IoT –based Adversarial Attack's Effect on Cloud Data Platform Service in Smart Building's Context

Petri Vähäkainu, Martti Lehto, Antti Kariluoto
Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
petri.vahakainu@jyu.fi
martti.lehto@jyu.fi
anjuedka@jyu.fi

Abstract IoT sensors and sensor networks are widely employed in businesses. The common problem is a remarkable number of IoT device transactions are unencrypted. Lack of correctly implemented and robust defense leaves the organization's IoT devices vulnerable to numerous cyber threats, such as adversarial and man-in-the-middle attacks or malware infections. A perpetrator can utilize adversarial examples when attacking machine learning (ML) models, such as convolutional neural networks (CNN) or deep neural networks (DNN) used, e.g., in DaaS cloud data platform service of smart buildings. DaaS cloud data platform's function in this study is to connect data from multiple IoT sensors, databases, private on-premises cloud services, public or hybrid cloud services into a metadata database. This study focuses on reviewing adversarial attack threats towards artificial intelligence systems in the smart building's context where the DaaS cloud data platform services under various information propagation chain structures utilizing ML models and reviews. Adversarial examples can be malicious inputs to ML models providing erroneous model outputs while appearing to be unmodified in human eyes. This kind of attack can knock out the classifier, prevent ML model from generalizing well, and from learning high-level representation, but instead to learn superficial dataset regularity. The purpose of this study is to investigate, detect, and prevent cyber-attack vectors, such as adversarial attacks towards DaaS cloud data platform.

Keywords artificial intelligence-based applications · artificial intelligence · cloud service · data platform · attack vectors · adversarial attacks

1 Introduction

Insecure IoT devices are a problem for cyber-physical systems (CPS), such as smart buildings. In this study, we touch subjects, for example, artificial intelligence (AI), cybersecurity, smart services, and the Data-as-a-Service platform (DaaS), and answer some of the cybersecurity needs of smart buildings. We focus on reviewing the most common adversarial attack techniques that pose a threat to defensive AI systems.

Artificial intelligence is a widely spread vogue word nowadays and is considered as the new "oil" of the future with the potential for significant societal impact. The term AI has been on the surface for many decades, and it was initially presented to mimic the cognitive functions of the human brain. AI can process a significant quantity of data, and it has novel applications. AI has been utilized in various fields, e.g., in construction, education, healthcare, space exploration, and transportation. In the healthcare sector, AI has succeeded in providing accurate diagnoses to prevent skin cancer, treatment recommendations, and provided surgical aid. In the field of smart buildings, AI can assist in finding anomalies and provide future forecasting in order to reduce maintenance costs. It can also function as a defense against cyber-attacks.

Data is paramount to have a properly functioning artificial intelligence. Data should be accessible and correct. Cybersecurity provides means to access the data in different ways. Effective cybersecurity controls provide a cyberspace infrastructure that is reliable and resilient. Lacking or absent controls lead to insecure cyberspace. According to Bayuk et al. (2013), cybersecurity is applied to prevent, detect, and recover from damage to confidentiality, integrity, and availability of information in cyberspace. In order to use all these factors, people, processes, and technologies are used.

In a cyber-physical system, smart sensors automatically measure usage, functions, and variables describing the state of a building (Schmidt & Åhlund, 2018). Due to decreased costs of cloud

computing, IoT sensors, and sensor networks, these techniques are becoming more common. Through provided benefits, they are gaining a foothold in a smart building's cyber-physical system context. Energy, electricity, and water consumption, inside temperature, humidity, and other relevant variables are examined and used to automatically adjust, for instance, the heating system of a smart building. A building can be considered smart, even if only some of these variables are measured.

IoT devices produce ample data. The data gathered from CPS systems provide a significant asset to organizations and artificial intelligence-based smart service providers. Utilizing the data through the DaaS platform and IoT sensors, prosumers can consume, develop, and implement smart services. The services provide added value and benefits to end-users, and they might increase cost savings and gains to organizations. Smart services can be, for example, smart snow plowing service, an AI-based predictive heating or air conditioning system, or a digital caretaker. The possible advantages of the gathered data in implementing smart services are almost endless. Cybercriminals will have a difficult time if the architecture of services is designed well, and data security, integrity, and availability (CIA) issues have been taken seriously.

