Data Analysis for Predicting Attacks in Smart Systems Using AI Techniques

Reem Mushari Albalawi
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
452009782@stu.ut.edu.sa

Shrouq Khalaf Alaazi
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
452010149@stu.ut.edu.sa

Thuraya Saeed Alshahrani
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
452010089@stu.ut.edu.sa

Tahani Mohammed Alanazi
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
452010097@stu.ut.edu.sa

Wejdan ALomrani
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
452010183@stu.ut.edu.sa

Majed Aborokbah
Faculty of Computers & Information Technology
University of Tabuk
Tabuk, Saudi Arabia
m.aborokbah@ut.edu.sa

***Abstract*— This paper presents an AI-driven framework to predict and mitigate DDoS attacks in smart cities using SoftwareDefined Networking (SDN), federated learning, and advanced machine learning algorithms. Parameter optimization significantly enhanced model performance. The Decision Tree algorithm, optimized via Research, achieved an accuracy of 0.9236 and an F1-Score of 0.929, outperforming Logistic Regression. Deep Learning models also improved with strategies like Early Stopping and ReduceLROnPlateau, with SimpleRNN showing the highest improvement, followed by GRU, LSTM, and CNN models. These results surpass previous research, demonstrating the effectiveness of the applied techniques and model preparation.**

Keywords*— *Smart Cities, Cybersecurity, Distributed Denial of Service (DDoS), Artificial Intelligence (AI), Software-Defined Networking (SDN), Federated Learning, Internet of Things (IoT), Real-time threat detection.

I. INTRODUCTION

To improve urban life, the smart city concept embraces various advanced technologies including, The Internet of Things (IoT), vast amounts of big data, and artificial intelligence. However, the rapid adoption of technology also increases threat levels, particularly in terms of cybersecurity. There are inherent dangers to the proper functioning of smart cities such as distributed denial of service (DDoS) attacks that pose threats to critical systems and operations across the transportation, healthcare, and energy management infrastructure. This paper explores the development of a unique AI technology aiming to hamper such threats before they occur.

II. BACKGROUND AND RELATED WORD

Integrating Software-Defined Networking (SDN) into smart cities has been researched as a possible way to improve the security and management of the networks. Past studies like "A Secure and Intelligent Software-Defined Networking Framework for Future Smart Cities to Prevent DDoS Attack, have designed DDoS attack learning-based frameworks using SDN Artificial Intelligence- Based Secured Power Grid Protocol for Smart City [4]. A human-centered artificial intelligence approach for privacy protection of elderly App users in smart cities. Another relevant one can be found in "Internet of Things for Smart Cities," in which the author argues the need for a single IoT architecture that allows for the integration of diverse devices and the analysis of vast amounts of data for effective urban services management. Further, the significance of artificial intelligence in protecting smart grid systems, as contained in "Artificial Intelligence Based Secured Power Grid Protocol for Smart City".

Encompasses the application of Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) as cyber threat detectors and energy data optimizers [3] [4] [7].

The idea of smart cities has also been scrutinized in relation to big data and urban governance for example in the paper titled 'RK SUMMRIZE'. The paper focuses on the use and application of digital technologies and big data analysis for improving urban management and addressing concerns of technocracy, privacy, and surveillance with the help of real-time analytics. It also touches upon the benefits of urban data analytics in real-time and the dangers that accompany such techniques [6].

These undertaken studies, elaborately articulate how to secure smart city infrastructure, that is IoT-enabled, through advanced, hybrid, and efficient real-time anomaly and intrusion detection systems with the use of developed computational intelligence models. Operationalizing ensemble models, such as Gradient Boosting Machines, Random Forest, AdaBoost, and deep models along with the grey wolf optimization techniques enhances the model's accuracy and efficiency. Deploying these models into fog computing environments and SCADA (supervisory control and data acquisition), Industrial sensor networks (ISN), and smart power grids assist in addressing the challenges associated with IoT systems supported by cyber-physical infrastructural systems operating in a more decentralized way. Principal component analysis, Kernel PCA-based feature selection, and explainable AI tools have further improved the detection and classification performance by enhancing feature interpretability while at the same time reducing the computational loads. Real-world validation for instance shows the streams of these models on the datasets (CICIDS2017, UNSW- NB15, MSU-ORNL) has often recorded accuracy rates of more than 95% making them ideal in real-world applications for simulating cyber security in critical smart city infrastructures and maintains scalability and resilience [1][9]]11].

