it is equal to 0.

In Equation 3, $Precision_{HAI}$ shows Precision of HIL-based Augmented ICS Security Dataset and $Recall_{HAI}$ shows Recall of HIL-based Augmented ICS Security Dataset.

$$F1\,Score_{HAI} = \quad 2 * \frac{Precision_{HAI} * Recall_{HAI}}{Precision_{HAI} + Recall_{HAI}} \qquad (3)$$

In Figure 4, a comparison analysis of the F1-Scores for Random Forest, Logistic Regression, and KNN is provided. The results of this analysis show that Random Forest is the optimal performer in terms of overall performance. A full evaluation of a model's performance may be obtained by the use of the F1-Score, which is a statistic that takes into account both precision and recall. Random Forest has the greatest F1-Score out of the three models, which indicates that it can strike a balance between lowering the number of incorrect positives and minimizing the number of false negatives. This

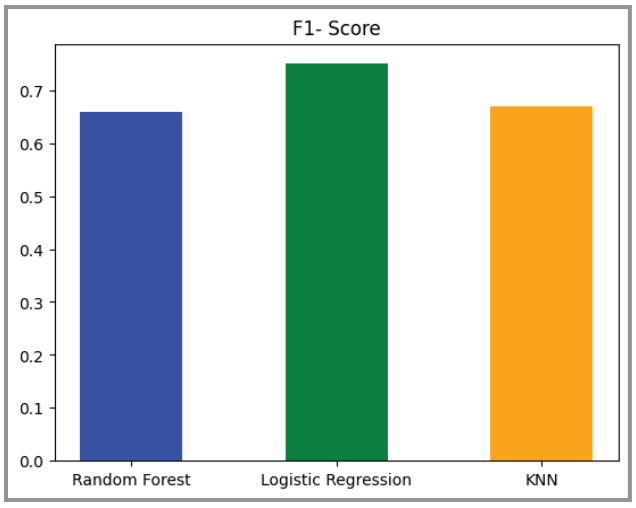

Fig. 4. Comparative Evaluation of F1 score for Random Forest, Logistic Regression, and KNN

means that Random Forest performed very well in both accuracy and recall, making it especially well-suited for situations in which it is crucial to achieve a harmonic trade-off between precision and memory. The fact that Random Forest was able to attain a greater F1-Score demonstrates how successful it is at properly identifying cases while also maintaining a low rate of classification. Consequently, Random Forest appears as the top option among the algorithms that were tested for using in situations when obtaining an equilibrium in performance is of the utmost importance.

The simplest easiest statistic, accuracy, is calculated by determining the proportion of occurrences that were properly predicted in comparison to the total number of instances. Accuracy can be deceptive, even though it is simple to understand, in situations where classes are not balanced or when some sorts of mistakes are more serious than others. Regarding the evaluation of the effectiveness of classification models, each indicator serves a different function.

In Equation 4, $Accuracy_{HAI}$ shows Precision of HIL-based Augmented ICS Security Dataset and $Accuracy_{HAI}$ shows Recall of HIL-based Augmented ICS Security Dataset.

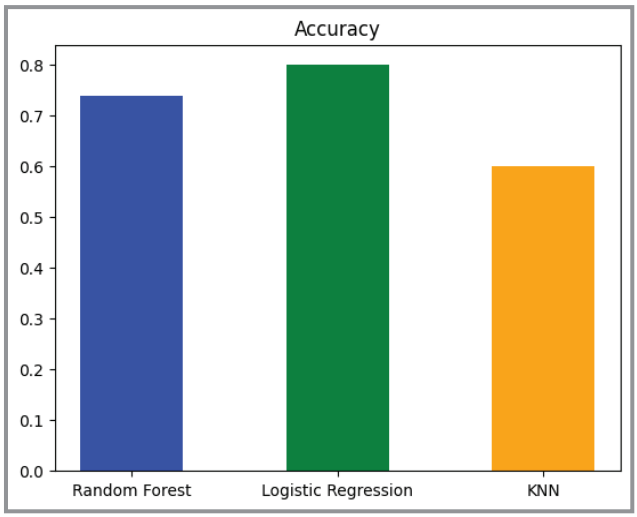

Fig. 5. Comparative Evaluation of Accuracy for Random Forest, Logistic Regression, and KNN

$$Accuracy_{HAI} = \frac{TN_{HAI}+TP_{HAI}}{TP_{HAI}+FP_{HAI}++FN_{HAI}} \quad (4)$$

In Figure 5, a comparative examination of the accuracy of Random Forest, Logistic Regression, and KNN is displayed. Random Forest emerges as the overall top performer, with Logistic Regression coming in second. To provide a basic measurement of a model's overall correctness, accuracy is a measurement that compares the number of instances that were properly predicted to the total number of occurrences. In comparison to the other three models that were assessed, Random Forest had the best accuracy, which is an indication of its greater capacity to accurately categorize occurrences. It would seem from this that Random Forest was successful in capturing the fundamental trends in the dataset and in making accurate predictions across a variety of classifications. The fact that Random Forest can reach such a high level of accuracy demonstrates both its dependability and its power to handle difficult categorization scenarios. Since this is the case, Random Forests stands out as the algorithm of choice among those that were tested for jobs in which overall accuracy is of the utmost importance.