Unfortunately, a significant amount of IoT devices and sensors are not secure. Their transactions are also unencrypted. They are not easily seen as vulnerable devices as a remarkable amount of them do not have any user interface. These rather new attack surfaces act as entry points for cybercriminals to conduct different kinds of cyber-attacks, who are continuously looking for new ways to exploit vulnerabilities. These perpetrators have, in their use, even more sophisticated attack vectors, such as artificial intelligence-based attacks, in looking for vulnerabilities they can exploit.

Several defense mechanisms presented in studies mainly focus on computer vision, and there exist only some mechanisms, which are designed for the cybersecurity applications in mind. There is an acute need for developing defensive mechanisms to detect and prevent cyberattacks targeting ML model classifiers used in the smart home context.

An adversarial attack is an attack vector created using artificial intelligence. Adversarial attacks are adversarial perturbations constructed deliberately by the perpetrator that are invisible to human observers, but mostly negatively affect deep neural network models. The increase of adversarial attacks towards Machine-Learning (ML) models has brought concern about machine learning security and privacy issues into daylight. In the smart building context, for example, there exist a chance that with adversarial attack perpetrator could fool the ML model and gain entry to the building.

This paper is constructed in the following way. The first chapter is the introduction, where we presented the main concerns caused by adversarial attacks in the smart building's context. In chapter 2, we present the background of artificial intelligence and machine learning. In chapter 3, we explain cybersecurity while in chapter 4, we talk about smart buildings and smart services. We also explain the Data-as-a-Service platform shortly. In chapters 5.1 and 5.2, we review by explaining the adversarial attack methods and their defensive methods, respectively. In chapter 6, we discuss our findings, and chapter 7 concludes our paper.

2 Artificial intelligence and machine learning

Artificial intelligence (AI) is a mathematical approach to estimate a function. It uses data to construct the probabilistic estimate of the behavior of data. The artificial intelligence can be expressed mathematically as f(x): R^n -> R^m, where f(x) is the function to be modeled, R^n represents the real input values and R^m represents the possible real output values. Since the workings of an AI are measured based on the outputs of the model compared to the target values with the given inputs; it can be used as a black-box model to find the corresponding output without caring about the actual form of the function. The black-box type of thinking can mislead to thinking that correctly working artificial intelligence models can add human-like capacities to its functioning. Strictly speaking, the quality and quantity of data, as well as the structure of the model and training time, affect how the AI makes its choices. In practice, artificial intelligence models tend to function poorly when put to work outside their respective fields or input domains. They are especially bad for innovating. For example, an AI capable of predicting the energy consumption of a townhouse will quite likely, after some

adjustments predict the energy consumption values of another similar townhouse adequately but fail when put into predicting stock market values.

Machine learning (ML) research aims to make artificial intelligence more generalizable. Machine learning focuses on teaching machines to learn and adapt on their own (Jordan & Mitchell, 2015), and this is why it is considered as a sub-domain of artificial intelligence. There exist many different methods to teach these learning algorithms, and a method, or rather a tactic, is to re-train the artificial intelligence model after a certain number of calculations or events. There exist three basic ways of training AI systems: supervised, unsupervised, and re-enforcement learning. Unsupervised and re-enforcement learning both belong in the ML field.

Nowadays, the trainable model is very often, neural networks (NN) model. The neural networks model is built from organized and connected layers, which are made of interconnected nodes. Simplistically, each connection has a weight that can be changed based on the inputs and activation functions, such as, rectified linear unit (ReLu) and hyperbolic tangent (tanh). NN model is considered trained when no significant improvement is recorded between the output values of the model and the target values. Some other performance metrics used to define the learning algorithms might also be employed.

According to Ibitoye, Shafiq, and Matraway (2019), in their test feed-forward neural networks (FNN) outperformed self-normalizing neural networks (SNN) in accuracy and precision; however, they did not do well against adversarial attacks. The reason might be that FNN does not normalize values and relies heavily on the ReLu activation function in the hidden layers, whereas, SNN transform all values between 0 and 1. From the self-normalized inputs, it becomes a bit more challenging to learn differences in the data in the hidden layers, and this difficulty becomes beneficial against adversarial attacks. The position is similar to defensive distillation, where the differences between probabilities of different features are purposefully minimized.