- *Case Stud*

Padova Smart City: A case study on the "Padova Smart City" project in Italy demonstrates the effectiveness of the proposed framework. The project integrates IoT and big data analytics to manage public services like traffic flow, air quality monitoring, and energy consumption. The results show that the proposed AI-based framework can significantly enhance the resilience of smart city infrastructures to cyber-attacks, while also improving the efficiency of urban service management [8].

III. Data Sets

Data preparation and cleaning play a pivotal role in enhancing the performance of machine learning algorithms, particularly in the context of smart city cybersecurity. Data collected from IoT networks and infrastructure is often prone to duplication or missing records. Addressing these issues reduces redundancy, improves data integrity, and boosts the efficiency of models in detecting Distributed Denial-of-Service (DDoS) attacks [5].

To ensure the data was cleaned, formatted, and prepared for cybersecurity analysis, the following steps were implemented:

1. *Removing Unnecessary Data:* Repetitive and irrelevant information was eliminated to reduce computational overhead and focus on essential data.
2. *Handling Missing Values:* Missing values were either imputed using estimated values or removed, ensuring data completeness.
3. *Outlier Detection and Management:* Outliers that could negatively impact model performance or produce inaccurate results were identified and subsequently removed or adjusted.
4. *Normalization:* Numerical values were scaled to a uniform range (e.g., 0 to 1), ensuring that features with larger values did not dominate the analysis.
5. *Standardization:* Data was transformed to have a mean of zero and a standard deviation of one, improving the model’s ability to analyze data effectively.

The visualization of preprocessing methods demonstrated that Min-Max scaling effectively standardized the data range while preserving the original distribution's characteristics, such as skewness. This insight is critical when preparing data for machine learning models that assume or benefit from normally distributed data.

A. *Dataset Overview*

The dataset used in this study contains 82,333 samples and 45 features, comprising both numerical and categorical attributes. These features represent key characteristics relevant to cyberattack detection, such as traffic statistics, protocol details, and payload attributes. The dataset is structured to support multiple tasks, including:

a. Binary Classification: Identifying attack versus normal traffic.

b. Multi-Class Classification: Differentiating between various attack types.

c. Regression: Predicting metrics such as detection time or attack severity.

B. Preprocessing Steps

The data preprocessing pipeline involved the following actions to enhance model performance:

a. Normalizing Numerical Features: Ensured consistency across features with varying scales.

b. Encoding Categorical Features: Transformed non-numerical data into numerical representations.

c. Addressing Class Imbalances: Balanced the dataset to prevent bias in machine learning models.

This rigorous data preparation ensures compatibility with advanced machine learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRU), making the dataset suitable for comprehensive cybersecurity analysis.

The plots in Figure 1 illustrate that Min-Max scaling efficiently normalizes the range of the data between 0 and 1 while maintaining the shape and skewness of the original distribution. This characteristic ensures that the distribution's essential features remain intact. Such a transformation is particularly valuable when preparing data for machine learning models that either require data to be in a specific range or perform better with distributions that closely resemble the original dataset's characteristics.

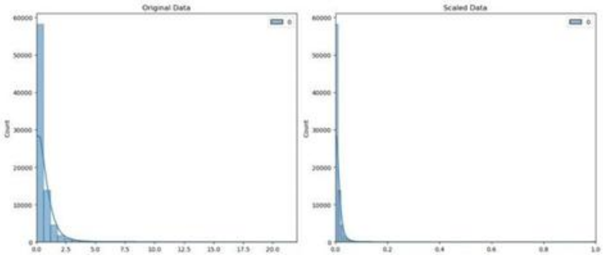

Fig.1. Scaled Data

The plots in Figure 2 demonstrate that Z-score normalization (or standardization) efficiently centers the data around a mean of zero and scales it to have a standard deviation of one. Importantly, it retains the original distribution's characteristics, including skewness. This method is especially beneficial for preparing data for machine learning models that require standardized input or are sensitive to features with differing scales.

