Machine Learning Based Security for Smart Cities

Gabriel Chukwunonso Amaizu *, Jae-Min Lee†, and Dong-Seong Kim†,
*ICT Convergence Research Center †Department of IT Convergence Engineering
Kumoh National Institute of Technology
Email: gabriel4amaizu@gmail.com, (ljmpaul, dskim)@kumoh.ac.kr

***Abstract*—The proliferation and wide usage of the Internet of Things (IoT) and related information and communication technologies (ICT) have led to the emergence of smart cities which comprises ubiquitous sensors, and heterogeneous network architectures. These cities are capable of relaying real-time information about the world which can then be used to improve the Qualify of Life (QoL). However, due to the unprecedented access to the city and personal data by smart city applications, there is an increase in both security and privacy threat. In this study, we propose a stacked generalization machine learning algorithm for the detection of cyberattacks in a smart city. The algorithm was tested using datasets from various smart city infrastructures. Simulation results show a high detection accuracy.**

***Index Terms*—Artificial Intelligence, IoT, Machine Learning, Security, Smart City**

I. Introduction

As the world's population continues to see a steady increase, and cities get more populated, the attention of the research community begins to shift towards the development of cities that caters to the increasing demands of its ecosystem and humans alike. The subsequent proliferation and advancements in the Internet of Things (IoT), big data, and related technologies, coupled with the growing need of addressing urban issues to improve the Quality of Life (QoL) of its residents, have led to the emergence of "smart cities" [1]. A smart city is a technologically advanced metropolitan region that collects data using various electronic technologies, voice activation methods, and sensors. The information obtained from that data is utilized to efficiently manage assets, resources, and services; in turn, that data is used to enhance operations throughout the city. In a smart city, information and communication technologies (ICT), as well as physical devices that are connected to the IoT network are incorporated into city operations and services to maximize efficiency, connect citizens and improve QoL.

Subsequently, smart cities have come to be an all-encompassing term for applications such as smart homes, smart health, smart buildings, smart grids, etc. The concept of a smart city involves an infrastructure of a variety of applications as can be seen in Fig. 1. Although the concept of smart cities has been greatly accepted and regularly improved by the research community, it is still plaque with cybersecurity and privacy issues. [2]. In order to realize the full potential and obtain maximum benefits of a smart city, these ever-growing security and privacy challenges need to be addressed. In this paper, we aim at achieving a more secure smart city by proposing a stacked generalization machine learning model. The model involves the use of random forest, gradient descent, and naive bayes as initial estimators, while a logistics regression was implemented as the final estimator taking its inputs from the initial estimators.

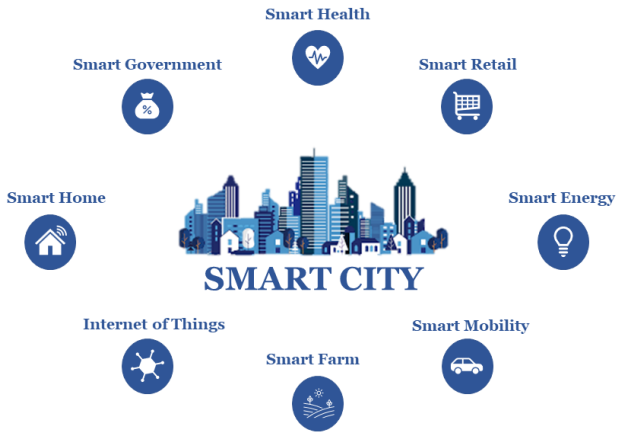

Fig. 1: Concept of smart city with its various applications

II. System Model

This section discusses the system model of this study. The system model of our proposed scheme can be seen in Fig. 2. It starts with the data preprocessing phase which involves the cleaning and balancing of the data. The data is then split into a train and test set where a stacked generalization model is used for training. A stacked generalization method makes use of a combination of estimators/classifiers in training. This method is known to reduce biases in machine learning models. It takes the predictions of individual estimator/classifiers and stacked them up together, then used them as an input to a final estimator/classifier to get a final prediction. The final estimator is typically trained using cross-validation. This final estimator is capable of predicting an attack on numerous smart city applications, hence creating a smart city environment that is secured.

In the proposed scheme, the individual estimators used were random forest, stochastic gradient descent, and naive bayes

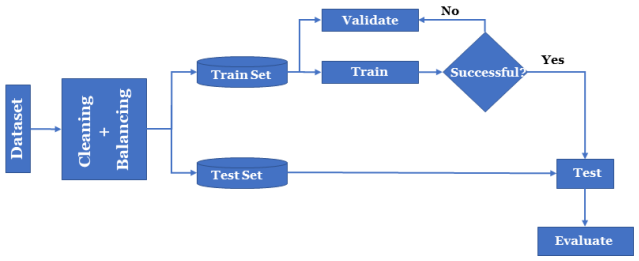

Fig. 2: System model depicting control flow during the entire training process.