The quality and quantity of data is the most crucial factor in the training of artificial intelligence models. Generative adversarial neural networks (GAN) can be used to increase the amount of data from only a small sample to a larger one where the samples are similar to the original samples. Generative adversarial neural networks are a machine learning system of at least two NN models put against each other. According to Radford, Metz & Chintala (2016), one of the neural networks (generator) creates outputs, and the other NN (discriminator) classifies inputs it gets from the generator and database. In the attempt to raise their performance metrics score, they will eventually improve; the classifier will label input data more accurately, and the distribution of the output of the generator will move closer to the distribution of the original data.

3 Cybersecurity

The word “cyber” comes from the Greek word κυβερεω (*kybereo*), which means to direct, guide, and control. Cyber refers to the digital world, which includes the surroundings and being present in our daily lives. Cybersecurity measures are associated with risk management, vulnerability patching, and improving system resilience. There are various research topics available in the field of cybersecurity that include techniques associated with detecting different network behavior anomalies and malware, and IT questions related to IT security. (Lehto, 2015, 3 - 29.) Cybersecurity is a collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance, and technologies that can be used in protecting the organization’s assets (Solms & Niekerk, 2013).

There is no globally accepted definition of cybersecurity, and even the term is widely used. Cybersecurity can be defined as a range of actions taken in defense against cyberattacks and their consequences and includes implementing the required countermeasures. Cisco defined cybersecurity as a practice of protecting systems, networks, and programs from digital attacks (Cisco). These kinds of cyberattacks are often focused on accessing, changing, or destroying sensitive or critical information, blackmailing money from customers, or interrupting normal business operations. Cybersecurity is built on the threat analysis of an organization or institution. The structure and

elements of an organization's cybersecurity strategy and its implementation program is based on the estimated threats and risk analyses. In many cases, it becomes necessary to prepare several-targeted cybersecurity strategies and guidelines for an organization. (Lehto, 2015, 3 - 29.)

Preparations to counter cyber threats are essential to be made and focus on building sufficient protection towards the adverse effects of threats. Successful preparations against cyber threats can be implemented by increasing common knowledge of cyber threats, improving operational capability, and maintaining security. Cyberattacks may not be prevented entirely. Therefore, the critical issue is being able to maintain the ability to function under attack, be able to stop the attack quickly and restore the organization's functions to a previous healthy state before the incident happened. In order to find a solution to these issues require proper legislation and widely open discussion. (Limnéll, Majewski & Salminen, 2014, 107.)

The threat, vulnerability, and risk are not separate concepts; instead, they are firmly intertwined. Risk can be seen as a function of threats exploiting vulnerabilities to obtain, damage, or destroy assets (resources). A threat can exploit a vulnerability, deliberately (intentional threats), or by accident (unintentional threats), and it can only exist if there is a vulnerability that can be exploited. Examples of intentional threats are, for example, malware, espionage, or privacy violation. The vulnerability uses weaknesses or gaps, for example, within technology or in a process related to the information, which can be taken advantage of threats in gaining unauthorized access to an asset. The risk lies in the intersection of assets, threats, and vulnerabilities, and it reflects a potential loss, damage, or destruction of an asset resulting from a threat exploiting a vulnerability. (Flores et al., 2017)

4 Smart buildings and services

Smart buildings are thought of as structures that try to optimize energy consumption with the use of sensors and a smart meter. The meter relays information exchange between the building and smart grid, where a smart grid is a conventional grid that is utilizing information technology for the optimization of energy consumption. Alam et al. (2014) Smart buildings with their corresponding IoT devices can be perceived as a type of cyber-physical (CP) system. According to Legatiuk & Smarsly (2018), sensors, actuators, and the building structure itself belong to the physical domain, and the functioning of the structure with proper controlling actions calculated in the cloud is known as the cyber domain of the cyber-physical system. There are different types of buildings and many kinds of usages for buildings; and therefore, many different kinds of users, we the authors define smart buildings as not only the combination of the two previous definitions but also as a union of technical aspects and humans-cognitive aspects. In other words, a smart building is a cyber-physical system that utilizes IoT technology to minimize energy consumption and maximizes inhabitants' satisfaction under various decision criteria.