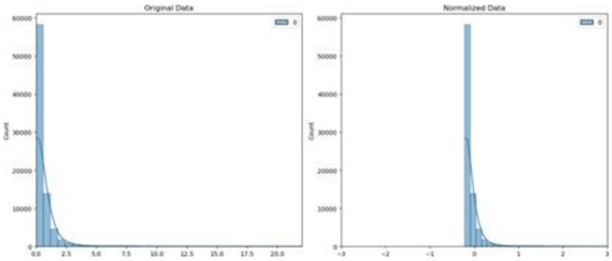

Fig.2. Normalized Data

IV. METHODOLOGY

This research adopts an AI-based methodology that integrates advanced machine learning techniques with SDN to address cybersecurity challenges in smart cities. The methodology includes the following key components:

A. *DDoS Detection and Mitigation*

As already mentioned, the use of machine learning algorithms such as LSTM and RNN helps in identifying anomalies in network traffic that are indicative and can lead to possible DDoS attacks. By using federated learning data is kept secure as the analysis is done on data that isn't centralized and exposes sensitive data [5].

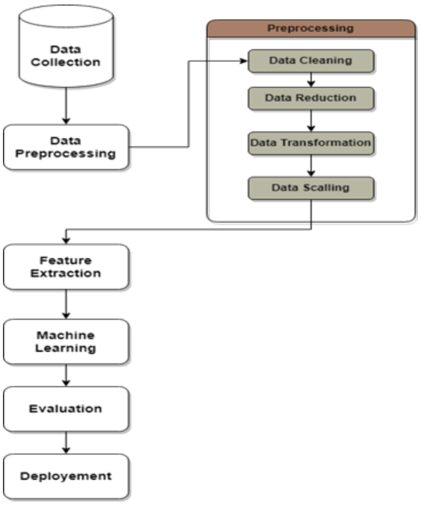

Fig.3. Data Curation Workflow

The diagram illustrates the data collection, preprocessing (including cleaning, reduction, transformation, and scaling), feature extraction, and machine learning stages, followed by evaluation and deployment.

B. *Privacy Protection for Vulnerable Populations*

In smart cities, elderly users' privacy concerns are addressed within the framework of a Human-Centered AI approach. This includes Participatory Privacy Protection Algorithms (PPPA-I and PPPA-II) to adjust the privacy requirements for Ambient Assisted Living (AAL) applications reducing the cognitive load of elderly users [4].

The following diagram illustrates the proposed framework that integrates SDN, Federated Learning, and Machine.

Learning algorithms to enhance cybersecurity in smart cities:

a. SDN Controller: Centrally manages network resources and directs traffic flow.

b. Federated Learning: Enables secure data sharing between Internet of Things (IoT) devices without the need for direct data exchange.

c. Anomaly Detection Models (LSTM/RNN): These machine learning models analyze network traffic to identify unusual activities, potentially indicating cyberattacks.

V. PROPOSED FRAMEWORK

The proposed Smart Cities Cyber Defense Framework utilizes SDN capabilities as well as federated learning and ML models including CNN, GRU, Decision Trees, and Logistic Regression to increase the cybersecurity posture of smart cities.

All network resources are controlled through the SDN controller, while federated learning enables secure data sharing across IoT devices. Different ML models, including LSTM, RNN, CNN, and GRU, are used for network traffic analysis which aids in identifying unusual activities and effective threat recognition. Furthermore, Decision Trees and Logistic Regression are used as supplementary tools for the detection and classification of anomalies as they are lightweight and interpretable options for real-time threat identification and mitigation.

The framework is adaptable to various smart city domains, such as traffic management, energy distribution, and healthcare.

SDN Controller helps to monitor traffic attributes as well as administrate control over the traffic network.

Federated Learning guarantees that models are deployed at several locations without the danger of exposing information. Temporal models such as LSTM, RNN, CNN, and GRU are utilized for the analysis of data validity by applying an outlier scoring method which determines anomalies in the dataset that correspond to abnormal activity.

Anomaly Detection, supported by Decision Trees and Logistic Regression, flags irregular patterns, while Threat Detection further analyzes these anomalies to recognize actual threats.

Privacy Protection safeguards personal data from leakage. Also, Traffic Data is studied to find probable attacking ways, and Mitigation techniques are applied to those discovered threats in order to lessen their effect using the advantages of these modern machine learning approaches.