TABLE I: Classification results for weather dataset.

	Accuracy	Precision	Recall	F1_score
Decision Tree	96.11	96.43	95.76	96.11
Naive Bayes	49.60	49.64	73.62	49.61
Stochastic Gradient Descent	53.17	53.19	51.28	53.16
Logistic Regression	52.46	52.29	54.02	52.417
Random Forest	96.86	92.93	83.69	88.07
Max Voting	74.69	69.69	85.05	74.11
Stacked Generalization	**97.39**	**97.94**	**97.57**	**97.77**

algorithm. This was due to their better individual performance as will be seen in the results. However logistic regression is used as the final estimator which produced the final predictions.

III. Results

Simulations were carried out using Jupyter Notebook running in anaconda on a Linux machine. In addition to the stacked generalization model, we have also presented results of various standalone machine learning classifiers including decision tree naive bayes, stochastic gradient descent, logistic regression random forest, and a max voting classifier comprising the aforementioned algorithms. Moreover, we have used datasets from four different smart city applications including smart weather, fridge, thermostat, and smart garage. The dataset used for this experiment can be found in [3] The results are contained in Tables I - IV, with each table corresponding to a single smart city application. The collected results include accuracy, precision, recall, and f1 score.

In Table I, the results on a weather dataset are shown. The stacked generalization model is seen to have better results in all metrics followed by the decision tree. Table II contains the classification results from the smart fridge dataset with the stacked generalization model also outperforming the rest on all metrics except for precision with NB coming having a higher value. The results in Tables III and IV shows that the proposed scheme (stacked generalization) also outperformed other classifiers when considering the four metrics used.

IV. Conclusion

The wide and rapid adoption of IoT and ICT applications has given rise to the concept of a smart city. Smart city involves the use of ubiquitous sensors and heterogeneous network architectures in relaying information about the world with the aim of improving QoL. These applications, however, have been prone and subjected to cyberattacks which tend to disrupt the use of such applications and by extension QoL. This paper proposes a stacked generalization machine learning model for the detection of cyberattacks on smart city applications. The model was tested using datasets from applications such as smart weather, fridge, thermostat, and smart garage. All registered an accuracy of over 90% for both training and testing.

TABLE II: Classification results for fridge dataset.

	Accuracy	Precision	Recall	F1_score
Decision Tree	85.39	72.92	85.39	78.67
Naive Bayes	82.44	**84.84**	82.44	83.43
Stochastic Gradient Descent	85.39	72.92	85.39	78.67
Logistic Regression	84.14	81.32	84.14	82.28
Random Forest	85.39	72.92	85.39	78.67
Max Voting	85.39	72.92	85.39	78.67
Stacked Generalization	**90.26**	83.01	**88.12**	**89.50**

TABLE III: Classification results for garage dataset.

	Accuracy	Precision	Recall	F1_score
Decision Tree	89.17	82.19	89	90.22
Naive Bayes	85.27	82.21	89.97	85.92
Stochastic Gradient Descent	85.73	77.78	85.75	87.50
Logistic Regression	52.46	52.29	54.02	52.41
Random Forest	89.17	82.19	89.18	90.22
Max Voting	89.17	82.19	89.18	90.22
Stacked Generalization	**96.86**	**92.93**	**90.69**	**90.07**

TABLE IV: Classification results for thermostat dataset.

	Accuracy	Precision	Recall	F1_score
Decision Tree	82.20	83.74	76.19	79.73
Naive Bayes	70.79	70.42	80.19	81.91
Stochastic Gradient Descent	70.82	70.39	89.24	84.46
Logistic Regression	70.47	70.33	72.68	71.48
Random Forest	82.75	83.16	80.78	61.95
Max Voting	77.06	75.20	73.89	83.19
Stacked Generalization	**92.56**	**92.81**	**91.14**	**91.97**

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01612) and Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003).

References

[1] S. Rani, R. K. Mishra, M. Usman, A. Kataria, P. Kumar, P. Bhambri, and A. K. Mishra, "Amalgamation of Advanced Technologies for Sustainable Development of Smart City Environment: A Review," *IEEE Access*, vol. 9, pp. 150 060–150 087, 2021.

[2] R. Khatoun and S. Zeadally, "Cybersecurity and Privacy Solutions in Smart Cities," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 51–59, 2017.

[3] N. Moustafa, "A new distributed architecture for evaluating ai-based security systems at the edge: Network ton_iot datasets," *Sustainable Cities and Society*, vol. 72, p. 102994, 2021.