Smart services might be best thought of as applications that automate the handling of data and somehow benefit the entire system or subsystem – be it a cyber-physical system or just a cloud-based system. The underlying technology for smart services is, besides web technologies, smart contracts. According to Maleshkova et al. (2016), these services should consume and produce semantic data, while having elements, such as context-based adaptation and rules, which make these services smart.

In order to rationalize the use of smart buildings, the mass of data generated by them and the users ought to be used. The big data holds valuable information, such as information about the condition of a structure, utilization information, and much more. The big data might be stored locally, or in the cloud, however access to it tends to be limited, and data rarely used. For the designed smart services Data-as-a-Service platform might offer a possible solution for the data access and use. The DaaS platform operates by indexing the different data sources, such as databases, and enables access to its data. Data does not need to be moved, which means companies and inhabitants can simultaneously offer their data for use or sell while guarding it. This would guarantee the data necessary for the development of smart services and the continued usage of the services.

Figure 1 presents a simplified illustration of the IoT Data Platform (Data-as-a-service, DaaS). Data can be gathered from various sources, such as IoT –sensors through middleware to Database, Data

Warehouse (DW), or Data Lake. Organizations may have various ways to store the data gathered from multiple sources into, for example, a Database or Data Warehouse. Organizations may also have private or hybrid cloud services created on their premises. The data stored can be accessed through the IoT Data Platform without physically transferring and storing it into one large DW or Data Lake. By utilizing the data through DaaS provides means for software developers to develop and implement smart services, such as AI-based predictive heating or air condition adjusting system, digital caretaker, or an intelligent snow plowing service. Data analysts can utilize and analyze the data to make predictions of the structural health or maintenance needs of the future.

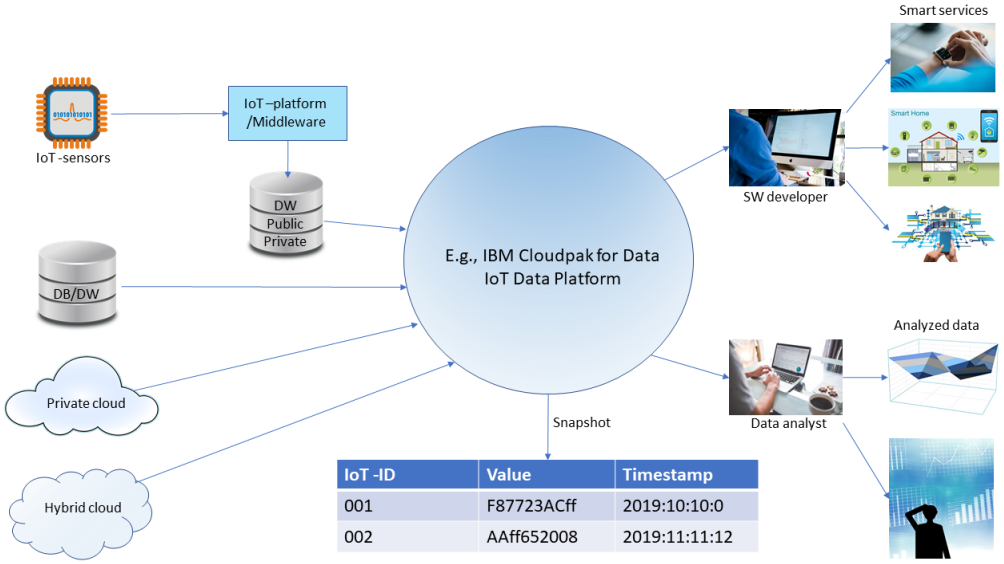

Fig. 1 Simplified illustration of IoT Data Platform (DaaS) –model.

5 Attacks and defenses

5.1 Adversarial attacks

An adversarial attack happens when an adversarial example is sent as an input to a machine-learning model. An adversarial example can be seen as an instance to the input with features that deliberately cause a disturbance in an ML model to deceive the ML model into acting incorrectly and into making false predictions. (Ibitoye et al., 2019) Deep learning applications are becoming more important each day, but they are vulnerable to adversarial attacks. Szegedy et al. (2013) presented that making tiny changes in an image; it is possible to cheat a deep learning model to classify the image incorrectly. Generally, the changes can be minimal and practically invisible to human eyes and can eventually lead to considerable differences in results between humans and trained ML models.