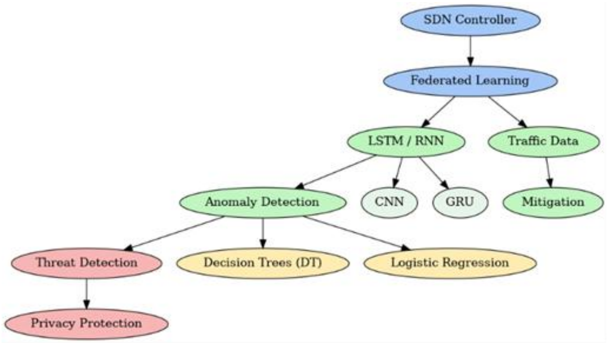

Fig.4. Federated Learning Model for Network Control: Threat Detection and Privacy Protection.

	Previous Models (Before Enhancement) [13]		**Proposed Models (Baseline)**		**Previous Models (After Enhancement) [13]**		**Proposed Models (After Enhancement)**	
	LR	DT	LR	DT	LR	DT	LR	DT
Accuracy	0.538	0.733	0.7376	0.8395	0.7232	0.8069	0.7461	0.9236
Precision	0.414	0.721	0.8131	0.9436	0.72	0.81	0.8327	0.9477
Recall	0.538	0.733	0.6777	0.7525	0.72	0.81	0.6723	0.9111
F1 Score	0.397	0.705	0.7392	0.8373	0.71	0.80	0.744	0.929

Fig.5. Comparison of Machine Learning Algorithms Previous and Proposed Research Results (Before and After

	Models Results Before Enhancement				**Models Results After Enhancement**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
RNN	0.8011	0.8009	0.8011	0.8004	0.8430	0.8452	0.8431	0.8434
LSTM	0.8081	0.8105	0.8081	0.8085	0.8189	0.8197	0.8190	0.8192
CNN	0.8012	0.8038	0.8012	0.7991	0.8414	0.8425	0.8414	0.8417
GRU	0.8110	0.8109	0.8110	0.8110	0.8457	0.8478	0.8458	0.8461

Fig.6. Deep Learning Models Results Before and After Enhancement

THE RESULT

A. Performance of Machine Learning Algorithms Before Optimization

The performance of machine learning algorithms, specifically Logistic Regression (LR) and Decision Tree (DT), prior to optimization, reflects a baseline in model effectiveness. As shown in Figure 5, the Decision Tree algorithm demonstrated higher accuracy (0.733) and F1-Score (0.705) compared to Logistic Regression, which achieved lower accuracy (0.538) and F1-Score (0.397). This result is consistent with prior research findings, where Decision Tree often outperforms Logistic Regression in datasets with non-linear decision boundaries.

Compared to previous research [13], our baseline results confirm similar trends, particularly the superiority of Decision Tree in balancing precision and recall. However, the need for parameter optimization is evident to further enhance these outcomes.

B. Performance of Machine Learning Algorithms After Optimization

After applying the GridSearchCV optimization technique, significant improvements were observed in both algorithms, with the Decision Tree showing the most notable enhancements. The optimized Decision Tree achieved an accuracy of 0.9236 and an F1-Score of 0.929, indicating a well-balanced model capable of generalizing effectively to unseen data. Logistic Regression, while improving to an accuracy of 0.7461 and F1-Score of 0.744, remained less effective compared to the Decision Tree.

In comparison to previous research post-optimization, our results exhibit a marked improvement. Specifically, the Decision Tree algorithm's performance aligns with state-of-the-art results in the literature, where optimization techniques such as GridSearchCV have consistently been shown to enhance model parameters effectively. The improvement highlights the importance of parameter tuning in achieving a balance between model complexity and predictive performance.

C. Performance of Deep Learning Algorithms Before Optimization

Deep learning models, including SimpleRNN, LSTM, CNN, and GRU, exhibited relatively high baseline performance metrics before applying optimization strategies, as demonstrated in Figure 6. Among these, GRU achieved the highest baseline accuracy (0.8110), closely followed by LSTM (0.8081). The F1-Score for these models ranged between 0.7991 (CNN) and 0.8110 (GRU), reflecting stable initial performance. Despite this, certain limitations in precision and recall indicate room for optimization to address potential overfitting or underfitting issues.