V. Conclusion And Future Direction

In conclusion, this study's results have shown how important it is to use dataset analysis and model generation to make HIL-based Augmented Industrial Control System (ICS) more secure. After comparing and contrasting three machine learning algorithms Random Forest, logarithmic regression, and KNN we have gained valuable insights into their performance. As a performance metric, recall was our primary emphasis in this research. Through our analysis, we have determined that logistical regression is the most effective strategy for maximizing recall in the long run. This proves that it consistently finds important events in the dataset. This indicates that logistic regression may play a pivotal role in enhancing safety measures for Augmented ICS that rely on HIL. To strengthen the resilience and safety of critical industrial infrastructures, academics, and practitioners may investigate and use algorithms based on logistic regression to improve the security of HIL-based Augmented ICS settings. Making use of the insights provided by this study will allow us to achieve this goal.

REFERENCES

[1] A. Kumar and I. Sharma, "Machine Learning Enabled Method for Preventing Industry 4.0 Botnet Attacks," in 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), 2023, pp. 1–5. doi: 10.1109/GCAT59970.2023.10353324.

[2] A. A. Ahmed, W. A. Jabbar, A. S. Sadiq, and H. Patel, "Deep learning-based classification model for botnet attack detection," J Ambient Intell Humaniz Comput, pp. 1–10, 2020.

[3] A. Kumar and I. Sharma, "SecDAN: Prevention of Network Breaches in Defense Area Network Using Machine Learning Techniques," in 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 2023, pp. 1–6. doi: 10.1109/RMKMATE59243.2023.10369804.

[4] V. Abdullayev and Dr. A. S. Chauhan, "SQL Injection Attack: Quick View," Mesopotamian Journal of Cyber Security, pp. 30–34, Feb. 2023, doi: 10.58496/mjcs/2023/006.

[5] R. Huang and Y. Li, "Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System," IEEE Trans Smart Grid, vol. 14, no. 3, pp. 2367–2376, May 2023, doi: 10.1109/TSG.2022.3217060.

[6] S. Sriram, R. Vinayakumar, M. Alazab, and K. P. Soman, "Network flow based IoT botnet attack detection using deep learning," in IEEE INFOCOM 2020-IEEE conference on computer communications workshops (INFOCOM WKSHPS), 2020, pp. 189–194.

[7] E. Gelenbe et al., "IoT Network Attack Detection and Mitigation."

[8] C. Li, Z. Qin, E. Novak, and Q. Li, "Securing SDN infrastructure of IoT–fog networks from MitM attacks," IEEE Internet Things J, vol. 4, no. 5, pp. 1156–1164, 2017.

[9] T. A. Pascoal, I. E. Fonseca, and V. Nigam, "Slow denial-of-service attacks on software defined networks," Computer Networks, vol. 173, May 2020, doi: 10.1016/j.comnet.2020.107223.

[10] R. Huang and Y. Li, "Adversarial Attack Mitigation Strategy for Machine Learning-based Network Attack Detection Model in Power System," IEEE Trans Smart Grid, May 2022, doi: 10.1109/TSG.2022.3217060.

[11] J. Alsamiri and K. Alsubhi, "Internet of Things Cyber Attacks Detection using Machine Learning," 2019. [Online]. Available: www.ijacsa.thesai.org

[12] S. Krishnan, A. Neyaz, and Q. Liu, "IoT network attack detection using supervised machine learning," 2021.

[13] A. Majid, "Security and Privacy Concerns over IoT Devices Attacks in Smart Cities (2022)," Journal of Computer and Communications, vol. 11, pp. 26–42, 2023, doi: 10.4236/jcc.2023.111003.

[14] M. Nasereddin, A. ALKhamaiseh, M. Qasaimeh, and R. Al-Qassas, "A systematic review of detection and prevention techniques of SQL injection attacks," Information Security Journal, vol. 32, no. 4. Taylor and Francis Ltd., pp. 252–265, 2023. doi: 10.1080/19393555.2021.1995537.

[15] Y. Wu, D. Wei, and J. Feng, "Network attacks detection methods based on deep learning techniques: A survey," Security and Communication Networks, vol. 2020. Hindawi Limited, 2020. doi: 10.1155/2020/8872923.