According to Biggio et al. (2013), the perpetrator may have perfect, limited, or zero-knowledge concerning the adversarial setting. In the case of perfect knowledge, the perpetrator has complete knowledge about features and trained models, including classifier type. When the perpetrator's knowledge is limited, she knows features and the classifier, but she does not know about the training data of the classifier. When the knowledge of the perpetrator is zero, she does not know about the type of the classifier and detector's model parameters. This kind of attack-type is considered a black-box attack.

Adversarial attack type can be exploratory, evasion, or causative (poisoning). Exploratory attacks are so-called white-box attacks, in which the adversary knows the classifier algorithm or training data. Exploratory attacks can also be black-box attacks, where the adversary does not know the classifier algorithm or have any knowledge about training data. In evasion attacks, the adversary focuses on specifying the data samples, which may be already misclassified by the target classifier. An example of an evasive attack is spam email generation, which can evade spam detection filters. The original

classifier may be trained with limited training data and then retrained with an additional one. The causative attack manipulates the data at the training time and causes misclassification consequences. These attacks can be used together or individually. (Sagduyu et al., 2019)

White-box attacks require perpetrators to understand the exact structure and parameters of the victim model (actual classifier) in decision-time attacks or learning algorithms in poisoning attacks. In the white-box scenario, the perpetrator has full access to the victim's model, and she knows the ML algorithm being used, including the model's required parameters. In the poisoning attacks, she knows the hyperparameters of the learning algorithm. The idea that the perpetrator has such detailed knowledge of the learning system seems suspicious. There may be indirect ways to obtain an adequate amount of knowledge about a learned model to apply a successful attack scenario. In case of malware evasion attack, a set of features may be public through published work, datasets used to train the detector are public, or there might be similar ones publicly available. The learner might also use a standard learning algorithm to learn the model, such as deep neural network, random forest, or support vector machine by using standard techniques to adjust hyperparameters. This may lead to the situation that the perpetrator can get a similar working detector as the actual one. (Vorobeychik & Kantarcioglu, 2018, 22 - 23)

In the most realistic attack scenario, the perpetrator has access to the victim model's parameters only through a limited interface. Therefore, there is a need to utilize additional strategies to implement attacks without a way to access the victim model's gradients. For example, the model's gradients are not available in the case of black-box attacks. The perpetrator can train another substitute model that is different from the target model to compute the gradients needed for the attack. If the substitute and targeted models operate similarly, there is a high probability that the targeted model will misclassify the adversarial example of the substitute model.

To implement this scenario, the perpetrator needs to collect and label her own training set. The scenario is relatively expensive to implement due to the need for a great number of real input examples and effort to label each example, but the benefit is a lack of need to have access to the victim's model. If the perpetrator is able to send queries to the victim's model by sending it inputs and by watching the returned outputs, she is able to send inputs generated by suitable algorithms to reverse engineer target model with small (or none) amount of training data. The perpetrator does not need to know the architecture used to create a victim's ML model, which can be based on technologies, such as utilizing a support vector machine or a neural network. Papernot e. al (2017).

In the case of black-box attacks, selecting and reducing the number of inputs the perpetrator sends to API to evade the detection can be challenging. Papernot et al. presented a strategy to produce synthetic inputs by using some collected real inputs. In literacy, many papers are focusing on research that utilizes images as datasets (e.g., MNIST or CIFAR). In that kind of case, the perpetrator can, for example, fetch several pictures of the target dataset and use the augmentation technique for each of the pictures to find new inputs that should be labeled with the API. The next step is to train a substitute by sequentially labeling and augmenting a set of training inputs. After the substitute is accurate enough, the perpetrator can launch white-box adversarial attacks, such as FGSM (Fast Gradient Sign Method) or JSMA (Jacobian Saliency Map Approach), to produce adversarial examples to be transferred to the targeted model. (Goodfellow et al., 2018)