D. Performance of Deep Learning Algorithms After Optimization

The application of advanced training strategies, including Callbacks such as Early Stopping and ReduceLROnPlateau, resulted in substantial performance improvements for the deep learning models. The SimpleRNN model demonstrated the highest percentage improvement, achieving a post-optimization accuracy of 0.8430 and F1-Score of 0.8434, indicating the effectiveness of the applied techniques in mitigating overfitting and enhancing generalization.

Other models, such as GRU and CNN, also benefited from these strategies, achieving accuracies of 0.8457 and 0.8414, respectively. Although the improvements were moderate, they underscore the stability of these architectures. The LSTM model maintained consistent performance with an accuracy of 0.8189 and an F1-Score of 0.8192, highlighting its robustness for sequential data tasks.

References

1. Rashid M.M., Kamruzzaman J., Hassan M.M., Imam T., Gordon S.Cyberattacks detection in IoT-based smart city applications using machine learning techniques. *International Journal of Environmental Research and Public Health, 17*(24), 9347. 10.3390/ijerph17249347
2. Alshahrani M.M.A secure and intelligent software-defined networking framework for future smart cities to prevent DDoS attacks. *Applied Sciences*, *13*(17), 9822. 10.3390/app13179822Tavallaee M., Bagheri E., Lu W., Ghorbani A.UNSW-NB15 Dataset. https://www.unsw.edu.au
3. Alshahrani M.M.A secure and intelligent software-defined networking framework for future smart cities to prevent DDoS attacks. *Applied Sciences*, *13*(17), 9822. 10.3390/app13179822
4. Sulaiman A., Nagu B., Kaur G., et al.Artificial intelligence-based secured power grid protocol for smart city. *Sensors*, *23*(19), 8016. 10.3390/s23198016
5. Ahmed S., Hossain F., Kaiser M.S., Noor M.B.T., Mahmud M., Chakraborty C.Artificial intelligence and Machine learning for Ensuring security in Smart cities. *Data-driven mining, learning, and analytics for secured smart cities* (pp. 23–47). Springer International Publishing AG. 10.1007/978-3-030-72139-8_2
6. Kitchin R.Real-time city? Big data and smart urbanism. *GeoJournal*, *79*(1), 1–14. 10.1007/s10708-013-9516-8
7. Elahi H., Castiglione A., Wang G., Geman O.A human-centered artificial intelligence approach for privacy protection of elderly app users in smart cities. *Neurocomputing (Amsterdam)*, *444*, 189–202. 10.1016/j.neucom.2020.06.149
8. Mengarah R., ALomrani W., Alshahrani T., Alanazi S., Alanazi T.Data Analysis for Predicting Attacks in Smart Systems Using AI Techniques. *University of Tabuk*
9. *Alharthi A., e. a.Robust Security Framework for IoT-Enabled Smart Cities: Leveraging Ensemble Machine Learning Techniques in Fog Computing Environments. https://www.researchsquare.com/article/rs-5197026/v1*
10. *Hussain S., e. a.A novel hybrid intrusion detection framework for securing smart city environments. Journal of King Saud University - Computer and Information Sciences, 37(6), 897–909. 10.1016/j.jksuci.2023.03.010*
11. *Liu Y., e. a.Expert Systems: Cybersecurity Enhancements for Smart City Architectures. Expert Systems, 0.1111/exsy.13556*
12. *ANOH N.G., KONE T., ADEPO J.C., M'MOH J.F., BABRI M.IoT intrusion detection system based on machine learning algorithms using the UNSW-NB15 dataset. International Journal of Advances in Scientific Research and Engineering, 10(1), 16–28. 10.31695/ijasre.2024.1.3*
13. *Kilichev D., T. D., Kim W.Next-generation intrusion detection for IoT EVCS: Integrating CNN, LSTM, and GRU models. Mathematics, 12(4), 571. 10.3390/math12040571*
14. *Hung-Chin Jang, C. C.Urban traffic flow prediction using LSTM and GRU. Engineering Proceedings, 55(1), 86. 10.3390/engproc2023055086*
15. *Chen W., Hao X., Kong C., Wu L.A short-term load forecasting method based on the GRU-CNN hybrid neural network model. Mathematical Problems in Engineering, 2020(2020), 1–10. 10.1155/2020/1428104*
16. *Kasongo S.M., S. Y.Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. Journal of Big Data, 7(1)10.1186/s40537-020-003*