A white-box attack uses the target model's gradients in producing adversarial perturbations. FGSM was introduced by Goodfellow et al. (2018) to generate adversarial examples against neural networks. FGSM can be used against any ML algorithms using gradients and weights. It provides low computational cost and, therefore, is cheap to use. The gradient needed can be calculated by using backpropagation, and if internal weights and learning algorithm architecture is known, FGSM is efficient to execute. (Co, 2018) FGSM fits well for crafting many adversarial examples with major perturbations, but it is also easier to detect than L-BFGS and JSMA; therefore, L-BFGS and JSMA are stealthier perturbations, but their drawback is higher computational cost than FGSM. Defense mechanisms can prevent a relatively considerable number of FGSM and JSMA attacks, but L-BFGS is a

brute-force based white-box method, which has a high success rate despite a defense technique if time is not a critical asset. (Goodfellow et al., 2018)

In the smart buildings' context, the perpetrator might attack against the cyber-physical system using an adversarial attack. Assuming that the defending artificial intelligence system is capable of learning. A way to alter the system would be to slowly input data that seems valid to the system via one or many internet-of-things devices. Slowly inputting false data will eventually change the distribution of data that is used in the training of the new and "improved" defensive artificial intelligence. The defensive AI model will also continue to encounter like data used in training during its operation. If the AI were to protect, for example, an intelligent heating system of the smart building, the heating system would end up functioning poorly. Depending on what is the perpetrators end-goal, the intelligent heater could stop heating on the minimum acceptable level during times when indoor heating would be most desired in the inhabitants' view. Other possibilities exist, as well. Perpetrator might also try to fool the smart building's entry control if it relies on image data. For example, Sharif et al. (2016) managed to pose as another person by wearing specially designed adversarial glasses. The use of glasses shows that perturbations can come from other things than altering images with typical adversarial techniques.

5.2 Adversarial attack defense mechanisms

Defending from adversarial attacks is challenging. Empirical studies state that conventional regularization strategies such as dropout, weight decay, and distorting training data with random noise do not present a solution to the problem. According to Samangouei et al. (2018), several defenses have been presented to reduce the effect of adversarial attacks. Defenses against adversarial attacks can be divided to the following areas: 1. modifying the training data to make the classifier more robust against attacks (e.g., adversarial training augmenting the training data of the classifier with adversarial examples), 2. the classifier training process adjustment to decrease the size of gradients, and 3. adversarial noise reduction from the input samples. These methods are efficient against white-box or black-box attacks but cannot cover both types of attacks. In addition, they are designed to avert specific attack models and therefore are not effective against new types of attacks.

Adversarial training injects perturbed inputs, such as adversarial examples, into training data to increase the robustness of the machine learning model. The goal of adversarial training is to defend from adversarial perturbations by training a classifier with adversarial examples. This method can also be applied to large datasets when perturbations are crafted using fast single-step methods. Adversarial training generally attains adversarial examples by utilizing an attack, such as Fast Gradient Signed Method (FGSM), and tries to build adequate defense targeting such an attack. The trained model can indicate poor generalization capability on adversarial examples originated from other adversaries. When combining adversarial training on FGSM adversary with unsupervised or supervised domain adaptation, the robustness of the defense could be improved. Unfortunately, the robustness of adversarial training is possible to evade by applying a joint attack with indiscriminate perturbation from other models. (Song et al., 2019)

The robustness that can be reached by adversarial training leans on the strength of the adversarial examples utilized. Training a model by using a fast non-iterative FGSM –attack produces a robust protection towards non-iterative attacks, but not against PGD –based adversarial attacks. PGD –based adversarial training can be considered as strong enough to sustain against powerful attacks, and it has been stated to be a state-of-the-art defense model. Despite shortcomings of adversarial training, it stays among one of the few efficient methods to strengthen a network towards attacks. Though its high computational complexity and cost can prevent or at least decrease utilizing it as a robust defensive method. (Shafahi et al., 2019) According to the research papers, a deeper understanding of adversarial training and a clear direction for further improvements is also principally missing.

Defensive distillation can be counted to one of the adversarial training techniques providing flexibility to an algorithm's classification process, making the model less prone to exploitation. Teenu & Tony (2019) presented a defense distillation method that can reduce the input variations making

adversarial crafting process more difficult, helping DNN to generalize the samples outside the training set and reducing the effectiveness of adversarial samples on DNN. The distillation method transfers the knowledge from one architecture to another by decreasing the size of DNN. In distillation adversarial training, one model can be trained to predict the output of probabilities of another model trained on an earlier baseline standard. Defense distillation provides the advantage of being compliant with yet unknown threats. Usually, the most efficient adversarial defense training methods demand interminable input of signatures of known vulnerabilities and attacks into the system. The distillation provides a dynamic method requiring less human intervention. As a drawback, if a perpetrator has a lot of computing power available and the proper fine-tuning, she can utilize reverse engineering to find fundamental exploits. Defense distillation models are also vulnerable to poisoning attacks in which a malicious actor corrupts a preliminary training database. (DeepAI)

Adversarial noise reduction from the input samples concerns the most image datasets and applications in that area. As time-series data (e.g., weather or sensor data) is generally uni-variate or multi-variate data, the noise present in the data is in the form of missing values or different kinds of signs. In this case, missing value related techniques, such as moving average or normalization, can be utilized. According to Moosavi-Dezfooli et al. (2019), measures to defend against adversarial perturbations have recently taken place by using stacked denoising auto-encoders to mitigate perturbations. The same method has been under research to denoise adversarial examples. Alternative generative models (e.g., GAN) have also been used to project malicious samples of diverse datasets. Unfortunately, these methods have mainly been applied to datasets such as MNIST, CIFAR, or ImageNet, and there is no guarantee that attack or defense strategies could work on other kinds of data. However, an interesting observation is that an ordinary JPEG and JPEG2000 compression algorithms can act as a potential defense measures against adversarial examples, and according to Ayademir et al. (2018) they both increase the classification accuracy of adversarial images and efficiently work in defiance of Fast Gradient Sign Method (FGSM) and Basic Iterative Method (BIM).

6 Conclusion

The merits of this paper are the following. We provided discreet background information concerning artificial intelligence and data as a platform service in the field of cybersecurity. We conducted a literary review on adversarial attack methods, machine learning, and artificial intelligence-based defenses. In this conclusion, we discuss our findings in the smart buildings' context and offer recommendations to the threat posed by adversarial attacks.

Protecting smart buildings is necessary to avoid hazards caused by cyberthreats, e.g., violations of privacy, data thefts, malicious acts of vandalism, insider threats, and others. Quality data from, for example, the cyber-physical system or IoT devices, is needed to train artificial intelligence solutions. However, open-source IoT data for research purposes is difficult to attain. Data-as-a-Service platform and smart services might offer a solution to usability questions by helping to automate the flow of data and transactions related to it. Recently, vulnerabilities to input samples, such as adversarial examples, have been found in deep neural networks.

One way to defend dynamically is to use an ensemble of multiple differently made AI models, which have been trained with quality data to combat either specific or generic attacks. The choice of training multiple models for the ensemble can, in some cases, improve robustness, like in the case of Ibitoye et al. (2019). We suggest that one might want to use defensive distillation to narrow the classification manifold. It would also be preferable to utilize GANs in the adversarial training to make the ML model to learn the differences between real inputs and the adversarial attacks. Adversarial noise removal, for example, with stacked denoising autoencoders, might also work to reduce malicious input effectiveness. However, applying powerful and robust defense mechanisms on ML classifiers may cause too high overhead that weakens the classifier performance. Challenges also arise when trying to transfer attacks generated for one model to other models. In the case of IoT devices, a dynamic and fit defense mechanism to detect and prevent sophisticated adversaries should be explored. At the very least, the defense should be separated from the controlling systems.

References

Ayademir, A. E, Temizel, A. & Temizel, T. T. (2018). The Effects of JPEG and JPEG2000 Compression on Attacks Using Adversarial Examples. arXiv:1803.10418 [cs.CV].

Bayuk J L, Healey J, Rohmeyer P, Sachs M H, Schmidt J, Weiss J. (2012). Cyber security policy guidebook, first edition. Wiley & Sons Inc, USA.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G. &Roli, F. (2013). Evasion Attacks Against Machine Learning at Test Time. arXiv:1708.06131v1 [cs.CR]

Cisco. What is Cybersecurity? Accessed 12.9.2019 https://www.cisco.com/c/en/us/products/security/what-is-cybersecurity.html.

Co., K., T. (2018). Bayesian Optimization for Black-Box Evasion of Machine Learning Systems. Imperial College London, Department of Computing.

DeepAI. What is Defensive Distillation? Accessed 9.10.2019 https://deepai.org/machine-learning-glossary-and-terms/defensive-distillation.

Flores, C., Guasco, T. & Leon-Acurio, J. (2017). A Diagnosis of Threat Vulnerability and Risk as IT Related to the Use of Social Media Sites When Utilized by Adolescent Students Enrolled at the Urban Center of Canton Canar. International Conference on Technology Trends, 199 – 214.

Goodfellow, I., McDaniel, P. &Papernot, N. (2018). Making Machine Learning Robust Against Adversarial Inputs. Communications of the ACM, 61(7), 56 – 66.

Ibitoye, O., Shafiq, O. &Matrawy, A. (2019). Analysing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks. Cornell University, arXiv:1905.05137 [cs.NI].

Jordan, M., I., Mitchell, T., M. (2015). Machile learning: trends, perspectives, and prospects. Sciencemag.org, Science, vol 349, issue 6245.

Legatiuk, D., Smarsly, K. (2018). An abstract approach towards modeling intelligent structural systems. 9th EWSHM 2018. Creative Commons CC-BY-NC licence https://creativecommons.org/licenses/by-nc/4.0

Lehto,M. (2015). Phenomena in the Cyber World. M. Lehto & P. Neittaanmäki (Edit.) Cyber Security: Analytics, Technology and Automation. Berlin: Springer.

Limnéll, J., Majewski, K. &Salminen, M. (2014). Kyberturvallisuus. Saarijärvi: Docendo.

Maleshkova, M., Philipp, P., Sure-Vetter, Y. & Studer, R. (2016). Smart Web Services (SmartWS) – the Future of Services on the Web.

Moosavi-Dezfooli, S-M., Shrivastava, A. & Tuzel, O. (2019). Divide, Denoise, and Defend against Adversarial Attacks. arXiv:1802.06806 [cs.CV].

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z., B., & Swami, A. (2017). Practical Black-box Attacks Against Deep Learning Systems Using Adversarial Examples. In Proceedings of the ACM Asia Conference on Computer and Communications Security, UAE. New York: ACM Press.

Radford, A., Metz, L., Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434v2 [cs.LG] 7 Jan 2016.

Samangouei, P, Kabkab, M. &Chellappa, R. (2018). Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. International Conference on Learning Representations, Vancouver Convention Center, BC, Canada.

Solms, R. & Niekerk, J. (2013). From Information Security to Cyber Security. Computers & Security, 38(13), 97-102.

Song, C., He, K., Wang, L. & Hopcroft, J., E. (2019). Improving the Generalization of Adversarial Training with Domain Adaptation. International Conference on Learning Representations, New Orleans, Lousiana, United States.

Sagduyu, Y., E., Shi, Y. & Erpek, T. (2019). IoT Network Security from the Perspective of Adversarial Deep Learning. Department of Electrical and Computer Engineering, Virginia Tech, Arlington, USA.

Schmidt M, Åhlund C. (2018). Smart buildings as Cyber-Physical Systems: Data-driven predictive control strategies for energy efficiency. Renewable and Sustainable Energy Reviews, Volume 90, 742-756. https://doi.org/10.1016/j.rser.2018.04.013

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L., S., Taylor, G. & Goldstein, T. (2019). Adversarial Training for Free! arXiv:1904.12843 [cs.LG].

Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M., K. (2016). Accessories to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. DOI: http://dx.doi.org/10.1145/2976749.2978392.

Szgedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013). Intriguing Properties of Neural Networks. arXiv preprint arXiv:1312.6199.

Teenu S. J & Tony T. (2019). DOI: 10.4018/978-1-5225-8407-0.ch007

Vorobeychik, Y. &Kantarcioglu, M. (2018). Adversarial Machine Learning. Synthesis Lectures of Artificial Intellignece and Machine Learning. Morgan & Claypool, USA.